TVA in the wild: Applying the theory of visual attention to game-like and less controlled experiments

Alexander Krüger*[1] Jan Tünnermann*[2]
Lukas Stratmann[3] Lucas Briese[1] Falko Dressler[3]
Ingrid Scharlau[1]

* both authors contributed equally

[1]Paderborn University, Faculty of Arts and Humanities, Psychology
[2]University of Marburg, Department of Psychology
[3]TU Berlin, School of Electrical Engineering and Computer Science

August 2020

1 Abstract

As a formal theory, Bundesen's theory of visual attention (TVA) enables the estimation of several theoretically meaningful parameters involved in attentional selection and visual encoding. As of yet, TVA has almost exclusively been used in restricted empirical scenarios such as whole and partial report and with strictly controlled stimulus material. We present a series of experiments in which we test whether the advantages of TVA can be exploited in more realistic scenarios with varying degree of stimulus control. This includes brief experimental sessions conducted on different mobile devices, computer games, and a driving simulator. Overall, six experiments demonstrate that the TVA parameters for processing capacity and attentional weight can be measured with sufficient precision in less controlled scenarios and that the results do not deviate strongly from typical laboratory results, although some systematic differences were found.

2 Introduction

In most areas, psychological research methods stress the value of strict control: According to common psychological thinking, drawing conclusions from data presupposes an appropriate choice and manipulation of independent variables, tight control of possible confounding factors, reduction of random noise, and randomization of participants. In more everyday settings—"in the wild"—such

control would not be possible which is one of the main reasons to study phenomena in the laboratory.

However, through the reproducibility crisis (e.g., Open Science Collaboration, 2015), we have learned that even data obtained in controlled laboratory studies are far less reliable than had been expected. We do not aim to get into details of what can be expected in terms of reproducibility and what the causes of the present crisis are (for a summary of potential problems with scientific practices in psychology see, e.g, Chambers, 2017), but want to stress one of those causes, a lack of cumulative theory (Muthukrishna & Henrich, 2004). Without theoretical frameworks and respective formal models, it is difficult to come up with precise and unambiguous predictions for yet unobserved situations that allow testing hypotheses. If these predictions are not precise and do not adhere to some formal framework, separating expected from unexpected results is difficult and undesired flexibility when interpreting results can hinder true progress. Theoretical frameworks also enable the consistent use of more than one empirical approach per research question and provide a means for integration across different disciplines.

In the present article, we explore a new approach in the field of visual processing and selective attention in which we turn around the usual scheme: Instead of adhering to the strictest laboratory settings and model-free null-hypothesis significance testing, we employ more flexible game-like experimental paradigms with rigorous models that formally link the results of the different experiments we conduct and existing findings in the literature. Our question is to which degree results obtained under such conditions match up with measurements obtained with a lab- and model-based agenda. Nested in this overall question, we also ask whether orientation and color salience bias temporal-order judgments as reported in earlier studies (Krüger et al., 2016, 2017).

Administering fundamental attention experiments in a more flexible, game-like manner potentially brings many advantages that might make up for the loss in experimental control and a possibly increased (but quantifiable) uncertainty in the results. For instance, more flexible experiments could be delivered via app stores and web browsers to large and diverse participant pools and motivational elements can be more easily integrated in game-like scenarios. In the end, easy large-scale access, flow and motivation, combined with a model-based evaluation could lead to a superior overall data quality despite losses in experimental control.

2.1 A simple task and model for investigating visual processing and selective attention "in the wild"

Formal models are an important means to foster cumulative theory and an important part of scientific progress. Many topics have a research history of several decades, and it is almost indefensible not to try to formalize the core knowledge that researchers have already obtained. Verbal research summaries are poor surrogates for this (Meehl, 1990); they may help to derive hypotheses that can be put to an empirical test, but are prone to inexactness and ambiguity and much less suited for describing complex relationships between possible influences.

Put very generally, the main function of a formal model in empirical sciences is to connect theoretical considerations and data. Models are customizations of theories such that these become applicable to some of the concrete properties of the phenomena by filling in the gaps between latent causes and data (Bailer-Jones, 2009). According to Bailer-Jones, two important parts of modeling are the theoretical model and the data model. The theoretical model is derived from the theory, and it is necessary as a model of the situation in which the data is observed. It represents the theoretical considerations in a specific situation. The data model is necessary to deal with the uncertainty arising from data collection in a specific experiment, for instance measurement uncertainty. Modeling thus provides a tight coupling of theory and data (Krüger et al., 2018).

For the present work, we derive a formal model from Bundesen's theory of visual attention (TVA; Bundesen, 1990; Bundesen & Habekost, 2008). Its parameters are defined theoretically and cognitively specific (e. g. Habekost, 2015). Different from unspecific parameters such as error rates or response times, TVA's parameters have a closely defined meaning that can be traced into cognitive functions. This makes them – and experimental tasks such as those presented below – very valuable for answering theoretical as well as applied questions. We will come back to this topic in the General Discussion. With a TVA-based model we set up the data model as a Bayesian parameter estimation scheme. This enables the estimation of attention and visual processing parameters of the individual participants and the whole group. Moreover, in this approach the uncertainty of the estimates is explicitly available for subject- and group-level estimates as well as for comparisons between conditions.

How does TVA model visual stimulus processing? TVA is a race model in which stimuli in the visual field compete for being encoded into visual short term memory (VSTM). Once stimuli are encoded, they can undergo further processing, being transferred to other memory systems or guiding behavior. The race for VSTM occurs in a two-wave procedure (Bundesen et al., 2005): In an unselective first wave, "attentional weights" are assigned to every stimulus x in the visual field:

$$w_x = \sum_{j \in R} \eta(x, j)\pi_j \tag{1}$$

where R represents feature categories that might be relevant in a task (e.g., colors if the task is to report, say, blue and red objects), $\eta(x, i)$ is the sensory evidence that stimulus x has feature j (e.g., medium for a pink stimulus being red), and π is the pertinence of feature j (e.g., high for target colors blue and red and low for other features).

In the second wave of processing (selective wave), these weights determine how processing resources (TVA parameter C) are distributed across the visual field, leading to a processing rate for each stimulus. This processing rate determines if and when a stimulus is encoded in VSTM. Higher rates lead to earlier and more certain encoding. At lower rates, stimuli proceed more slowly and are less likely to be encoded because VSTM might be filled up before they finish processing. The VSTM capacity (TVA parameter K) is typically smaller than four items.

The formal calculation of processing rates will be explained later in section 3.2 in the context of the present experimental paradigm.

Summing up, TVA makes a quantitative connection between latent but theoretically interesting distinct components of attentional processing such as attentional weights, VSTM, and processing speed. It has been used with different experimental paradigms: whole and partial report (e.g., Bundesen & Habekost, 2008), attentional dwell time (Petersen et al., 2013), or temporal-order judgments (Tünnermann et al., 2017). Recent developments have been summarized by Bundesen, Vangkilde, and Petersen (2015). Moreover, TVA integrates different views, such as behavioral and neuronal interpretations of visual processing (Bundesen et al., 2005). It is also used in the clinical context, facilitating new diagnostic applications (for a review, see, Habekost, 2015).

In the present article, we send TVA into the wild by relaxing stimulus control on the one hand and embedding the task into game-like dynamic scenarios on the other hand as motivated above. Experiments 1, 2, and 3 contrast a typical lab experiment with one running in a browser on mobile devices. Experiments 4, 5, and 6 implement both factors by using a game engine and a gaming task (flying and driving a bike). One might object that none of this releases TVA into real life or real wild. However, compared to typical psychological experiments with their strict control, we take, to say the least, several large steps, and prepare the ground for possible further progress. Before turning to the individual experiments, we describe the temporal-order judgment (TOJ) paradigm, which is at the core of all experiments, in general and how exactly a model of TOJs can be derived from TVA.

3 General Method

3.1 Experimental paradigm and relationships between the experiments

The experimental paradigm is a temporal-order judgment, an easy experimental task with a long tradition of lab-based research (Sternberg & Knoll, 1973) in which the participants indicate which of two stimuli appears first or, alternatively, second. Readers familiar with TVA may wonder why we did not use one of the more common tasks such as whole and partial letter report or combiTVA (Bundesen, 1990; Vangkilde et al., 2012). One reason is that the TOJ task is very simple and can be done by very different groups including children (Petrini et al., 2020), animals (Wada et al., 2005) and neurophysiologically impaired persons (Rorden et al., 1997). The other reason is that the more common TVA tasks presuppose a large set of equally recognizable well learned stimuli, for instance letters. The present method works with virtually any material as long as two asynchronous stimuli can be presented

In the TOJ, we use a flicker instead of the more common onset of stimuli (e.g., Tünnermann, 2016) because it is better suited for estimation of attention parameters in multi-element displays (Krüger et al., 2016). The flicker is realized

by a temporary change of the stimulus display. It is implemented by an offset and re-onset after a brief delay of a few hundredth of a second. The two stimuli are clearly identifiable because they are marked by a special feature such as size or orientation, or, in one of the experiments (Experiment 6), are pointed out to the participants as such.

The flicker of the two targets (depicted in abstract form in Figure 1) is separated by an interval that we call SOA (stimulus onset asynchrony) which is in accordance with the TOJ literature although, strictly speaking, it is a flicker onset asynchrony in our experiments. The range of SOAs is chosen so that the judgment accuracy of the participants varies between chance (at an SOA of 0 ms) and few mistakes at the largest SOA.

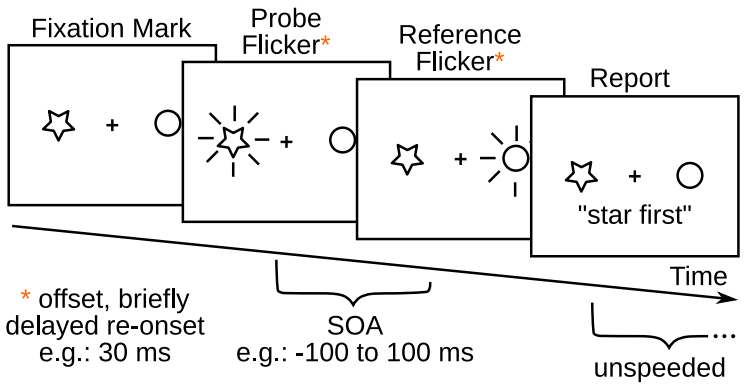

Figure 1: Illustration of a typical TOJ task. This example sequence shows a negative SOA, at which the *probe* stimulus (here a star) leads. At positive SOAs the *reference* (circle) would lead

.

In all experiments, we increase the salience of one stimulus in an experimental condition and keep it equal for both stimuli in a control condition. The salience boost is known to bias attention and consequently influence processing speed and order judgments. The resulting pattern of attentional weights—balanced in control conditions and biased in favor of attended stimuli—is well documented in laboratory experiments (Krüger et al., 2016, 2017). Whether and to what degree this pattern can be detected "in the wild" is a research question we address here. Even in lab settings, the distribution and activation of attentional resources likely are subject to influences beyond the experimental stimulation. Some attention might be directed at irrelevant parts of the apparatus (e.g., the layout of response keys), and motivation might modulate the overall resources dedicated to correct execution of the task. Such influences might be both stronger and more diverse outside the lab, especially when the task is embedded in less restrictive contexts. They might also work in both directions: Motivation, for instance, could be higher, leading to more pronounced effects in less restrictive settings while the distribution of attention over the experimentally relevant stimuli might

be diluted by uncontrolled influences of the environments. It is outside the scope of the present article to tease apart all the different influences on attention and quantify how their impact differs between lab and "wild" settings. Instead, we are interested in whether we can find the known patterns caused by salience-induced attention biases despite the variable and unknown influences. The model-based estimation of meaningful attentional parameters aids the comparison of the outcomes. However, the degree to which our experiments deviate from the typical lab setting varies between the experiments and some include a typical lab-version of the task as a control. We present the experiments ordered by decreasing restrictiveness and experimental control and increasing complexity of the overall task.

In Experiments 1 to 3, we conduct—with different salience manipulations—a typical lab TOJ task both in the lab and on an uncontrolled selection of mobile devices. While the mobile experiments retain several aspects of the typical presentation (e.g. non-interactive one-shot trials followed by keypress responses), other factors become less controlled (e.g., stimulus size and viewing distance, or how well web browsers on mobile devices implement the intended timing). Experiments 4 to 6 embed the TOJ task in interactive environments. Experiments 4 and 5 use games, which introduce several dynamic aspects such as changing object sizes and positions caused by the apparent ego-motion although the experimental displays are still artificial (just more dynamic) multi-element arrangements. Experiment 5 explores the influence of the additional factors adaptive vs. constant gaming speed and SOA size on the attention measurements. During these experiments, the responses are given by controlling game elements with the computer keyboard. Experiment 6 finally leaves computer keyboards or response boxes behind and places the participants on an actual bicycle on a roller trainer which is connected to a traffic simulation. Here participants perform TOJs by navigating the bike over simulated objects that implement the flicker task. The objects with the experimental salience manipulation compete with a multitude of visual influences experienced in dynamic scenes.

3.2 Formal model

How can insights on visual processing and attention be gained from TOJ data? Fitting traditional psychometric functions (based on the logistic function or the cumulative Gaussian distribution) provides only relative latency measures (e.g., "stimulus x is perceived as appearing 20 ms earlier than stimulus y") and discrimination performance measures. These components provide no direct information about how attention is distributed across the stimuli and how fast these are processed (for instance, it cannot be distinguished whether attended stimuli are processed faster or if unattended ones are processed slower; Tünnermann et al., 2015).

The model of TOJs used to estimate relative attentional weights and processing speed parameters from all experiments in the present study can be derived from Bundesen's (1990) TVA (Tünnermann et al., 2015):

According to TVA, the probability that a stimulus x is encoded until time t

is given by

$$F(t) = \begin{cases} 1 - e^{-v_x(t-t_0)} & \text{if } t > t_0 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where v_x is the rate with which a stimulus is encoded into VSTM and t_0 is a threshold time. For presentations up to t_0, no effective encoding occurs. In TVA, v_x can be further decomposed into overall processing capacity and relative attentional weights. We use the term "attentional weight" to refer to TVA's relative attentional weight (cf. Equation 3 to 8). These parameters are composed of more fine-grained ones that model pertinence (low-level filters, e.g. "red is important, blue not") and bias (report category biases, e.g. "letters are to be reported, numbers must be ignored"). The formal model of how these fine-grained components produce attentional weights can be found, for instance, in Tünnermann et al. (2015). For the present study, it is sufficient to relate TVA's encoding to TOJs. Assuming that temporal order is judged based on the VSTM entry order, the probability of reporting the probe stimulus as appearing first ($P_{p^{1\text{st}}}$) can be calculated as follows:

$$P_{p^{1\text{st}}}(v_p, v_r, \text{SOA}) = \begin{cases} 1 - e^{-v_p|\text{SOA}|} + e^{-v_p|\text{SOA}|}\left(\frac{v_p}{v_p+v_r}\right) & \text{if SOA} < 0 \\ e^{-v_r|\text{SOA}|}\left(\frac{v_p}{v_p+v_r}\right) & \text{if SOA} \geq 0. \end{cases} \tag{3}$$

Here, v_p and v_r are the processing rates of the probe and reference stimuli with which they race for VSTM encoding. The first part of the first case, for the SOAs smaller than zero (in which the probe leads), is the probability that the probe is encoded before the reference even starts racing for VSTM ($1 - e^{-v_p|\text{SOA}|}$). The second describes the probability that this has not happened ($e^{-v_p|\text{SOA}|}$) and that the probe wins the race for VSTM when both, probe and reference, race together ($v_p/(v_p + v_r)$); Luce's choice axiom (Luce, 1977). The second case, for SOAs larger than zero (in which the reference leads), the probability of "probe first" judgments is $e^{-v_r|\text{SOA}|}$, the probability that the reference is not already encoded before the probe is shown and that the probe wins when both race together ($v_p/(v_p + v_r)$). Note that this model does not include t_0. Because t_0 is assumed to be equal for both stimuli (adding equal latencies), it cancels out in the equations (cf. Tünnermann et al., 2015).

For the present study, it is advantageous to re-parametrize this model so that instead of individual processing rates v_p and v_r the overall processing rate C and attentional weights (w_p^*) can be estimated. According to TVA, the overall processing rate C is the sum of the processing rates of the individual encodings:

$$C = \sum_{x \in S} \sum_{r \in S} v(x, i). \tag{4}$$

Moreover, TVA states that the individual rates $v(x, i)$ with which stimuli x are encoded into VSTM as members of category i are calculated as

$$v(x, i) = \eta(x, i)\beta_i \frac{w_x}{\sum_{z \in S} w_z} \tag{5}$$

where $\eta(x, i)$ is the sensory evidence that stimulus x is a member of category i and β_i is a bias for encoding stimuli as members of category i. The w are attentional weights and S refers to all stimuli in the visual field. In the following we will use v_p and v_r for simplicity to refer to $v(p, i)$ and $v(r, i)$, the rates with which probe and reference are encoded as members of their report categories.

The experiments in the present study (in line with other TVA-based TOJ experiments) are designed so that both $\eta(x, i)$ and β_i are the same for the probe and reference stimuli. That is, neither of the two has a higher η (e.g., better visibility) nor is its report category more important (higher β) than the other. Therefore, only the attentional weight part ($\frac{w_x}{\sum_{z \in S} w_z}$) of Equation 5 is different for the two stimuli. Consequently we can assume that the processing rates are equal when they are divided by the attentional weights part:

$$v_p \frac{w_p + w_r}{w_p} = v_r \frac{w_p + w_r}{w_r} \tag{6}$$

This can be rearranged to

$$v_p + v_p \frac{w_r}{w_p} = \underbrace{v_r + v_p}_{C} \tag{7}$$

where $v_p + v_r$ (all individual processing rates) can be substituted with C (cf. Equation 4). Taking further into account that $w_p + w_r = 1$ (because there are only two stimuli that acquire attentional weights) the equation can be simplified to

$$v_p = C \cdot \underbrace{\frac{w_p}{w_r + w_p}}_{w_p^*} \tag{8}$$

That is, the processing rates v_p (and similarly v_r) in Equation 3 can be replaced by the term above. The main benefit of this is that when fitting experiments with more that one condition, a common C can be estimated for both conditions whereas individual w_p^* are estimated for each condition. Note that w_p^* is a relative weighting (Krüger et al., 2017; Tünnermann & Scharlau, 2018c) that expresses the attentional advantage of the probe stimulus relative to the attentional weighting of all modeled stimuli. For the sake of simplicity, we call w_p^* the "attentional weight". The model structure is further detailed below.

3.3 Bayesian parameter estimation

The TVA-based TOJ model derived above is embedded in a hierarchical Bayesian parameter estimation. The model structure is depicted in Figure 2. Note that the inner structure is contained twice to model one "neutral" condition (N) and one "salient" condition (S) in agreement with the design of most of the experiments of this study. A common C (TVA's overall processing rate) is estimated across the conditions. Earlier research has shown that C remains unchanged under attentional manipulations (e.g. Krüger et al., 2016) and hence we can pool the

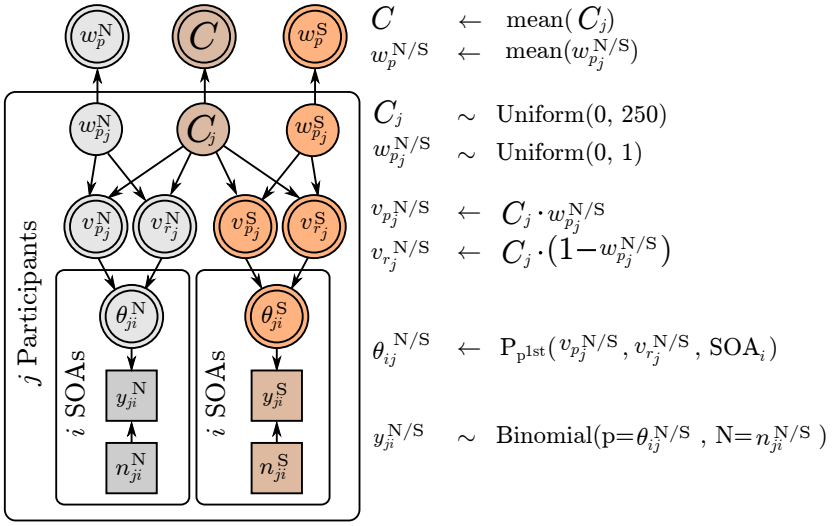

Figure 2: Hierarchical model structure (left) and priors and deterministic dependencies (right).

information about C from both conditions. However, for each condition an individual attentional weight of the probe stimulus w_p^* is estimated (note that w_r^*, the reference's weight can be obtained as $1 - w_p^*$). For neutral conditions, w_p^* is expected at .5, indicating that attention is equally divided between probe and reference. For salient conditions, w_p^* is expected to be larger than .5, indicating that attention is biased toward the salient probe stimulus. As indicated in Figure 2, overall estimates of the w and C parameters are obtained as the means of the corresponding distributions estimated for the participant.

The models were implemented in pymc3 (Salvatier et al., 2016) and estimated using the NUTS sampler (Hoffman & Gelman, 2014) with 5000 samples (in each of four chains). For attentional weights, uniform priors with the range zero to one (equal a priori probabilities for all possible attentional weights) were used. For the overall processing rate C a uniform range corresponding to zero to 250 Hz was used (0–0.25 items/ms). This reflects a conservative choice, given that earlier studies found C typically around 70 Hz. Where not indicated otherwise, estimates in the results section refer to group-level estimates obtained as means of the participant-level posteriors. We report the full posterior distribution together with the mode (as a point estimate of a parameter) and the boundaries of 95 % Highest Probability Density (HPD).

4 Experiment 1

In this first experiment, we move the experimental setup into the "wild" but keep the experimental paradigm identical to typical lab-based TOJ experiments

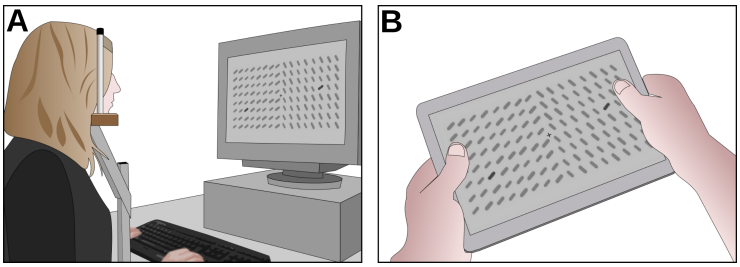

Figure 3: **A** Experimental setup in the lab condition with accurate timing on a CRT monitor, controlled viewing distance with a head rest. **B** Browser condition on a mobile device with less control of stimulus timing, viewing distance and posture.

(Krüger et al., 2016, 2017). In fact, we include a within-participants control condition performed in the usual lab setup. In the experimental condition, we have participants perform a typical lab task in the web browsers on their own mobile devices under typical office conditions. Why might researchers be interested in using typical tasks but collecting data "in the wild"? After all, we are losing, at least to some degree, control over the exact circumstances under which the experiment takes place, such as the general viewing conditions, and spatial and temporal aspects of the presentations. The use of web-based setups enables crowd-sourcing large datasets and can help to avoid WEIRD samples(western, educated, from industrialized, rich, and democratic countries; Henrich et al., 2010). Moreover, when researchers have no or limited access to laboratories, online experiments can be a way out. Our experiments predated COVID-19 but certainly the pandemic lead to an urgent need for online experiments in most psychology labs.

Especially researchers who use psychophysical tasks, such as the TOJ in the focus of the present study, typically shun away from giving up control of the most basic variables such as accurate spatial and temporal stimulus presentation. Here, we ask whether one can still obtain useful insights from such "uncontrolled" data. The formal model and Bayesian estimation scheme enable the estimation of meaningful parameters and quantification of uncertainties. The rationale of Experiment 1 is to compare highly controlled lab-based TOJ data collection with less constrained data collection to check whether under such conditions the typical patterns can be replicated, and to assess the uncertainty of the estimates and quantify the model fit.

In this experiment, we had participants perform a typical TOJ experiment. In the experimental condition, we increased the salience of one target whereas salience was kept balanced in neutral control trials. In a "lab" condition, participants performed the task in a usual lab setting with controlled viewing distance to a time-accurate CRT monitor. In the browser condition, participants used their own laptops, tablets, or mobile phones with various display sizes and uncontrolled viewing distance. Based on earlier studies, we expect w to be increased for

a more salien target. We have no specific expectations concerning the influence of the controlled vs. noncontrolled environment although we expect any possible changes to be at most medium (cf. Semmelmann & Weigelt, 2017). Because we are interested to which degree the parameters from lab-based experiment and experiments outside the lab reflect the same construct, we report their correlations.

4.1 Method

4.1.1 Participants

From earlier experiments that combined the TOJ paradigm with TVA modeling (Krüger et al., 2016, 2017), we know that we need approximately 30 participants to reach appropriately precise parameter estimates in experiments manipulating salience and we aimed at this number in all the experiments. Due to organizational reasons, more participants were available in some of the experiments. Because the outcome of Bayesian parameter estimation does not depend on sampling intentions or stopping rules (Dienes, 2011), we made use of the opportunity to further increase precision of parameter estimation and included these participants in the experiments and analyses.

All experiments were approved by the ethics committee of Paderborn University. Thirty-seven persons (20 male and 17 female; $M_{\text{age}} = 24.08$, range 19–62) participated. All participants were students or members of Paderborn University. Each participant gave informed written consent, reported normal or corrected-to-normal visual acuity and received course credit or was payed 8 Euro per hour. In the browser condition, across all participants, 26 different combinations of devices, operating systems (and versions) and browsers (and versions) were used. If only the major release version of operating system and browser are taken into account, the number of different combinations is 23. Two participants completed only one instead of two sessions of the lab condition. Because the Bayesian analysis accounts for the fact that these data sets are less informative, we include these participants in the analysis. One participant produced no usable data set due to a configuration problem with the monitor (see Footnote 1).

4.1.2 Apparatus

The lab condition was conducted on a Microsoft Windows 10 PC with a 22" Iiyama Vision Master Pro512 (40.4 cm × 30.3 cm) CRT monitor. A resolution of 640 × 480 pixels was used with 32-bit colors and a refresh rate of 100 Hz. The experimental paradigm was implemented with OpenSesame (Mathôt et al., 2012) and PsychoPy (Peirce, 2007). Presentation was time-synchronized with the monitor's vertical retrace signal. When the PC detected a mismatch between the programmed SOAs and estimates of the realized SOAs (based on internal clock time stamps and the monitor synchronization) which was larger than 1 ms, the trial was discarded and repeated later. Such repetition occurred on less than

0.17 % of the trials[1].

The browser condition was conducted on mobile devices the students brought to the lab, or, if they had no working mobile device available, in the browser of a PC (Ubuntu, Firefox 70.0, 22" TFT-display, Acer V223W Ab, with 1680×1050 px resolution). It was programmed using the JavaScript library jsPsych (de Leeuw, 2015). The scripts were transpiled with Babel.js, bundled with WebPack.js, and served by JATOS (Lange et al., 2015). Pre-loaded images were used as stimuli.

The same SOAs were used in both conditions. These SOAs were divisible by 10 ms which allowed accurate presentation on the 100 Hz lab monitor. However, it was expected that most mobile devices in the browser condition will not be able to display these SOAs accurately, fostering rather uncontrolled timing.

The viewing distance on the lab computer was 50 cm, for the browser condition it was not fixed. On the lab computers and participants' laptops, the `Q` key and the `P` key were used for collecting the order judgments. On mobile devices, participants had to touch the corresponding side of the screen. The lab condition took place in a dimly lit experimental booth, the browser condition under variable daylight conditions in an office room.

4.1.3 Stimuli

An array of 16×8 bars with a fixation point in its center was shown (see Figure 3). On the lab computer, the array encompassed 41° × 21° of visual angle. Length and width of the bars were 1.37° and 0.32° plus a small jitter of 0.32° and 0.12°. The background color was a light gray (#c1c1c1), the bars were colored dark gray (#7f7f7f), and the fixation mark was black. In the browser condition, the layout was the same but sizes varied with the uncontrolled display size and viewing distances.

Two positions, one on the left and one on the right, contained reference and a probe stimuli. Both were darker than the other stimuli. Depending on the condition (salient vs. nonsalient), the probe could be salient in its orientation (90° difference from the background elements). The two positions were randomly chosen among the inner positions in each hemifield. (The outer columns close to the screen border and the fixation mark were excluded). Background orientation was chosen randomly in each trial (equal for all background bars on a display half) and differed by 90° between the left and the right half of the display.

After the display had been presented for 300 ms plus a jitter (450 ms), probe and reference flickered briefly by offsetting and onsetting again after 30 ms. The flickers were separated by SOAs of ±100 ms, ±70 ms, ±50 ms, ±30 ms, ±10 ms, or 0 ms. SOAs were repeated 20 times each. A video of the experiment can be found at https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3.

[1]Except for one participant who did not finish the experiment because trials were continuously repeated due to a monitor configuration problem.

4.1.4 Procedure

Participants performed in four sessions, two of the lab conditions and two of the browser conditions, in random order. Throughout each trial, participants were instructed to fixate a point in the center of the display that was visible from the beginning of each trial. There was no fixation control.

Observers judged which of the two flicker events appeared earlier, the event on the left or the event on the right. Responses were given with the `Q` key and the `P` key on the computer and by touching the corresponding half of the display on mobile devices. The next trial started automatically. A short training of 10 trials allowed familiarization with the task. The main parts contained 440 trials and a break was offered every 44 trials. The experiment lasted approximately 25 min.

4.2 Results and Discussion

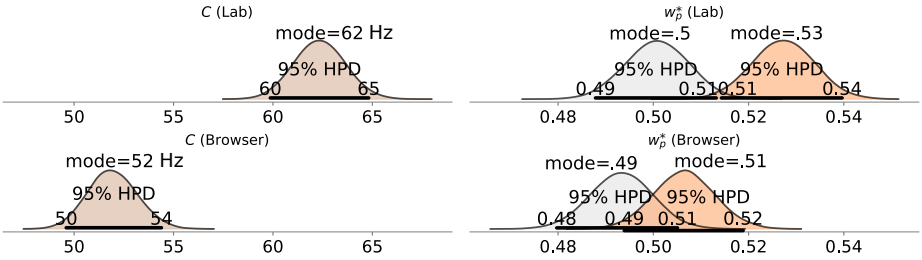

Figure 4: Results of Experiment 1: Means for overall processing capacity, C, and attentional weight of salient and nonsalient probe, w_p^*, in the lab condition, top, and the browser condition, bottom.

We first report the C and w values and their difference between conditions. Afterwards, we look into correlations for each value between the lab and the browser condition. Finally, we look at effect sizes of the salience effect in the two conditions and compare the model fits.

We found a C difference between the lab and browser in overall capacity (see Figure 4). In the lab condition, the mean overall processing rate C across all participants was estimated at 62.32 Hz [95 % HPD: 59.86, 64.82] with an SD of 21.76 [95 % HPD: 19.26, 24.44]. In the browser condition, the mean overall processing rate C across all participants was estimated at 51.91 Hz [95 % HPD: 49.60, 54.39] with an SD of 32.26 [95 % HPD: 28.16, 36.53]. The difference in C between the conditions is 10.41 Hz [95 % HPD: 6.92, 13.83].

The overlap of the w_p posterior distributions is relatively large. In the lab condition, the mean attentional weight of the nonsalient probe across all participants was estimated at .50 [95 % HPD: .49, .51] with an SD of 0.05 [95 % HPD: 0.04, 0.06]. The mean attentional weight of the salient probe was .53 [95 % HPD: .51, .54] with an SD of 0.06 [95 % HPD: 0.05, 0.07]. This probe weight of the salient target is .03 higher than that of the neutral one [95 % HPD: .01, .04].

In the browser condition, the mean attentional weight of the nonsalient probe across all participants was estimated at .49 [95 % HPD: .48, .51] with an SD of 0.06 [95 % HPD: 0.04, 0.07] and the mean attentional weight of the salient probe at .51 [95 % HPD: .49, .52] with an SD of 0.06 [95 % HPD: 0.05, 0.07]. The probe weight of the attended stimulus is .01 higher than that of the neutral one [95 % HPD: .00, .03]. Comparing the attentional weight of the salient probe between lab and browser, the difference is estimated at .02 [95 % HPD: .00, .04].

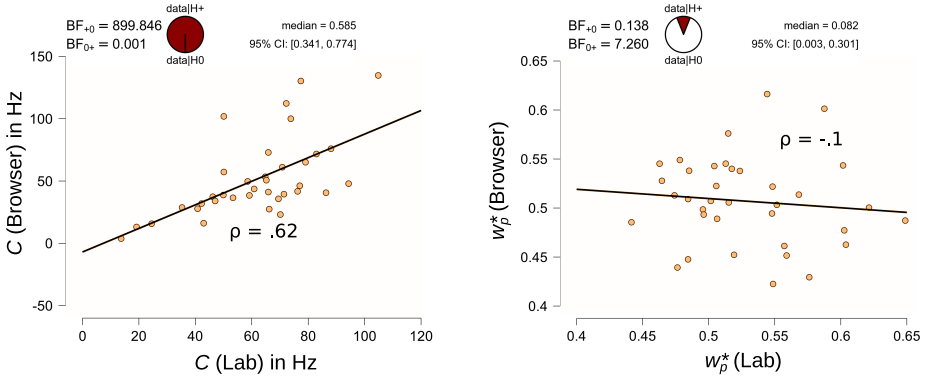

Figure 5: Results of Experiment 1: The C estimates of the lab and browser session are positively correlated while no such correlation is evident for the w_p^* estimates (salient probes only); BF = Bayes Factor.

TVA's overall processing capacity C shows a strong positive correlation between the lab and the browser condition supported by a very high Bayes factor (see Figure 5, left panel; posterior modes of the participant-level estimates enter the analysis and are shown in the scatter plot). The Bayes factor (calculated with JASP JASP Team, 2019) quantifies how much more it is probable that the data is positively correlated than that it has no positive correlation. Concerning w_p^* of the salience condition, there is no evidence in favor of a correlation (Bayes factor below one; see Figure 5, right panel). Hence, although different in magnitude, C remains an index of the participants' processing capacity even under the rather uncontrolled presentation conditions. The lack of a correlation in w_p^* is most likely due to the small size of the effect in the browser condition (cf. Figure 4, lower right distribution pair) which may not be able to overrule the random fluctuations introduced by the less controlled setting on the mobile devices.

To understand the difference between the same experiments in the lab and in the wild and how it may influence the possibility to find effects known from the lab, we calculated standardized effect sizes (Cohen's d) for the salience effect on w_p^* and the probability of superiority. The latter is calculated from Cohen's d and gives the probability that in a random individual we measure a higher w_p^* in the salient condition than in the neutral one (cf. Grisson & Kim, 2005). As Table 1 (first row) shows, effect size in the lab is larger than in the wild. The same is true for the probability of superiority. Still, the mobile condition

allows for a small effect and a .59 superiority which might suffice for a variety of questions and when large enough samples are available.

Experiment	ES w_p* (lab)	ES w_p* (wild)	PS w_p* (lab)	PS w_p* (wild)
1 (Orientation)	0.48 [0.16, 0.84]	0.23 [-0.09, 0.54]	.69 [.56, .80]	.59 [.47, .71]
2 (Red–green)	1.18 [0.85, 1.55]	1.19 [0.85, 1.59]	.89 [.81, .94]	.89 [.81, .95]
3 (Yellow–blue)	1.54 [1.18, 1.95]	1.05 [0.74, 1.40]	.95 [.89, .98]	.86 [.78, .93]

Table 1: Effect size and probability of superiority together with their boundaries of 95 % Highest Probability Density in the lab and wild conditions in Experiments 1 to 3.

The reader may be interested in how well the theoretically derived model fits the presumably more variable data recorded "in the wild" compared to the lab recordings. Thus we provide a comparison of the goodness of fit for Experiments 1–4 because they have a lab and "in the wild" condition. Note, first, that although we want our model to fit the data well, we also want it to be theoretically sound which is why theory-free optimization of model fit (e.g. by adding parameters) is no option for us. One important criterion for model fit is a visual check. Figure 19 shows that for most participants the model curves well describe the change of data points across SOAs. Another criterion is quantitative model fit. How to quantitatively describe and test model fits is a very difficult question. A widely used method to analyze the model fit are posterior predictive checks that allow for sampling of a posterior predictive p-value (Conn et al., 2018). Similar to classical p-values, the value states the probability of observing the present or more extreme data under the fitted model. A low probability reflects lack of fit. However, from a Bayesian perspective, the value is also a gradual indication of goodness of fit. Values around .5 indicate good model fit (see Berkhof et al., 2000, for details). The model check requires a discrepancy measure for which we use a χ^2 measure (c.f. Berkhof et al., 2000; Wichmann & Hill, 2001). We compare the p-values in the lab and "wild" conditions and separately for large and small SOAs, as more deviations can be expected at the smaller ones, which are prone to problems in the presentation timing.

The results for the present experiment (Table 2, first row) indicate that goodness of fit is strongly reduced in the browser condition, and it is also strongly reduced for the small SOAs. That the lowest values are observed at smaller SOAs is in line with findings that the central parts of TOJ psychometric functions reflect additional processes which are not included in the model we use for this study. An extensive assessment and discussion from a TVA perspective can be found in Tünnermann and Scharlau, 2018b. In the browser condition, additional deviations are evident (both in small p-values and the subject level plots in Figure 19) which reflect the additional noise from conducting the experiment "in the wild". While these deviations might be statistically strong, we believe that they do not interfere substantially with the estimation of the relevant parameters. As can be seen in Figure 19, central deviations have little impact on the overall shape of the fitted psychometric function.

Experiment	p-value Lab			p-value "Wild"		
SOAs	all	large	small	all	large	small
1 (Orientation)	.038	.75	$< .001$	.004	.018	$< .001$
2 (Red–green)	.548	.556	.481	.097	.249	.12
3 (Blue–yellow)	.235	.451	.111	.018	.166	.004
4 (Dragonfly)	.002	.003	.016	.047	.126	.051

Table 2: Posterior predictive check p-values of model fit for Experiments 1 to 4, separately for the lab and the browser condition as well as small (Exp. 1–3: $|SOA| <= 30$ ms, Exp. 4: $|SOA| <= 50$ ms) and large SOAs (remaining SOAs).

To compare the reliability of parameter estimates in lab and wild settings, we conducted split-half tests, which are reported in Appendix Section A.6. Concerning the C parameter, a $\rho = .8$ indicates a good reliability "in the wild". In fact, the score is higher than the $\rho = .62$ obtained for the lab condition. The correlations of the halves are supported by high Bayes factors (cf. Table 3, row 1). Concerning the w_p^* estimates for the salient condition, the picture is a different one. The estimated reliability is low ($\rho = .25$ in the lab and $\rho = .13$ in the wild) and these estimates and the correlation are rather uncertain (low Bayes factors, see Table 3, row 1). To anticipate the outcomes of these tests for the other experiments: the picture is similar. The C estimates are reliable, whereas the estimated reliability of the w_p^* is rather low (half of the estimates are below $\rho = .5$), see Tables 3–5. However, two things must be kept in mind: The reliability might be underestimated. The number of trials in these experiments was already low and is further divided as part of this analysis. Only half of the trials belong to the salience condition, and only half of those go into one sub-sample for the split-half test. This leaves only 110 trials (in this experiment and less in some others) to estimate the parameters of the individuals. Given the small w_p^* salience effect, especially in this experiment, the estimates might be too noisy to produce a better reliability score. In practice, this can be counteracted by including more trials (perhaps by replacing the neutral condition with more salience trials). The second thing to keep in mind is that the purpose of the present reliability estimation is to see whether or not the estimates obtained from experiments conducted "in the wild" are less reliable than those form lab experiments. Taken together, tables 3–5 show no indication of a generally reduced reliability under less controlled conditions.

To sum up the results of Experiment 1, we found a difference in w in favor of the salient target in both conditions, larger in the lab than the browser condition. This pattern goes along with a strong difference in overall processing capacity C. Mean C is substantially larger in the lab than in the browser condition. Although both C estimates are reasonable for healthy adult participants, the difference is in need of explanation. Most likely, the different presentation on smaller screens and less controlled lighting reduces the C "in the wild". Processing rate C is a stimulus-dependent measure anyway and researchers should be aware

that in uncontrolled conditions it can be expected to be lower. One might want to counter this by increasing its visual impact (e.g., larger stimulus size, higher contrast). Parameter C does however correlate highly between the two conditions, indicating that it essentially probes the same process. The w_p^* from the salient condition do not correlate between the lab and "wild" condition. This is somewhat unexpected. However, random noise in the individual estimates might overrule the rather small rather small salience effect.

Effect sizes and probability of superiority are larger in the lab than in the browser condition. Model fit is reduced in the browser condition, but the model is probably still useful. Before too strong conclusions are drawn, we want to test the reliability of these findings. This is the purpose of Experiments 2 and 3.

5 Experiment 2

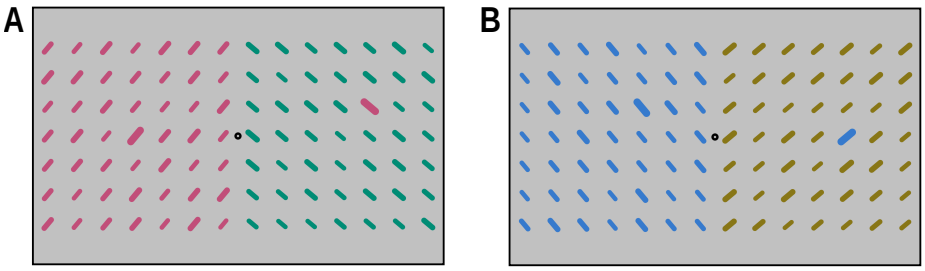

Figure 6: Exemplary displays with color salience used in Experiments 2 and 3. The slightly larger line segments constitute the probe and reference elements that flicker (separated by the stimulus onset asynchrony) to produce the temporal-order judgment stimulation.

In Experiment 1, the TOJ paradigm and TVA-based modeling worked reasonably well with experiments run in a web browser on varying and unselected (mobile) devices. However, C estimates (and to a lesser degree in terms of distribution overlap w_p^* estimates) seem to be biased toward smaller values. The attention effect on the probe attentional weight w_p^* was only just detectable. Experiments 2 and 3 test the reliability of this pattern and expand the comparison to the feature dimension color which might be affected more strongly by relaxing control than stimulus orientation.

Experiment 2 considered color salience of two colors with complementary a values in the CIELAB color space ($L = 50, a \in \{50, -50\}, b = 0$). Note that a regular screen was used without color calibration and hence the differences in LAB values only approximately adhere to the uniformity of the CIELAB color space. However, given that colors are diametrically apart on the chromacity axes, a strong hue difference that should lead to a relative salience effect apparent in TVA's attentional parameters is guaranteed. Except for the salience feature, the experimental procedure is the same in as Experiment 1; minor modifications are

described in the following.

5.1 Method

5.1.1 Participants

Thirty persons (9 male and 21 female; $M_{\text{age}} = 23.41$, range 19–45) participated. All participants were students or members of Paderborn University. Each participant gave informed written consent, reported normal or corrected-to-normal visual acuity and no color vision deficits and received course credit. All completed two sessions in each of two conditions, except two participants, who were only available for one session in the browser–mobile condition.

5.1.2 Apparatus

This time, both conditions used the browser-based implementation. In the browser–PC condition, the experiment was run on the lab PC with CRT monitor from Experiment 1. In the browser–mobile condition, again the experiment was conducted on the (mobile) devices participants brought to the lab.

5.1.3 Stimuli

Stimuli were the same as in the preceding experiment, except for the following differences: The probe's orientation did not differ from the surrounding background bars' orientations. All bars had the same randomly chosen orientation and differed by 90 degrees between the left and the right side. The two targets were slightly larger than the background elements in order to make them easily distinguishable as targets (see Figure 6. Stimuli were either red (LAB: $L = 50, a = 50, b = 0$; RGB: #c14e79) or green (LAB: $L = 50, a = -50, b = 0$; RGB: #008c75), and the probe target could be salient by having the alternative color (red among green or green among red). The reference target always had the same color as the background elements. A video of the experiment can be found at https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3.

5.1.4 Procedure

The procedure was the same as in Experiment 1. The experiment lasted approximately 25 min.

5.2 Results and Discussion

In the browser–PC condition the mean overall processing rate C across all participants was estimated at 57.81 Hz [95 % HPD: 55.07, 60.45] with an SD of 28.81 [95 % HPD: 25.29, 32.59]. In the browser–mobile condition the mean overall processing rate C across all participants was calculated as 34.48 Hz [95 % HPD: 32.90, 36.11] with an SD of 16.02 [95 % HPD: 14.23, 17.71]. The difference in C is 23.33 Hz [95 % HPD: 20.22, 26.48]. In the browser–PC condition the neutral

mean attentional weight of the probe across all participants was estimated at .50 [95 % HPD: .49, .51] with an SD of 0.05 [95 % HPD: 0.04, 0.06], and that of the salient probe at .58 [95 % HPD: .56, .59] with an SD of 0.07 [95 % HPD: 0.06, 0.09]. The estimated weight of the salient stimulus is .08 higher than that of the neutral one [95 % HPD: .06, .09]. In the browser–mobile condition the neutral mean attentional weight of the probe across all participants was again estimated at .50 [95 % HPD: .48, .51] with an SD of 0.05 [95 % HPD: 0.04, 0.06], that of the salient probe at .57 [95 % HPD: .55, .58] with an SD of 0.06 [95 % HPD: 0.05, 0.08]. The estimated weight of the salient stimulus is .07 higher than that of the neutral one [95 % HPD: .05, .09]. Comparing the attentional weight of the salient probe between the PC and mobile conditions, the difference is estimated at .01 [95 % HPD: -.01, .03].

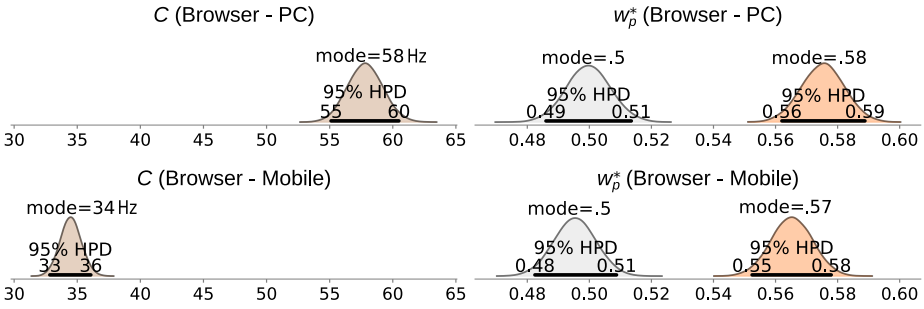

Figure 7: Results of Experiment 2: Means for the overall processing capacity, C, and attentional weight of salient and nonsalient probe stimulus, w_p^*, in the browser–PC condition, top, and the browser–mobile condition, bottom.

The overall result pattern in Experiment 2 is very similar to Experiment 1. We found an increased attentional weight of a salient stimulus. This is in accordance with earlier studies (Krüger et al., 2016, 2017). The increase was very similar for the conditions run on the PC and on an unselected mobile device.

Worth noting is the difference in the parameter C. Again, C is much smaller when doing the experiment on a mobile device, and the mean of 34 Hz is very low.

In contrast to Experiment 1, not only the C estimates but also the w_p^* estimates (of the salience condition) from the browser-PC and browser-mobile condition show strong positive correlations (see Figure 8). Hence, it seems that if strong salience effect can be established "in the wild", the pattern of individual differences found in the lab can be largely reproduced.

Effect sizes are very large, and so is probability of superiority (see Table 1, second row)). What is more—and different from Experiment 1—both effect size and probability of superiority are the same for the lab and the mobile condition.

The model fit assessment (Table 2, second row) indicates that goodness of fit is reduced in the browser–mobile ("Wild") condition, compared to the lab condition. It is not as low as in Experiment 1. In fact, the browser–mobile condition seems to have a better fit than the lab condition of Experiment 1.

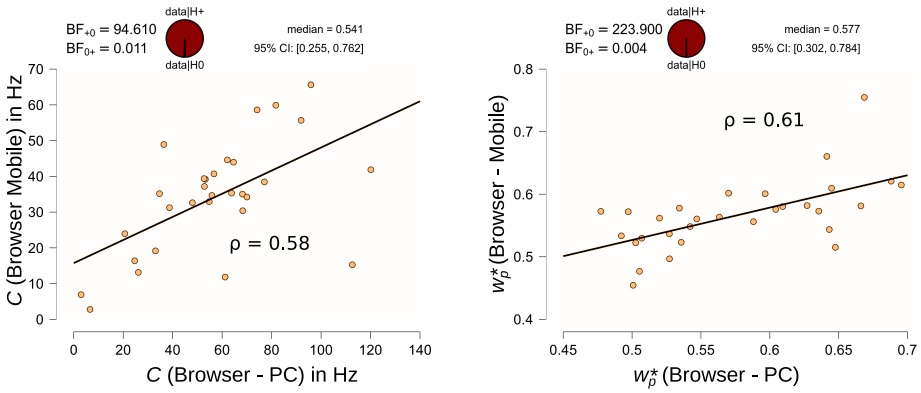

Figure 8: Results of Experiment 2: As in Experiment 1, in Experiment 2 the C estimates of the browser–PC and browser–mobile session are positively correlated; the w_p^* values (salient probes only) show a similar correlation in this experiment; BF = Bayes Factor.

We also point to Appendix Figure 20 for plots individual data and fits and the split-half reliability tests, which are reported in Appendix Section A.6.

Because the variety of mobile device participants bring to the lab and to further vary the salient feature, we run another version of the experiment with other colors as Experiment 3.

6 Experiment 3

Experiment 3 is a replication of Experiment 2 with different color values (yellow and blue; LAB: $L = 50, a = 0, b \in \{50, -50\}$). As everything except for the color axis in the CIELAB color space was the same, we expect the same results as in Experiment 2.

6.1 Method

6.1.1 Participants

Thirty-two persons (2 male and 30 female; $M_{\text{age}} = 22.31$, range 14–35) participated. Except for one person, all participants were students or members of Paderborn University, gave informed written consent, reported normal or corrected-to-normal visual acuity and no color vision deficits and, if students, received course credit. All performed two sessions per condition, except one participant who only completed one of the browser–PC sessions and another person who completed only one of the browser–mobile sessions.

6.1.2 Apparatus

The apparatus was the same as in the preceding experiment.

6.1.3 Stimuli

Stimuli were the same as in the preceding experiment, except for the colors which were determined setting $a = 0$ and $b \in \{50, -50\}$ in the CIELAB color space. The exact color values were #887616 (yellow; LAB: $L = 50, a = 0, b = 50$) and #367ACD (blue; LAB: $L = 50, a = 0, b = -50$). A video of the experiment can be found at https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3.

6.1.4 Procedure

Procedure was the same as in Experiment 2, and the experiment again lasted approximately 25 min.

6.1.5 Results and Discussion

In the browser–PC condition, the mean overall processing rate C across all participants was estimated at 61.31 Hz [95 % HPD: 58.29, 64.24] with an SD of 34.15 [95 % HPD: 28.44, 40.53]. In the browser–mobile condition the mean overall processing rate C across all participants was estimated at 45.27 Hz [95 % HPD: 47.49, 43.09] with an SD of 27.69 [95 % HPD: 24.27, 31.33]. The difference in C between these conditions is 16.04 Hz [95 % HPD: 12.38, 19.81].

In the browser–PC part, the neutral mean attentional weight of the probe across all participants was estimated at .51 [95 % HPD: .49, .52] with an SD of 0.06 [95 % HPD: 0.04, 0.07], the salient weight as .60 [95 % HPD: .59, .61] with an SD of 0.06 [95 % HPD: 0.05, 0.08]. The probe weight of the salient stimulus is .09 higher than that of the neutral one [95 % HPD: .08, .11].

In the browser–mobile condition, the neutral mean attentional weight of the probe across all participants was calculated as .50 [95 % HPD: .49, .51] with an SD of 0.05 [95 % HPD: 0.04, 0.06], the salient one as .57 [95 % HPD: .55, .58] with an SD of 0.07 [95 % HPD: 0.06, 0.08], with a difference of .07 [95 % HPD: .05, .08]. Comparing the attentional weights of the salient stimulus between the two browser conditions, the difference is estimated at .03 [95 % HPD: .02, .05]. For the split-half reliability tests, see Appendix Section A.6.

The correlation of parameters estimated in the browser–PC condition and those from the browser–mobile condition show the same pattern. While C strongly correlates, having a high Bayes factor, the w_p^* correlation is weaker and much less certain (see Figure 10). The weak correlation and reduced certainty reflected in the Bayes factor is unexpected. Experiment 3 was identical to Experiment 2, except for the color contrast used to establish salience. We further looked into this by conducting sequence analyses that show how evidence in favor or against the correlation accumulates when participants are iteratively added into the analysis (see Appendix Figure 26). As Appendix Figure 26C reveals, several participants do not change the evidence level much. Nevertheless, it

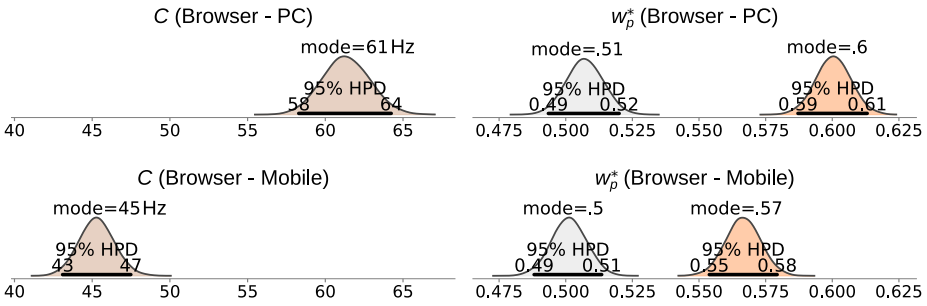

Figure 9: Results of Experiment 3: Means for processing capacity, C, and attentional weight of salient and nonsalient probe stimulus, w_p^*, in the browser–PC, top, and browser–mobile, bottom, conditions.

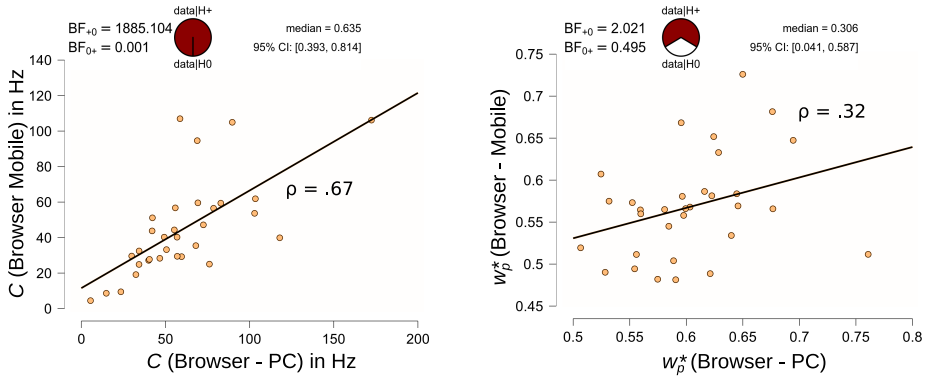

Figure 10: Results of Experiment 3: As in Experiments 1 and 2, in Experiment 3 the C estimates of the browser–PC and browser–mobile session are positively correlated; the w_p^* values (salient probes only) show a somewhat weaker correlation; BF = Bayes Factor.

accumulates towards strong evidence until participant 29 is added. Participant 29 has an exceptionally high w_p^* in the browser–PC condition (.76) and a very low w_p^* in the browser–mobile condition (.51). A possible explanation is that rendering one of the colors (perhaps the yellow) depended strongly on the display properties. Since no display (neither lab nor mobile) was color calibrated, it is well possible that the colors rendered substantially different on different devices. If this was the case, it is noteworthy that the red–green contrasts used in the previous experiment appears to be particularly robust (cf. the correlation in Figure 10 and the corresponding sequence analysis in 26B).

As in Experiment 2, effect sizes (see 1 are very large, and so is probability of superiority, though they are somewhat less impressive in the wild than in the lab.

The result of the model fit (Table 2, third row) indicates that goodness of fit

is reduced in the browser–mobile condition, especially for the small SOAs. Small SOAs also show a deviation in the lab condition. Individual data and fit plots are provided in Figure 21.

Overall, Experiment 3 confirms the pattern of results in Experiment 2. Firstly, salience results in a substantial weight increase. The increase was smaller on an unselected mobile device, but still in the range of expected values. Again, C is strongly reduced when doing the experiment on a mobile device, and the mean of 45 Hz is comparably low. Effects sizes and probability of superiority are large.

7 Some comments on Experiments 1 to 3

We now turn to potential reasons for the substantial reduction of C in the conditions with unselected (mobile) devices brought by the participants in Experiments 1 to 3. In general, the discrepancy could have two origins: Participants could indeed have a reduced C available for the task on a mobile device. They might spend more of their processing capacity on their surrounding (which is less distracting in a typical lab setup) and stimuli occupy smaller parts of the visual field. This is in line with a reported reduction of C if the environment is monitored for a change additionally to a concurrent task (Poth et al., 2014). However, the second possible origin seems more likely: Many devices might not be able to present the stimuli with sufficient precision. Especially the lower frame rate of most devices could be crucial. If at short SOAs, frames are dropped they can even be rendered as SOA zero (both targets start flickering in the same frame). This effectively shortens the SOAs which lowers the precision of the order judgment and ultimately C. On the other hand, lags in computation could lengthen SOAs depending on the exact methods of time keeping in the devices. While the participant-level plots of conditions conducted on mobile devices generally agree with the typical data pattern and show good fits, some indications of timing problems at the small SOAs can be seen. For instance, among others, participant 23 of Experiment 1 shows data points clustering near .5 at smaller SOAs. At larger SOAs the points follow the expected pattern (see Figure 19).

Figure 11 shows a comparison of the discrepancies between lab PC and unselected devices. For this purpose, the participant-level differences from Experiments 1 to 3 in the estimates of the two conditions were grouped by the device configuration that was used in the mobile conditions. After pooling over different version numbers, 16 configurations remained. Some contain only one or a few participants but others contained up to 27 (for the most popular configuration "Phone iOS Safari"). The varying degrees of certainty in the estimates are captured in the varying width of the distributions.

One insight of this analysis is that, different from what the results in Figure 9 suggest, the C estimates measured with different devices vary, and some are smaller and others larger than the C estimates from the lab condition. While C estimates are smaller on smartphones and laptops, they are increased in Windows and Ubuntu laptops. The w_p^* parameters from the lab and browser condition are

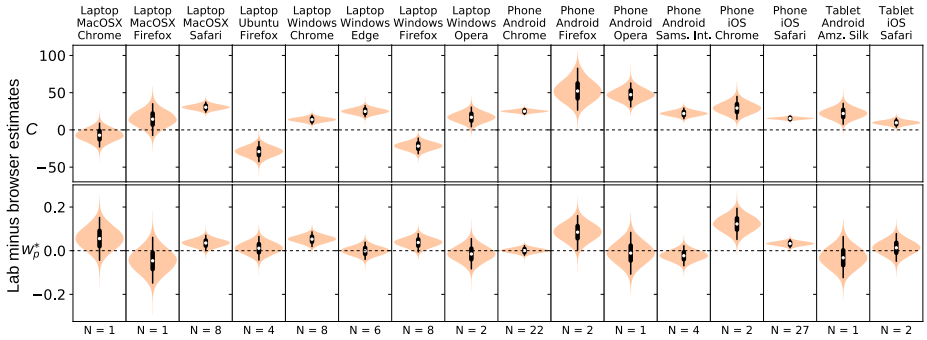

Figure 11: Comparison of the differences in the C and w_p^* estimates for different devices (operating systems and browsers).

often similar. If they differ, the w_p^* are smaller in the browser condition. While it is out of scope of the present study to investigate the technical reasons for the implementation behaving differently on the different systems, this analysis provides a starting point for researchers who want to run similar (timing critical) experiments on mobile devices. The mismatches not only provide hints on where potential problems originate from (e.g. larger C: realized SOAs are too long; smaller C: SOAs too short) but also an overview about which system configurations can be expected "in the wild".

To sum up Experiments 1 to 3, using mobile devices with experiments implemented in the browser can be a useful tool. In the context of TVA-based TOJ analysis, the overall processing rate C seems to be reduced, but measurements in the wild correlate with those from the lab. Effects on the attentional weights w_p^* can also be detected. However, depending on the salience manipulation, they were much smaller (Experiment 1 with orientation salience), somewhat smaller with yellow–blue color salience (Experiment 3) or the same as in the lab (Experiment 2 with red–green color salience). The degree to which the lab and "wild" estimates of individuals correlated followed the same pattern, with the strongest correlation for the red–green color salience. In many contexts, smaller effects and weaker correlations might be acceptable. However, researchers who set out for especially precise measurements should optimize experiment implementations for the expected target devices (e.g., to guarantee reliable timing) and—if their research question allows—turn to reliable attention manipulations, such as the red–green color contrast.

While, overall, the three experiments show that there are differences in how well effects can be measured in the wild compared to typical lab conditions, the experiments also show a reassuring and sometimes even very high precision of measurement in the wild. The assessment of model fit showed that data recorded outside the lab deviates more from the predictions of the fitted model than data from the lab. In both lab and "wild" recordings the fit is worse at smaller SOAs. The reason for this is likely two-fold: central regions of TOJ

psychometric functions might reflect processes not included in the current model (Tünnermann & Scharlau, 2018b, cf.). Moreover, "in the wild" such SOAs are more severely affected by timing problems that can be expected on many devices. However, visual assessments of the model fit (Appendix Figure A.1–A.3) indicate that even though these deviations are statistically detected, they do not seem to interfere with with the general quality of the fit and appear not relevant for the practical estimation of TVA parameters. (From a theoretical point of view, they indicate where modelers could start to improve models in future work).

8 Experiment 4

The first three experiments demonstrated that data collection outside the lab can reproduce the typical result patterns and the expected differences between experimental and control condition. In the fourth experiment we embed the TOJ task in a game environment. Computer games keep players interested even though the actual task can be very repetitive. The present experiment frames a salience experiment as a game. Results are compared to a control experiment in which the game elements have been removed. The dynamic movements of the game stimuli and their changing retinal size and position introduce additional degrees of freedom which are typically excluded in lab TOJs. While this can be seen as further loss of control over the stimulation, the game-like character of the experiment could help to keep the observer's motivation high. As explained earlier, it is not the scope of the present study to disentangle such influences. However, we would like to point out that influences can be both disadvantageous and beneficial and hence we do not necessarily expect a degradation in data quality when moving towards more game-like experiments.

The game was established as a race in which the participant, or player, controlled a dragonfly that moved forward with speed depending on the judgments. Participants used a computer mouse to steer the dragonfly through a tunnel. The dragonfly's speed was controlled by the game. Our focus was on keeping participants engaged throughout the experiment. To this end, their primary task was to reach the end of the tunnel as quickly as possible and before another computer-controlled dragonfly.

In regular intervals, players encountered walls in the tunnel with a grid of holes representing the stimuli. For the TOJ task, two slightly larger holes (a nonsalient reference and a salient probe) flickered. If the player flew through the hole that flickered second, they received a short boost, giving them an advantage over the other dragonfly. If a hole was missed, the dragonfly crashed into and thus shattered a piece of the wall, resulting in a loss of momentum. The player's actions, e.g., lateral motion or correct TOJ were accompanied by game-typical sounds. The sound of moving wings reacting to the dragonfly's lateral acceleration, as well as chimes sounding when a player flew through a target hole, or the sound of crumbling bricks when they crashed into a wall, all serve the purpose to enhance immersion and to give the player immediate feedback for their actions. In later levels, players were increasingly challenged

by stronger wind gusts coming from random horizontal and vertical directions, which they had to counteract.

The control experiment showed the same walls and required the judgment without the dragonfly race and without the sounds. Because probe and reference never occupied the same quadrant of the display, participants of the control experiment indicated the quadrant containing the stimulus that flickered second by a keypress (the four sectors were mapped on numpad keys).

Our hypotheses were that the overall visual processing capacity for the game would be reduced in comparison to the standard experiment, but still in the range of typical visual capacity of healthy adults (20 Hz–60 Hz, see, e.g., Wiegand et al., 2014). We furthermore expected an effect of salience on the attentional weights in both conditions with no specific expectations as to its exact value (a stronger attentional weight in the game would, however, be odd in so far as salience experiments are particularly designed so that the salience manipulations should exhibit maximal effectiveness).

8.1 Method

8.1.1 Participants

There were 31 participants (15 male and 16 female; $M_{\text{age}} = 22.41$, range 18–35) in the game and 31 participants (18 male and 13 female; $M_{\text{age}} = 23.09$, range 18–35) in the control experiment. All participants gave informed written consent, completed one session, reported normal or corrected-to-normal visual acuity and received course credit or a payment of 8 Euro per hour.

8.1.2 Apparatus

The experiment was conducted on a Microsoft Windows 10 PC with dedicated graphics card and a Master Pro512 22 inches (40.4 cm × 30.3 cm) CRT monitor. A resolution of 1024 × 768 pixels was used with 32-bit colors and a refresh rate of 100 Hz. The viewing distance was 50 cm. Input devices were an optical mouse (game) and a keyboard (control experiment). The game was implemented using the Unreal game engine, the control experiment was implemented with PsychoPy (Peirce, 2007).

8.1.3 Stimuli and Design

In this experiment, gaming was realized as a between-participants factor, salience as a within-participants factor.

Each wall consisted of a circular grid of stimuli (see Figure 12 for screenshots) with 9 holes in the horizontal and vertical direction. (The outermost rows and columns had 5 holes, those next to them 7 holes in order to fit them into the circular tunnel.) Two holes in each wall were marked as targets by being slightly but visibly larger than the others. In the experimental condition, the hole for the probe stimulus was rotated by 90° in relation to the other stimuli. The SOA values were 0 ms, ±50 ms, ±100 ms, and ±150 ms. For each trial, the probe

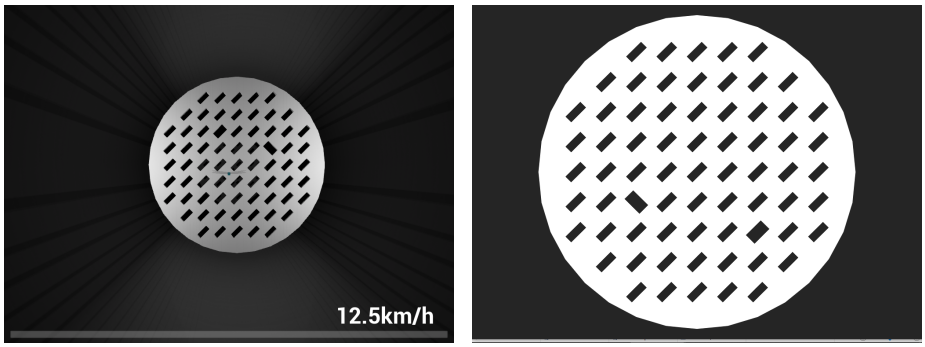

Figure 12: Example stimulus display of Experiment 4. Game on the left (the camera followed the small dragonfly in the center), control experiment on the right. The control experiment had a centered fixation mark preceding the display.

and the reference stimulus appeared in random but different quadrants of the grid. Each participant completed 50 training trials and afterwards another 10 levels with 50 trials each. A video of the experiment can be found at https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3.

8.1.4 Procedure

The procedure of the TOJ is sketched in Figure 1. In order to ensure that a TOJ was made and participants did not simply choose the first flicker, they had to choose the hole that flickered second. In the game, the judgment had to be made by steering the dragonfly through the hole that flickered second. There was no fixation mark. Participants did not receive any instructions regarding eye movements so that they were free to solve the task or play the game as they see fit. Like in the description of Experiment 1, it is worth noting that moving the eyes while the jittered flicker occurs likely results in not seeing the transitory change. Still, we cannot provide evidence for or against a possible influence of eye movements on the results. In the control experiment, participants indicated their judgment by choosing a quadrant of the display by a key press.

8.2 Results and Discussion

The game allowed to steer the dragonfly so that it could pass the wall at neither the reference nor the probe position. These trials were not evaluated because they do not correspond to a clear judgment. This was the case in less than 1.5 % of the trials.

In the game, the mean overall processing rate C across all participants was estimated at 55.49 Hz [95 % HPD: 52.97, 58.04] with an SD of 26.15 [95 % HPD: 23.14, 29.04]. The mean attentional probe w_p^* across all participants was estimated at .53 [95 % HPD: .52, .54] with an SD of 0.05 [95 % HPD: 0.05, 0.06].

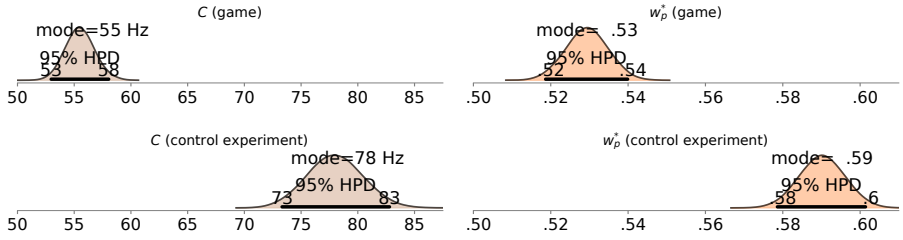

Figure 13: Results of Experiment 4: Game and control experiment means for processing capacity, C, and attentional weight of salient probe stimulus, w_p^*.

In the control experiment, the mean overall processing rate C across all participants was estimated at 77.97 Hz [95 % HPD: 73.32, 82.79] with an SD of 44.35 [95 % HPD: 36.42, 52.69]. The mean attentional probe w_p^* across all participants was estimated at .59 [95 % HPD: .58, .60] with an SD of 0.06 [95 % HPD: 0.05, 0.07].

Figure 13 shows the posterior estimates of both parameters for the game and the control experiment. In the game, the attentional weight is reduced. Also, the visual processing speed is diminished in comparison to the control experiment. Individual fits are shown in the Appendix Figure 22 for the game and Appendix Figure 23 for the control experiment.The split-half reliability tests are reported in Appendix Section A.6.

Both findings are in line with a TVA-based explanation. The game required a further task while observing the flicker. A detrimental effect of a second task on visual processing capacity has been shown in a TVA-based study before (Poth et al., 2014). This makes a reduced visual capacity plausible while an increase would have been a problem for a TVA-based explanation. Also, the reduced attentional weight can be explained in TVA by the necessity to observe the position of the dragonfly as well as both flicker locations.

The pattern of model fits shown in Table 2 (fourth row) deviates from Experiments 1–3 in a number of aspects: First, the p-values of the lab condition are lower than for the game condition, indicating that the model provides a better fit for the "in the wild" (game) condition. However, the goodness-of-fit measure is heavily affected by the fact that in this experiment, the SOAs were repeated many times (80 repetitions per SOA)—the high power per SOA requires highest precision of the fit and even slight deviations are penalized harshly (this seems akin to any significant test yielding significance if N is chosen high enough). Moreover, the steep psychometric functions lead to many observations close to the convergence levels where little variance is allowed. The small p-values might reflect that there are subtle influences on the course of the psychometric function which are not included in the present model. Taking the participant-level plots (Appendix Figure 23 and 22) into account, these fluctuations are clearly below what is relevant for the current study. However, again the model check can inform future modeling. For instance, lapse parameters (Tünnermann

& Scharlau, 2018a, cf.) could account for occasional "random" response errors, to deal with the overly strict requirement to meet the convergence axes.

For practical reasons, this experiment was conducted as a between-subjects design. While this might reduce statistical power, it does not introduce a systematic bias. Given the sufficiently large sample size (our earlier experiments of this study and those published elsewhere show that sufficiently precise parameter estimates can be obtained with 30 participants) and the presence of a salience effect we believe the outcomes are valid.

Summing up, assessing attention with TVA parameters works comparably well in the present game-like situation. The quantitative deviation is in a direction that is consistent with a TVA-based explanation of the additional attentional requirements during the game. Turning the logic of comparison upside-down—comparing not to justify the means but comparing to quantify the difference of the attentional requirements—the TVA-based approach potentially allows a quantitative estimate of the increase in attentional requirements between game and control experiment. The following experiment is designed to corroborate these findings with stronger gaming features.

9 Experiment 5

Experiment 5 used the same logic as Experiment 4. Given that Experiment 4 confirmed that the typical result pattern can be found outside the the restricted lab paradigm in game-like scenarios, we did not run a typical lab version of the task as a control. Instead we used the available resources to add further gaming features and explore the influence of varied task difficulty by varying a gaming factor (speed) and experimental timing (the SOA). Participants steered a spaceship through a tunnel. Again, there were grids in this tunnel with holes, two of which were marked as targets by their size and flickered as the spaceship approached the grid (see Figure 14 and the video available at https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3). The participants performed TOJs by choosing the hole that flickered second and steering their spaceship through it. In one condition, flying difficulty was increased when the participant performed well and decreased when they made mistakes. This factor, adaptive flying difficulty, was crossed with another difficulty factor, SOA timing, which we studied exploratively for reasons explained in the next paragraphs.

In an unpublished experiment with a similar game (Briese, 2019), the C values we estimated were very large. This finding is remarkable since in such dynamical situations, we cannot assess overall C because the participants have to distribute some of their resources to other parts of the game such as steering or monitoring their scores. So if equal capacity is assumed between game and experiment, a smaller C would have been plausible because a part of the fixed C would be applied to the game elements.

Two possible (post-hoc) explanations may account for the large Cs. Eleven of the twelve participants were experienced gamers. They might have acquired processing routines that enlarge C (or have had large visual processing capacity in

the beginning and therefore turned into heavy gamers). The other explanation—which is more interesting in the context of the present paper—is that the experiment had unusually small SOAs to compensate for the experience of the gamers. The increased difficulty that goes along with small SOAs might have pushed the players into strategic changes that showed up in a large C.

There are not many studies on what influences TVA's C parameter (exceptions are clinical and psychopharmacological studies, see, e.g., Habekost, 2015; Vangkilde et al., 2011). Two gaming studies indicate the possibility that gaming experience may increase C in some conditions (Schubert et al., 2015; Wilms et al., 2013). Studies of temporal expectancy (Vangkilde et al., 2012) and alertness (Haupt et al., 2015; Matthias et al., 2010; Petersen et al., 2017) provide evidence that timing is an important factor. Performing a concurrent tasks while monitoring the environment for events reduces the visual processing capacity (Poth et al., 2014). A theoretical explanation how external factors may affect C is offered by the three components expectancy, subjective importance, and general level of alertness (Bundesen, Vangkilde, & Habekost, 2015).

Because temporal factors are likely to affect C, we compared two blocked conditions, one with typical SOAs and one with slightly smaller ones. Keep in mind, however, that we still only measure a part of C; as the experimental manipulations could also affect the distribution of C over subtasks, we should be circumspect with all conclusions. Note also that we are not primarily interested in the influence of gaming experience on attention but study it only as a possible reason for the high C values reported by Briese (2019).

9.1 Method

9.1.1 Participants

Thirty persons (18 male and 12 female; $M_{\text{age}} = 24.77$, range 20–54) participated. Twenty-four participants were students or members of Paderborn University. All participants gave informed written consent, completed one session, reported normal or corrected-to-normal visual acuity and received course credit or a payment of 8 Euro per hour. On average, the participants' reports on gaming result in a mean of 9.2 hours per week (range 0–42 hours, SD = 9.5 hours, median = 6 hours). Fifteen participants reported to prefer action games, 8 reported to prefer logic games, and the 7 who did not play computer games did not indicate a preference.

9.1.2 Apparatus

Apparatus was the same as in Experiment 4.

9.1.3 Stimuli and Design

Two experimental factors were varied blockwise. SOA size s ms was either large ($s \in \{0, \pm 50, \pm 100\}$) or small ($s \in \{0, \pm 30, \pm 70\}$). Gamification was varied by using adaptive spaceship speed v units/s $\in [1000, 5000]$ in the adaptive conditions

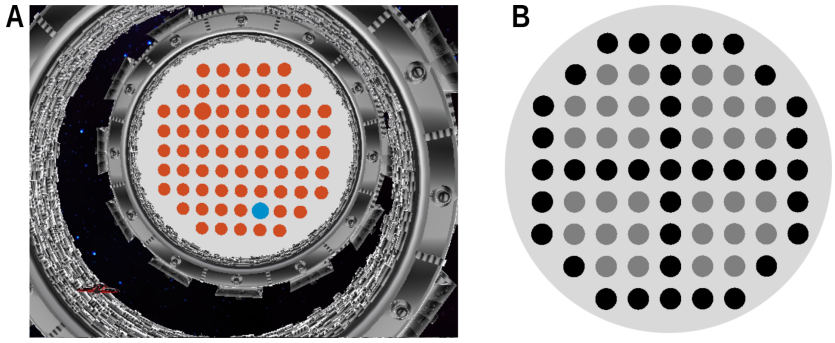

Figure 14: **A** Example of a stimulus display of Experiment 5. The spaceship is a small reddish object in the lower left. During the experiment its color is set to white. **B** illustrates the possible probe and reference locations in gray. Locations that always contained background elements are shown in black.

(v_{start} :=1000) and a constant speed v := 1000 units/s in the nonadaptive conditions. In each of the four blocks, the SOAs were repeated 50 times, which resulted in 250 trials per pass and 1000 in total. Block order was varied between participants.

Comparable to Experiment 4, each wall consisted of a circular grid of stimuli with 9 holes in the horizontal and vertical direction (except for the outer rows and columns, as in Experiment 4), see Figure 14**B**. The background elements were either red (RGB: #cf4b22) or blue (RGB: #008fcb) in a trial. The reference stimulus always had the same color as the background elements, whereas the probe always had the other color. Both targets were slightly larger than the background elements. Probe and reference were always in different quadrants; the allocation and exact position were randomly chosen among the allowed locations. An example is shown in Figure 14**A**. After the wall had been presented for 150 ms (±20 ms) plus a jitter (j ms $\in [10, 100]$), probe and reference flickered briefly by offsetting and onsetting again after 80 ms. Calm pieces of jazz and lounge were used as background music, also known as "elevator music."

In the nonadaptive speed condition, the spaceship flew at a constant speed of v := 1000 units/s. In the adaptive speed condition, spaceship speed was in the range of 1000–5000 units/s and depended on the performance of the participants. After each trial, speed could be increased by 20 or 45 units/s or decreased by 100 or 50 units/s, depending on aggregated performance. The distance between the walls was dynamically coupled to the speed of the spaceship so that adaptive speed affected only the time the participant had to make their judgment. A video of the experiment can be found at https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3.

9.1.4 Procedure

The task was explained to the participants in the main menu where the spaceship flew through the walls on its own, illustrating the game. An example wall was shown for one training trial. The experiment started without any additional training. Like in Experiment 4, participants did not receive any instructions regarding eye movements so that they were free to play the game as they see fit.

The judgment had to be made by steering the spaceship with the mouse through the appropriate hole. As in Experiment 4, the correct choice was the hole that flickered second. Audio feedback was given if the participant hit one of the target holes. One block took 6 to 8 minutes. The complete experiment lasted approximately 35 min.

9.2 Results and Discussion

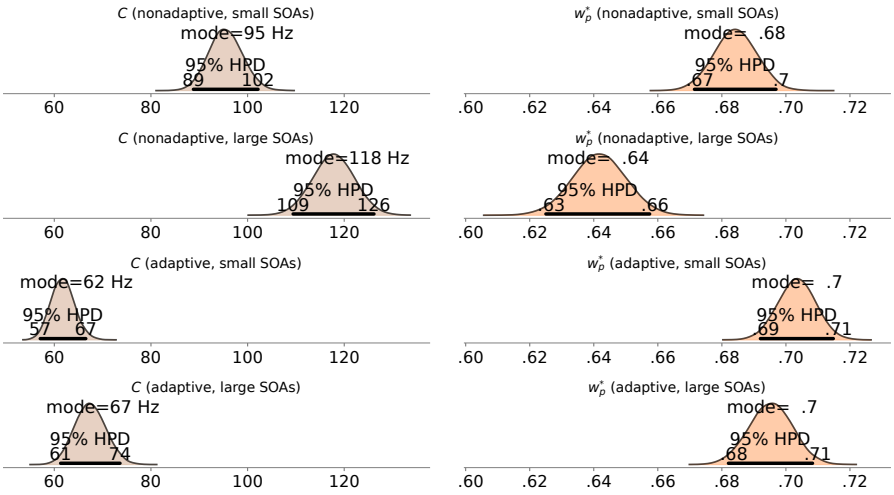

Figure 15: Results of Experiment 5: Means for processing capacity, C, and attentional weight of salient probe stimulus, w_p^* in the four conditions.

The C values in the adaptive conditions were smaller than those in the nonadaptive conditions (adaptive, large SOAs: 67.39 Hz [95 % HPD: 61.31, 73.60] with an SD of 43.67 [95 % HPD: 32.90, 54.51; adaptive, small SOAs: 61.70 Hz [95 % HPD: 57.08, 66.51] with an SD of 31.82 [95 % HPD: 25.12, 39.61; nonadaptive, large SOAs: 117.70 Hz [95 % HPD: 109.42, 126.16] with an SD of 65.17 [95 % HPD: 57.34, 73.13; nonadaptive, small SOAs: 95.40 Hz [95 % HPD: 88.79, 102.17] with an SD of 44.59 [95 % HPD: 36.80, 52.70; see Figure 15, left column). The difference between the adaptive and the nonadaptive conditions is 50.31 Hz [95 % HPD: 60.90, 40.20] between the large SOA conditions and 33.69 Hz [95 % HPD: 25.46, 41.79] between the small SOA conditions. Both differences are large and the HPDs exclude zero. This finding accords well

with the idea that in the adaptive condition, participants had to spend more of their capacity on strategic changes. The individual fits are shown in Appendix Figure 24 and the split-half reliability tests are reported in Appendix Section A.6.

The experimental factor SOA size had an influence in the nonadaptive condition with C being larger with larger SOAs, that is, less difficult trials. The difference is 22.30 Hz [95 % HPD: 33.24, 11.78] In the adaptive condition, the distributions of C overlap strongly and the difference is close to zero, -0.01 Hz [95 % HPD: -0.03, 0.01] (see Figure 15).

In all conditions, w_p^* deviates from a neutral distribution of .5 (adaptive, large SOAs: .70 [95 % HPD: .68, .71] with an SD of 0.09 [95 % HPD: 0.08, 0.11]; adaptive, small SOAs: .70 [95 % HPD: .69, .71] with an SD of 0.09 [95 % HPD: 0.08, 0.10]; nonadaptive, large SOAs: .64 [95 % HPD: .63, .66] with an SD of 0.11 [95 % HPD: 0.09, 0.12]; nonadaptive, small SOAs: .68 [95 % HPD: .67, .70] with an SD of 0.10 [95 % HPD: 0.08, 0.11]. This means, firstly, that the effect of salience is present in all attentional weights. Secondly, the weight values of .64 to .7 are comparably large. Similarly large weights have been reported by Krüger et al. (2016). Thirdly, w_p^* deviates stronger from .5 the less overall capacity there is. As the weight is a relative value, one explanation might be that with large C, there is less competition between the two targets. This idea is, however, speculative.

Cs in the nonadaptive condition were large, replicating the finding of Briese (2019), but many of our participants cannot be regarded as heavy gamers. To exploratively test for the influence of gaming experience, we calculated Pearson correlation coefficients between self-reported hours of gaming per week and the mean of the participant's C distribution for all four conditions with JASP (JASP Team, 2019). Interestingly, there was a positive correlation in both nonadaptive conditions (small SOAs: $\rho = .34$ [95 % HPD: .04, .6]; large SOAs: $\rho = .43$ [95 % HPD: .09, .66]). This was absent in the adaptive trials (small SOAs: $\rho =$ -.12 [95 % HPD: .24, -.44]; large SOAs: $\rho =$ -.02 [95 % HPD: .33, -.36]). Testing for a positive correlation revealed a Bayes factor of $BF_{+0} = 2.1$ and $BF_{+0} = 6.0$ for the small and large nonadaptive SOAs condition, respectively. The adaptive conditions with small SOAs ($BF_{+0} = 0.1$) and with large SOAs ($BF_{+0} = 0.2$) showed comparable evidence in the opposite direction (indicated by a Bayes factors smaller than 1) for the null hypothesis. These results point in the same direction as our explanation, but the Bayes factors around 4 show that the evidence is moderate ($BF \geq 3$) rather than strong ($BF \geq 10$; Lee & Wagenmakers, 2014). It seems that all persons were pushed towards lower C values by adaptive speed, but that this was more pronounced for the ones with higher gaming experience, see Figure 16. If this description is correct, we can conclude that the high overall C values are indeed partly driven by the gamers, but that these participants cannot utilize their advantage under typical (difficult) gaming conditions. Note however that the influence of gaming experience on attention parameters is beyond the scope of the present paper.

To sum up, Experiment 5 again showed that, within the TOJ-TVA paradigm, C and w can be assessed in gaming tasks that are less artificial to participants and possibly less tedious than the standard psychological experiments. The

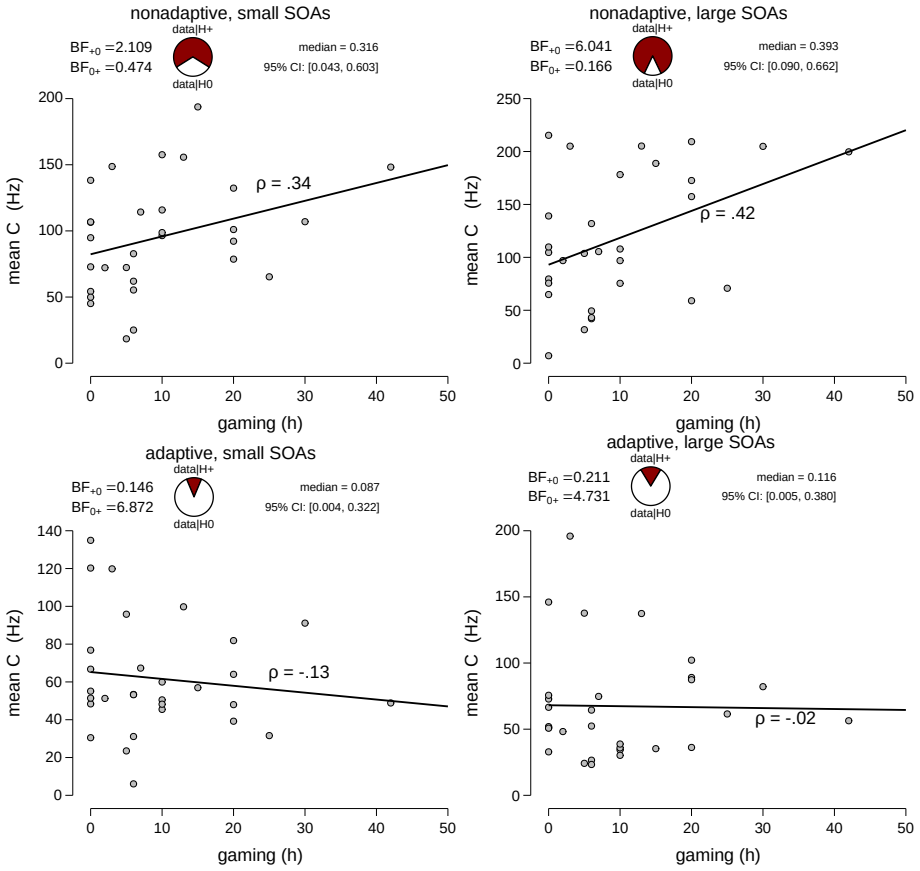

Figure 16: Results of Experiment 5: Scatter plots of self-reported hours of gaming per week and the mean of participants' C estimates per condition; BF = Bayes Factor.

experiment also shows that there might be problems in explaining the parameter values. One problem are the large C values for which there is no ready explanation and the apparently complementary pattern of C and w that is somewhat suspicious with respect to the assumed independence of C and w.

10 Experiment 6

Up to here, all experiments kept participants seated in an office chair in front of a PC or other device to perform the task with button presses or touch responses. In Experiment 6, we move one step further into the wild by putting participants in a physically more active situation: they had to sit on a bicycle and pedal through a virtual traffic scenario (cf. Heinovski et al., 2019, for technical details

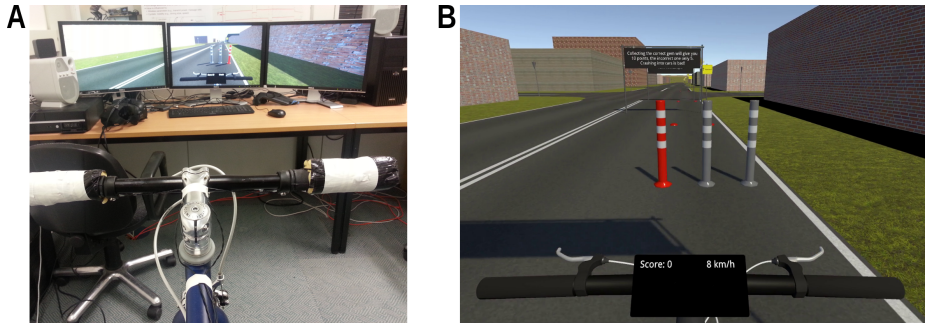

Figure 17: **A** Setup of the virtual cycling environment as it was used in Experiment 6. **B** Screenshot of one training trial. The probe stimulus (orange pylon) and the reference stimulus (one of the gray pylons) flickered and the cyclists had to collect the one they perceived as flickering second.

on the simulation). Recently, we reported TOJ experiments performed in this setting (Stratmann et al., 2019), which we now extend by introducing a salience manipulation similar to the other experiments reported in the present article.

Participants steered the bicyle through a simple, grid-shaped street scenario. The path they had to follow was indicated by street signs at the intersections. At the intersections, there was also car traffic, generating additional (and rather realistic) attentional load. The TOJ was integrated as follows: At regular intervals, hovering diamond-shaped objects appeared above the street and flickered. The participants had to steer through the diamond that flickered first. Collecting the correct diamond was rewarded by adding points to the participants' score, and collisions with cars were punished by subtracting points. As an experimental factor, traffic density was varied in the sections between junctions. Density was either low (no cars) or high (3.6 cars on average). The results indicated that the cycling TOJ task can be used to measure TVA's capacity parameter C: With approximately 69 Hz, processing capacity was higher in the low traffic condition than in the high traffic condition where it was estimated at 58 Hz. The values are plausible in terms of overall capacity, and the difference accords with studies that reported an influence of traffic density on mental workload (Strayer et al., 2003; Vlakveld et al., 2015).

In the present study, we extend the experiment with a salience manipulation which we expect to impact the relative weight w_p^*. To that aim, the diamonds were exchanged for highway cones, one of which we made more salient by giving it orange-colored instead of gray stripes. We assume that because of the increased salience, more capacity is assigned to this cone, indicated by a larger w_p^*. Overall processing capacity is expected to stay the same in both conditions. Traffic density was similar to the "high" condition of Stratmann et al., 2019.

We made further changes to the original design that were not assumed to affect the results: There were three cones, two of which flickered; the third one was never relevant. In accordance with the preceding experiments, but different

from the earlier study, they had to drive through the cone that flickered second.

10.1 Method

10.1.1 Participants

Thirty persons (21 male and 9 female; $M_{\text{age}} = 25.23$, range 20–35) participated. With the exception of one, all participants were students or members of Paderborn University. All participants gave informed written consent, completed one session, reported normal or corrected-to-normal visual acuity and received course credit or a payment of 8 euros per hour.

10.1.2 Apparatus

As shown in Figure 17**A**, we used the same hardware and simulation framework as in our previous virtual cycling experiment (Stratmann et al., 2019). Participants sat on a bicycle which was mounted on a bicycle stand to keep the bike stationary and to let the rear wheel rotate with some resistance. The visualization component was shown at a distance of 1.5 m from the bicycle handle on a triple monitor setup consisting of three 24″ 1920 × 1200 monitors with a 60 Hz refresh rate.

10.1.3 Stimuli

Two small red bumps on the street marked the beginning of each trial. When participants navigated centrally through the bumps, three cones appeared at a distance of 4.5 m, at least two of them gray-striped (nonsalient) and, in half of the trials, one orange with gray stripes (salient; see Figure 17**B** for a screenshot of the stimuli). The three 0.93 m tall cones were separated by 0.5 m and two of them flickered in an interval defined by the SOA. The SOA values were 0 ms, ±16.7 ms, ±33.3 ms, ±50 ms, ±66.7 ms, and ±83.3 ms. Trials were repeated 30 times for an SOA of 0, and 28, 26, 20, 16, and 8 times respectively for the other positive or negative SOA and for each condition (one salient stimulus or none). This results in an overall number of 452 trials split across two blocks, i.e., two runs through a cycling parcours. Because the length of each parcours is constant with 281 pairs of red bumps each marking a trial, participants typically collected more trial repetitions. This acted as a useful buffer to allow for repeating trials in which the cyclist missed collecting a cone, for example. Traffic was dynamically generated for the next intersection as soon as a participant entered a new 15 m long road segment. A video of the experiment can be found at https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3.

10.1.4 Procedure

Participants cycled through the parcours. At each crossroads, a sign indicated which direction to take. In each street section between two crossroads, three pairs of bumps each indicated the beginning of a trial. Immediately after cycling

through the bumps, the flickering cones appeared and the cyclist had to quickly steer though the cone that had flickered second. Again, participants did not receive any instructions regarding eye movements. A short training of 10 trials with feedback in the form of awarded points shown on the virtual bike computer allowed familiarization with the task. Additionally, after each of 6 levels and after the training, participants were shown a pause screen with both the collected points and the maximum points they could have reached. The experiment lasted approximately 60 min.

10.2 Results and Discussion

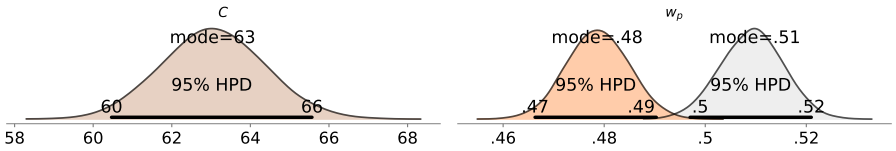

Figure 18: Results of Experiment 6: Means for processing capacity, C, and attentional weight of salient probe w_p^* in the experimental condition (orange cone, orange curve) and in the neutral condition (all-gray cones, gray curve).

Figure 18 shows the posterior estimates for the mean overall processing rate C and the mean attentional probe weight w_p^* across all participants. C was estimated at 62.5 Hz [95 % HPD: 59.9 Hz, 65.1 Hz] with an SD of 33 Hz [95 % HPD: 29.9 Hz, 37.5 Hz]. In the salient condition (orange cone), w_p^* was estimated at .48 [95 % HPD: .47, .494] with an SD of 0.06 [95 % HPD: 0.048, 0.07]. In the nonsalient condition, w_p^* was estimated at .51 [95 % HPD: .497, .522] with an SD of 0.05 [95 % HPD: 0.035, 0.056]. Individual fits are shown in the Appendix Figure 25 and the split-half reliability tests are reported in Appendix Section A.6.

Contrary to our expectation, we measured a lower attentional weight on the probe stimulus if it was colored bright orange instead of gray. Possibly the task of collecting the second cone that flickered rather than the first introduces complexity for participants that would warrant a longer training session. Furthermore, it is possible that the third (nonsalient) cone, of which we initially thought that it would make no difference, inclined the participants to spread their attention more equally over the three objects. As they needed to detect which of the gray cones was the target, they may have dedicated more attention to the gray than the orange cone. Another possibility that has to be tested in further experiments is that the cycling task requires more capacity, e.g., for reading the signs or detecting and avoiding other cars, reducing possible differences between attentional weights on the cones.

11 General Discussion

The present approach deviates from what experimental psychologists typically do: creating highly controlled environments to test hypotheses and develop theories to understand behavior and mind. This deviation may strike the reader as a particularly untimely approach as recent replication attempts (e.g., Open Science Collaboration, 2015) revealed substantial problems even with such strictly controlled experiments. However, we have a particular justification for this attempt that relies on a simple task, TOJ, in combination with a formal theory of attention, TVA.

TVA is based on the identification of common mechanisms of attention that explain results from experiments on selection and recognition described in the formal language of mathematics. The formal specification allows for severe testing. During this testing—in highly controlled environments—researchers reported good fits between theoretical prediction and observed data (beginning with Shibuya & Bundesen, 1988), and TVA has been shown to produce a reliable (Habekost et al., 2014) and meaningful formal (Logan, 2004) description of a person's attention.

In the present work, we have shown that TVA can also be used to investigate and describe attention during activities much less restricted than the experimental designs commonly used with TVA. Theoretically, this goal should be attainable because TVA and the formal TOJ model provide a basis for discerning expected from unexpected results. Practically, this goal has been promising because of the simplicity of the TOJ. It is simpler than the letter report design commonly used with TVA, can be done while being engaged in a second task and allows to use almost any visual stimulus material instead of being restricted to overlearned stimuli such as letters — for instance the holes in a grid or diamonds dangling in the air used in Experiments 4 to 6.

One way to characterize the presented experiments is as a game: We asked participants to engage in an activity that was not explicitly tailored for the measurement of attention but included rules that contained a TOJ. Degree of control was relaxed depending on the experiment: pursuing a second task while partaking in a TOJ-based attention measurement (riding a bicycle, playing a racing game), or allowing the participants to use their own device while not being isolated in an experimental booth. This is clearly not "in the wild", but an important step towards a measurement of attention in a real-life activity.

The estimated TVA parameters are within a reasonable range: The overall visual processing capacity, C, is comparable to the range of C estimates for healthy adults obtained with non-TOJ based experiments (Wiegand et al., 2014). In general, gaming reduces available processing capacity (Experiment 4), especially when the task gets rather difficult (adaptive speed condition in Experiment 5), although high capacity was also possible (nonadaptive speed condition in Experiment 5). The reduction is in accordance with TVA-based explanations. From a theoretical point of view, the C value does not have to be the same for an individual during different tasks because it depends on variable factors such as the sensory evidence of a stimulus belonging to a certain category. More

importantly, attention may draw processing capacity away from the stimuli and towards game-relevant locations and features. There are attempts within the TVA community towards more fine-grained models in which location components are included (Nordfang et al., 2018), but so far it remains an open question whether these could be set up in practice to capture all potentially relevant influences, even for a somewhat controlled environment like a computer-based 3D simulation.

The attentional weights could in most experiments be manipulated with the same ease and similar results as in laboratory settings. The weights of salient stimuli are close to the values reported in previous studies (Krüger et al., 2016, 2017) where the task and the salience manipulation on multi-element displays were very similar.

It is worth noting that it is not clear how eye movements may have affected the attentional weights. In Experiment 1-3, participants were instructed to fixate a central mark and there was no obvious advantage of moving the eyes. Eye movements could even be detrimental because the other target would be shifted into the periphery and unlucky timing might cause saccadic suppression of flicker events. Thus, we believe there was little incentive for performing eye movements. By contrast, eye movements were particularly likely in Experiments 4–6 where observers acted in more natural scenes. It remains unclear how eye-movements interact with the TVA-based estimates. There is as yet little theoretical and empirical research on this topic (some initial perspectives are provided by Schneider, 2013). In general, the fundamental mechanisms of TVA are based on objects and their features. Thus, it appears reasonable that these would be somewhat invariant with respect to eye movements. However, more recent research showed that attentional weights include a location-specific component (Nordfang et al., 2018) for which this would not be case. In future work, mobile eye tracking combined with experiments similar to the ones reported in this study might identify the relationship between TVA components and eye movements in dynamic scenes.

Even though not incompatible with TVA, some results remain unexplained. For instance, the high values for visual processing capacity in the nonadaptive conditions of Experiment 5 appear odd when compared to the values of the similar game in Experiment 4. Although there are possible TVA-based explanations (for instance the amount of sensory evidence that the colored stimuli provided) these are as yet not tested.

The reversed results in the cycling scenario, where we found an attentional weight bias in favor for the nonsalient target, are an example for a finding that is unexpected in the light of similar but lab-based experiments. Again, a TVA-based explanation is possible, but not tested. The weight distribution might have been caused by the fact that there were two possible nonsalient cones without any advance information about which one was the target so that the drivers distributed their attention more equally across the three important objects or even directed more attention to nonsalient cones to discern as quickly as possible which of the two is task relevant. The dynamic environment in such an interactive driving simulation contains many differences to typical lab experiments (many of

them would also be found in the real world). For instance, targets are visible for longer times and their projection moves across the visual field with their apparent size increasing. The changes over time invite participants to dynamically and possibly strategically direct their attention across the scene with more goals than just performing the TOJs. Identifying possible scenarios of what might be going on can lead to new hypotheses which then can be tested under lab conditions, leading to findings. It is likely that this uncovers phenomena that are easily missed if one only conducts experiments under strict control. In this way, experiments "in the wild" can help cumulative theory advancements and the generalizability of phenomena.

In the first three experiments we directly compared experimental conditions with lesser and higher degrees of experimental control which otherwise were identical in design, procedure and participants. We found that the model-based attention components were affected by experimenting "in the wild" in different forms which has some implications on how such measurements might be used by researchers. Allowing people to do experiments on any available mobile device biases the estimates of capacity towards lower values than found in similar but lab-based tasks because C depends on factors such as the visual evidence, which can be different in the lest restrictive settings. For instance, under normal light conditions, the effective stimulus contrast could be reduced. Moreover, at the variable viewing distance, retinal size of the stimuli could be reduced. Concerning w_p^*, some salience manipulations seem to transfer robustly into "the wild" when used in the flicker-TOJ-task. Oriented line segments however seem unsuited. The mean effect of salience is at best small and the expected correlation between lab and "wild" results is missing. The color salience manipulation worked fine outside of the lab, especially for a red–green contrast). Perhaps salience from local orientation contrasts relies more on the exact retinal size of the elements. Receptive fields dealing with local orientation might deal with the image statistics in a more narrow range that occurs in typical natural textures whereas relevant color contrasts can appear at a wide range of scales.

To sum up our findings, C is a reliable index of processing speed. The estimates from lab experiments correlated with those from less restrictive experiments, which were lower but in a reasonable range. Correlations between the salience-enhanced attentional weight w_p^* from the same people in the two conditions varied from nonexistent to large, apparently depending on the overall strength of the salience manipulation. According to our findings, it is possible to test scientific hypotheses in this way, but it might be difficult to get reliable w_p^* estimates for a person when rather subtle attention manipulations are used. The split-half reliability tests in Appendix Section A.6 show similar reliability for "in the wild" and lab conditions, although the reliability for the w_p^* estimates (salience condition) was overall low. For researchers interested in precise w_p^* estimates it is advisable to use more trials than we have.

A TOJ-based TVA assessment of attentional parameters looks promising enough for situations with relaxed experimental control, be it less control of the apparatus or more complex and dynamic visual scenes. Moreover, bodily activities (riding a bicycle), or engaging in an activity which is not directly

related to the TOJs does not lead to very different outcomes compared to usual lab-based research.

As mentioned in the passing in the Introduction, one further benefit of the method presented here is that it may be easily used in applied research and for questions where real-life adherence is of crucial importance. To give an example, TVA-based measures can be used to pinpoint attention-related deficits in clinical populations (for a review see Habekost, 2015) or in aging (e.g. Habekost et al., 2013; Künstler et al., 2018). Laboratory TVA tasks often show a decline of TVA parameters with age. Despite TVAs specificity in assessing attention-related parameters, some studies have shown that this typical laboratory assessment may underestimate the cognitive capacities of, e.g., elderly persons (e.g. Wiegand & Wolfe, 2020). Using a more realistic hybrid search task, Wiegand and Wolfe (2020) found no evidence for an age-related decline, except for overall response time. One possible explanation of these contradictory findings is that performance decline is overestimated in laboratory tasks because older adults may be little trained to presence their cognitive functions in real-world than tasks. The same problem impairs assessment in clinical populations. More generally, attentional abilities may generally be underestimated in the repetitive and barely motivating assessment of TVA parameters, and it is likely that there are groups of participants (for instance children) who suffer from this more than others. The approach we followed in the present paper may help to establish more realistic tasks conditions that would allow to investigate these possibilities and it would complement successful approaches to test TVA in more realistic scenarios by using, for instance, head-mounted displays (Foerster et al., 2019).

11.1 Conclusion and Outlook

The present article pushed toward the application of simple psychophysical tasks followed by model-based analysis in more natural, real-life scenarios—"the wild". We have used classical paradigms with less experimental control (Experiments 1 to 3) and even game-like scenarios to frame TOJs (Experiments 4 to 6). The results confirm that research conducted in this manner—using a simple task and a fine-grained model—is similarly informative as typical lab-based experiments and offers new possibilities. On the one hand, research conducted in this manner could make use of more diverse and more representative participant samples. A mobile phone or tablet can be carried into a café or a street market as easily as a questionnaire. On the other hand, research with children (and possibly also adults) can highly benefit from the game-like character that keeps the task interesting and participants motivated. Moreover, including TOJs in driving simulations provides a semi-naturalistic task. Researchers can investigate how different user interfaces or various environmental factors (e.g., the traffic in the scene or a distracting person in the passenger seat) interfere with attentional resources and guidance. Moreover, in addition to the binary judgments, rich data beyond the usual key presses, such as driving trajectories or other action related components, can be recorded and analyzed.

Arguably, we have not arrived in the real wild yet. Which challenges must be met to extend this work toward even more natural scenarios, possibly embedded in the real world? Because the TOJ task is very simple, it seems possible to integrate TOJ stimulation in the real world. A step forward could be achieved by augmented reality glasses, which display TOJ events with probe and reference stimuli on top of the natural visual scene, probing attention at certain locations of interest. It also might be possible to glue wireless battery-driven LEDs to many objects of interest (vehicles, tools, paintings). The judgment could be indicated by button presses, gestures, or maybe verbally.

A main challenge is to record a sufficient number of TOJ trials under such conditions. In the game-like scenarios we used in this article, the task repeatedly occurred in rather close succession. This would not only be difficult to implement in the real world, the highly repetitive artificial task interrupting normal behavior would itself appear rather odd. Very rare events collected over long time spans could be used. This might be a possibility for overall processing capacity C which is thought to be more or less stable but it would be difficult to capture the momentary (and constantly changing) attention distribution via w_p^*. Bayes estimation, as we used for the analyses of this study, can work with rather few data. At least, it explicitly captures the uncertainty that results from a low number of repetitions and allows to formally integrate prior knowledge obtained from earlier experiments. Hence, evidence can be accumulated in small pieces. The parameter distributions estimated in the present study can provide a basis for power analyses via simulations (e.g., see Kruschke, 2010). Such simulations could help to determine how many more participants one might need when performing TOJs under less controlled conditions and potentially with a smaller amount of trials. A further methodological recommendation for researchers who want to apply TOJs in the wild is to use a hierarchical version of this model and well informed priors, at least if they are not or less interested in discovering discrepancies (as we are) but rather want to shield off weirdness from the wild via shrinkage and prior knowledge about typical values (Kruschke & Vanpaemel, 2015).

In the long term, research should target to model real-world tasks (e.g., avoiding obstacles while walking) in a similar fine-grained manner as TVA can be used to model TOJs and similar artificial tasks. Then relevant events could be recorded during normal behavior and submitted to model-based analysis.

Data and Videos

The data collected in the six experiments of this study can be downloaded at: https://osf.io/sdk8r/?view_only=8aed2f9c6ca54d18b0456a4ce9662cd3. The page also contains demo videos of all the experiments.

Author Contributions and Acknowledgments

We are very grateful to Björn Luchterhandt for programming Experiments 1–3 and to Ngoc Chi Banh, Miriam Körber, and Björn Luchterhandt for conducting the experiments and doing preliminary analyses as well as preparing some of the figures. Alexander Krüger, Jan Tünnermann and Ingrid Scharlau planned the experiments. Alexander Krüger and Jan Tünnermann modeled the data. Lukas Stratmann programmed and analysed Experiments 4 and 6 under the supervision of Alexander Krüger (Experiment 4) and Ingrid Scharlau and Falko Dressler (Experiment 6). Lucas Briese programmed and conducted Experiment 5 under the supervision of Alexander Krüger and Ingrid Scharlau. All authors contributed to the final manuscript.

In the case of Experiment 6, research was supported in part by the project Safety4Bikes funded by the German Federal Ministry of Education and Research (BMBF) under grant number 16SV7672.

References

Bailer-Jones, D. M. (2009). *Scientific models in philosophy of science*. Pittsburg: PA, University of Pittsburgh Press.

Berkhof, J., Van Mechelen, I., & Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*, 337–354.

Briese, L. (2019). *Gamification of psychological experiments*. Paderborn, Germany, Paderborn University.

Bundesen, C., Vangkilde, S., & Habekost, T. (2015). Components of visual bias: A multiplicative hypothesis. *Annals of the New York Academy of Sciences*, *1339*, 116–124.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*, 523–547. https://doi.org/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., Habekost, T., & Kyllingsbæk, S. (2005). A neural theory of visual attention: Bridging cognition and neurophysiology. *Psychological Review*, *112*, 291–328. https://doi.org/10.1037/0033-295X.112.2.291

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

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ, Princeton University Press.

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to bayesian model checking for ecologists. *Ecological Monographs*, *88*(4), 526–542. https://doi.org/https://doi.org/10.1002/ecm.1314

de Leeuw, J. R. (2015). JsPsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290. https://doi.org/10.1177/1745691611406920

Foerster, R. M., Poth, C. H., Behler, C., Botsch, M., & Schneider, W. X. (2019). Neuropsychological assessment of visual selective attention and processing capacity with head-mounted displays. *Neuropsychology*, *33*(3), 309–318. https://doi.org/http://dx.doi.org/10.1037/neu0000517

Grisson, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. MahWah, NJ, Lawrence Erlbaum.

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

Habekost, T., Petersen, A., & Vangkilde, S. (2014). Testing attention: Comparing the ANT with TVA-based assessment. *Behavior Research Methods*, *46*(1), 81–94. https://doi.org/10.3758/s13428-013-0341-2

Habekost, T., Vogel, A., Rostrup, E., Bundesen, C., Kyllingsbæk, S., Garde, E., Ryberg, C., & Waldemar, G. (2013). Visual processing speed in old age. *Scandinavian Journal of Psychology*, *54*(2), 89–94. https://doi.org/http://dx.doi.org/10.1111/sjop.12008

Haupt, M., Ruiz-Rizzo, C., A. L. amd Sorg, & Finke, K. (2015). Phasic alerting effects on visual processing speed are associated with intrinsic functional connectivity in the cingulo-opercular network. *NeuroImage*, *196*, 216–226.

Heinovski, J., Stratmann, L., Buse, D. S., Klingler, F., Franke, M., Oczko, M.-C. H., Sommer, C., Scharlau, I., & Dressler, F. (2019). Modeling cycling behavior to improve bicyclists' safety at intersections: A networking perspective, In *20th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (wowmom 2019)*, Washington, D.C., IEEE. https://doi.org/10.1109/WoWMoM.2019.8793008

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

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.

JASP Team. (2019). JASP (Version 0.11.1)[Computer software]. https://jasp-stats.org/

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 (Online First)*, *5*(1), 1–27. https://doi.org/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. https://doi.org/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. https://doi.org/10.3758/s13414-017-1325-6

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Kruschke, J. K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models (J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels, Eds.). In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The oxford handbook of computational and mathematical psychology*. Oxford, UK, Oxford University Press.

Künstler, E. C. S., Penning, M. D., Napiórkowski, N., Klingner, C. M., Witte, O. W., Müller, H. J., Bublak, P., & Finke, K. (2018). Dual task effects on visual attention capacity in normal aging. *Frontiers in Psychology*, *9*. https://doi.org/http://dx.doi.org/10.3389/fpsyg.2018.01564

Lange, K., Kühn, S., & Filevich, E. (2015). "just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS One*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK, Cambridge University Press.

Logan, G. D. (2004). Cumulative progress in formal theories of attention. *Annual Review of Psychology*, *55*, 207–234. https://doi.org/10.1146/annurev.psych.55.090902.141415

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*, 215–233.

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

Matthias, E., Bublak, P., Müller, H. J., Schneider, W. X., Krummenacher, J., & Finke, K. (2010). The influence of alertness on spatial and nonspatial components of visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 38–56. https://doi.org/10.1037/a0017602

Meehl, P. M. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Muthukrishna, M., & Henrich, J. (2004). A problem in theory. *Nature Human Behavior*, *3*, 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nordfang, M., Staugaard, C., & Bundesen, C. (2018). Attentional weights in vision as products of spatial and nonspatial components. *Psychonomic Bulletin & Review*, *25*, 1043–1051. https://doi.org/10.3758/s13423-017-1337-1

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Peirce, J. W. (2007). PsychoPy: Psychophysics software in python. *Journal of Neuroscience Methods*, *162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Petersen, A., Petersen, A. H., Bundesen, C., Vangkilde, S., & Habekost, T. (2017). The effect of phasic auditory alerting on visual perception. *Cognition*, *165*, 73–81.

Petersen, A., Kyllingsbæk, S., & Bundesen, C. (2013). Attentional dwell times for targets and masks. *Journal of Vision*, *13*(3), 34–34. https://doi.org/10.1167/13.3.34.

Petrini, K., Denis, G., Love, S. A., & Nardini, M. (2020). Combining the senses: The role of experience- and task-dependent mechanisms in the development of audiovisual simultaneity perception. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/http://dx.doi.org/10.1037/xhp0000827

Poth, C. H., Petersen, A., Bundesen, C., & Schneider, W. X. (2014). Effects of monitoring for visual events on distinct components of attention. *Frontiers in psychology*, *5*, 930.

Rorden, C., Mattingley, J. B., Karnath, H.-O., & Driver, J. (1997). Visual extinction and prior entry: Impaired perception of temporal order with intact motion perception after unilateral parietal damage. *Neuropsychologia*, *35*(4), 421–433. https://doi.org/http://dx.doi.org/10.1016/S0028-3932(96)00093-0

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

Schneider, W. X. (2013). Selective visual processing across competition episodes: A theory of task-driven visual attention and working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628), 20130060.

Schubert, T., Finke, K., Redel, P., Kluckow, S., Müller, H., & Strobach, T. (2015). Video game experience and its influence on visual attention parameters: An investigation using the framework of the theory of visual attention (TVA). *Acta Psychologiga*, *157*, 200–214. https://doi.org/http://dx.doi.org/10.1016/j.actpsy.2015.03.005

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241–1260. https://doi.org/10.3758/s13428-016-0783-4

Shibuya, H., & Bundesen, C. (1988). Visual selection from multielement displays: Measuring and modeling effects of exposure duration. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 591–600. https://doi.org/1037//0096-1523.14.4.591

Sternberg, S., & Knoll, R. L. (1973). The perception of temporal order: Fundamental issues and a general model. *Attention and Performance IV*, 629–685.

Stratmann, L., Buse, D. S., Heinovski, J., Klingler, F., Sommer, C., Tünnermann, J., Scharlau, I., & Dressler, F. (2019). Psychological feasibility of a virtual

cycling environment for human-in-the-loop experiments (C. Draude, M. Lange, & B. Sick, Eds.). In C. Draude, M. Lange, & B. Sick (Eds.), *Jahrestagung der Gesellschaft für Informatik (INFORMATIK 2019), 1st Workshop on ICT based Collision Avoidance for VRUs (ICT4VRU 2019)*, Kassel, Germany, GI. https://doi.org/10.18420/inf2019_ws21

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, *9*(1), 23–32.

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

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

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*(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*, *14*, 39–50. https://doi.org/0.5709/acp-0237-0

Tünnermann, J., & Scharlau, I. (2018b). Stuck on a plateau? a model-based approach to fundamental issues in visual temporal-order judgments. *Vision*, *2*(3), 1–29.

Tünnermann, J., & Scharlau, I. (2018c). Stuck on a plateau? model-based analysis of temporal-order judgments. *Vision*, *2*(3), 29. https://doi.org/10.3390/vision203002

Vangkilde, S., Bundesen, C., & Coull, J. T. (2011). Prompt but inefficient: Nicotine differentially modulates discrete components of attention. *Psychopharmacology*, *218*(4), 667–680. https://doi.org/10.1007/s00213-011-2361-x

Vangkilde, S., Coull, J. T., & Bundesen, C. (2012). Great expectations: Temporal expectation modulates perceptual processing speed. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 1183–1191.

Vlakveld, W. P., Twisk, D., Christoph, M., Boele, M., Sikkema, R., Remy, R., & Schwab, A. L. (2015). Speed choice and mental workload of elderly cyclists on e-bikes in simple and complex traffic situations: A field experiment. *Accident Analysis and Prevention*, *74*, 97–106.

Wada, M., Moizumi, S., & Kitazawa, S. (2005). Temporal order judgment in mice. *Behavioural Brain Research*, *157*, 167–175.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Wiegand, I., Töllner, T., Dyrholm, M., Müller, H. J., Bundesen, C., & Finke, K. (2014). Neural correlates of age-related decline and compensation

in visual attention capacity. *Neurobiology of Aging*, *35*(9), 2161–2173. https://doi.org/10.1016/j.neurobiolaging.2014.02.023

Wiegand, I., & Wolfe, J. M. (2020). Age doesn't matter much: Hybrid visual and memory search is preserved in older adults. *Aging, Neuropsychology, and Cognition*, *27*(2), 220–253. https://doi.org/http://dx.doi.org/10.1080/13825585.2019.1604941

Wilms, I. L., Petersen, A., & Vangkilde, S. (2013). Intensive video gaming improves encoding speed to visual short-term memory in young male adults. *Acta Psychologiga*, *142*, 108–118. https://doi.org/http://dx.doi.org/10.1016/j.actpsy.2012.11.003

A Appendix

A.1 Devices, operating systems, and browsers used in Experiment 1

[1] BLN-L21 / Android 7.0 / Opera Mobile 53.1.2569
[2] Generic Smartphone / Android / Chrome Mobile 74.0.3729
[3] Generic Smartphone / Android / Chrome Mobile 76.0.3809
[4] Generic Smartphone / Android 7.0 / Chrome 77.0.3865
[5] Generic Smartphone / Android 7.0 / Chrome Mobile 76.0.3809
[6] Generic Smartphone / Android 7.0 / Chrome Mobile 77.0.3865
[7] Generic Smartphone / Android 7.0 / Firefox Mobile 68.0
[8] Generic Smartphone / Android 8.0.0 / Chrome Mobile 77.0.3865
[9] PC / Mac OS X 10.11 / Firefox 68.0
[10] PC / Mac OS X 10.13.6 / Chrome 76.0.3809
[11] PC / Mac OS X 10.14.6 / Safari 13.0.1
[12] PC / Mac OS X 10.15 / Safari 13.0.2
[13] PC / Ubuntu / Firefox 69.0
[14] PC / Windows 10 / Chrome 76.0.3809
[15] PC / Windows 10 / Edge 17.17134
[16] PC / Windows 10 / Firefox 44.0
[17] PC / Windows 10 / Firefox 68.0
[18] PC / Windows 10 / Firefox 69.0
[19] PC / Windows 8.1 / Chrome 75.0.3770
[20] Samsung SM-A520F / Android 8.0.0 / Samsung Internet 10.1
[21] Samsung SM-J510FN / Android 7.1.1 / Chrome Mobile 69.0.3497
[22] iPad / iOS 12.4.1 / Mobile Safari 12.1.2
[23] iPhone / iOS 11.4.1 / Mobile Safari 11.0
[24] iPhone / iOS 12.1 / Mobile Safari 12.0
[25] iPhone / iOS 13.1.1 / Mobile Safari 13.0.1
[26] iPhone / iOS 13.1.2 / Mobile Safari 13.0.1

A.2 Devices, operating systems, and browsers used in Experiment 2

[1] Generic Smartphone / Android / Chrome 77.0.3865
[2] Generic Smartphone / Android / Chrome Mobile 77.0.3865
[3] Generic Smartphone / Android 7.1.1 / Chrome Mobile 77.0.3865
[4] Kindle / Android 5.1.1 / Amazon Silk 77.3.1
[5] PC / Mac OS X 10.13.6 / Safari 12.1.1
[6] PC / Mac OS X 10.14.5 / Safari 12.1.1
[7] PC / Mac OS X 10.14.6 / Safari 13.0.2
[8] PC / Mac OS X 10.15 / Safari 13.0.2
[9] PC / Windows / Firefox 69.0
[10] PC / Windows 10 / Chrome 77.0.3865
[11] PC / Windows 10 / Edge 16.16299
[12] PC / Windows 10 / Edge 17.17134
[13] PC / Windows 10 / Edge 18.18362
[14] PC / Windows 10 / Firefox 69.0
[15] PC / Windows 10 / Opera 63.0.3368
[16] Samsung SM-G950F / Android / Samsung Internet 10.1
[17] Samsung SM-J600FN / Android / Chrome Mobile 69.0.3497
[18] iPhone / iOS 12.3.1 / Mobile Safari 12.1.1
[19] iPhone / iOS 12.4 / Mobile Safari 12.1.2
[20] iPhone / iOS 13.1 / Chrome Mobile 78.0.3904
[21] iPhone / iOS 13.1.2 / Mobile Safari 13.0.1
[22] iPhone / iOS 13.1.3 / Mobile Safari 13.0.1

A.3 Devices, operating systems, and browsers used in Experiment 3

[1] Generic Smartphone / Android / Chrome Mobile 76.0.3809
[2] Generic Smartphone / Android / Chrome Mobile 78.0.3904
[3] Generic Smartphone / Android / Firefox Mobile 68.0
[4] Generic Smartphone / Android 7.1.1 / Chrome Mobile 78.0.3904
[5] Generic Smartphone / Android 8.0.0 / Chrome Mobile 70.0.3538
[6] Generic Smartphone / Android 8.0.0 / Chrome Mobile 78.0.3904
[7] PC / Ubuntu / Firefox 70.0
[8] PC / Windows / Firefox 69.0
[9] PC / Windows 10 / Chrome 77.0.3865
[10] PC / Windows 10 / Chrome 78.0.3904
[11] PC / Windows 10 / Edge 17.17134
[12] PC / Windows 10 / Firefox 70.0
[13] PC / Windows 10 / Opera 64.0.3417
[14] Samsung SM-A320FL / Android 8.0.0 / Samsung Internet 10.1
[15] iPad / iOS 12.3.1 / Mobile Safari 12.1.1
[16] iPhone / iOS 12.3.1 / Mobile Safari 12.1.1
[17] iPhone / iOS 12.4 / Mobile Safari 12.1.2
[18] iPhone / iOS 13.1 / Chrome Mobile 78.0.3904
[19] iPhone / iOS 13.1 / Mobile Safari 13.0.1
[20] iPhone / iOS 13.1.2 / Mobile Safari 13.0.1
[21] iPhone / iOS 13.1.3 / Mobile Safari 13.0.1
[22] iPhone / iOS 13.2 / Mobile Safari 13.0.3

A.4 Participant-level data and predicted TOJ curves

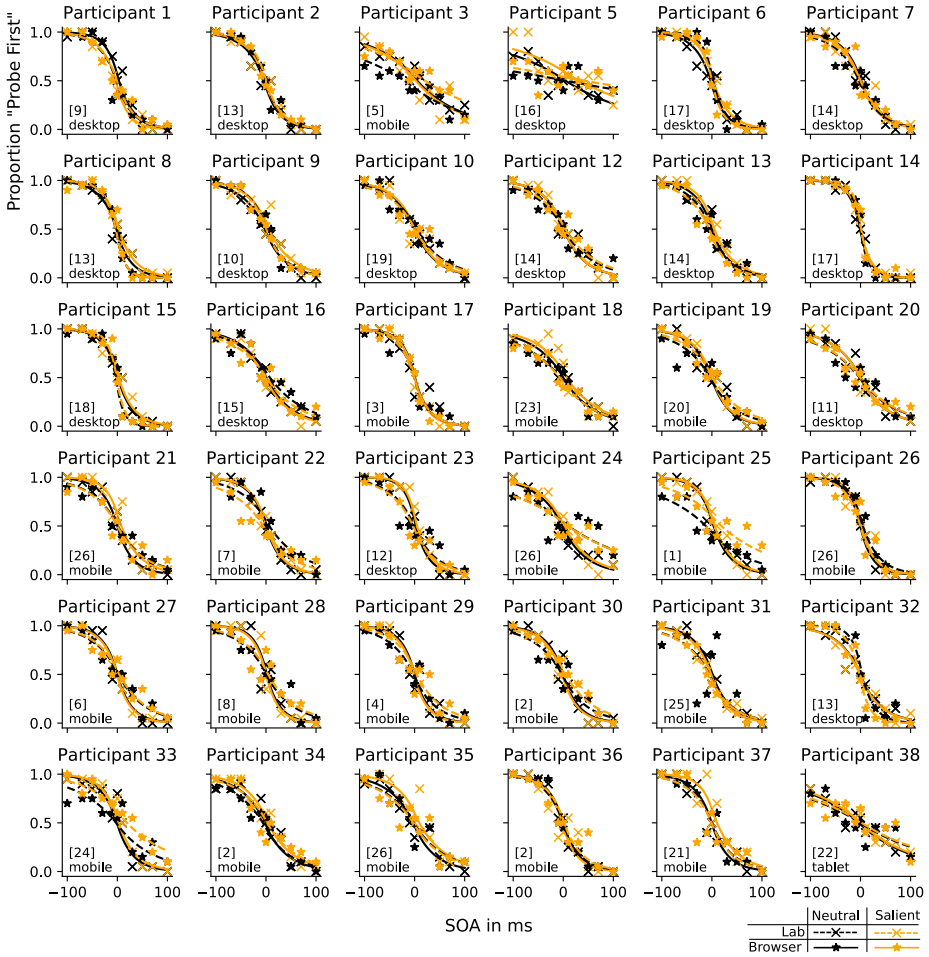

Figure 19: Results of Experiment 1: Individual data (points) and model estimates (curves; based on Equation 3 and the posterior modes of the C and w_p^*) for the four conditions. The numbers in square brackets identify the device and browser used in the browser–mobile condition, referring to the list in Appendix A.1.

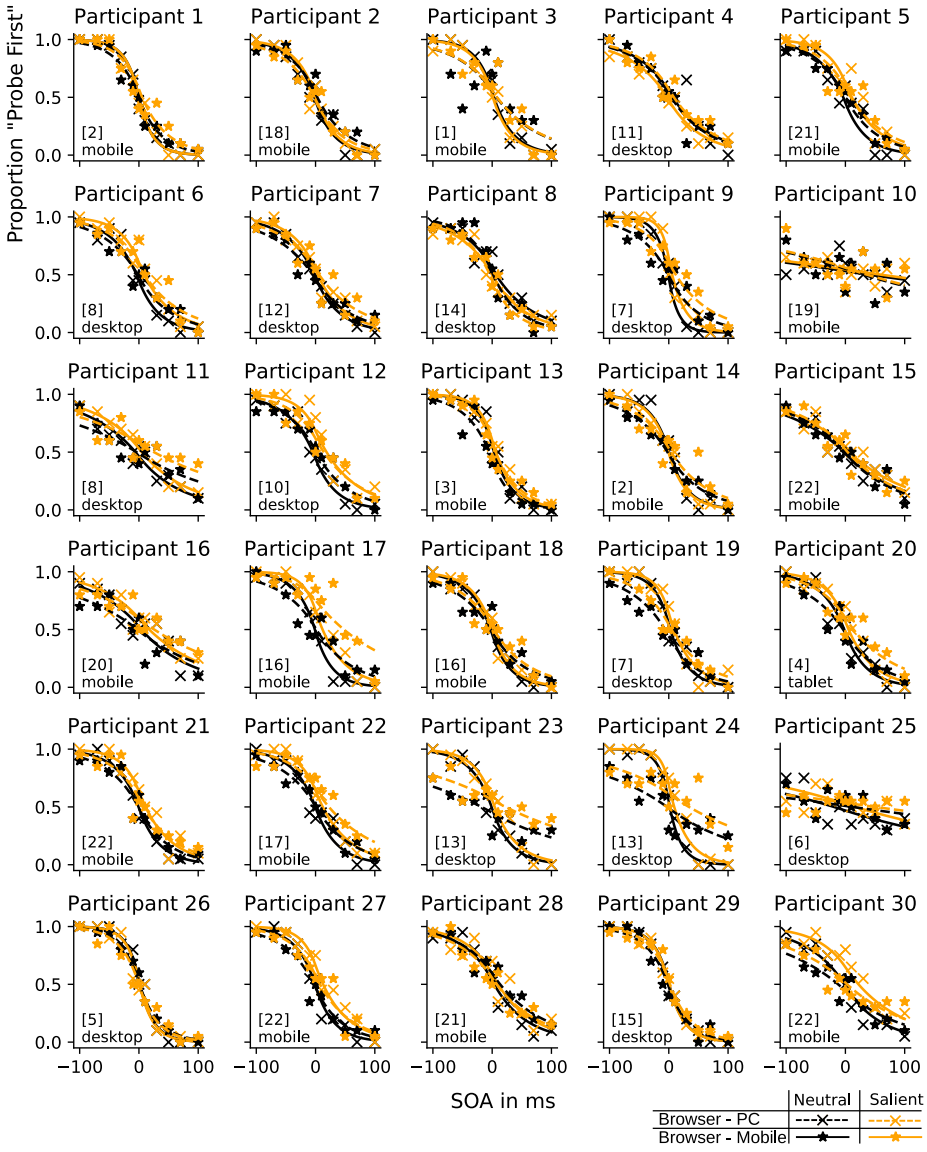

Figure 20: Results of Experiment 2: Individual data (points) and model estimates (curves; based on Equation 3 and the posterior modes of C and w_p^*) for the four conditions. The numbers in square brackets identify the device and browser used in the browser–mobile condition, referring to the list in Appendix A.2.

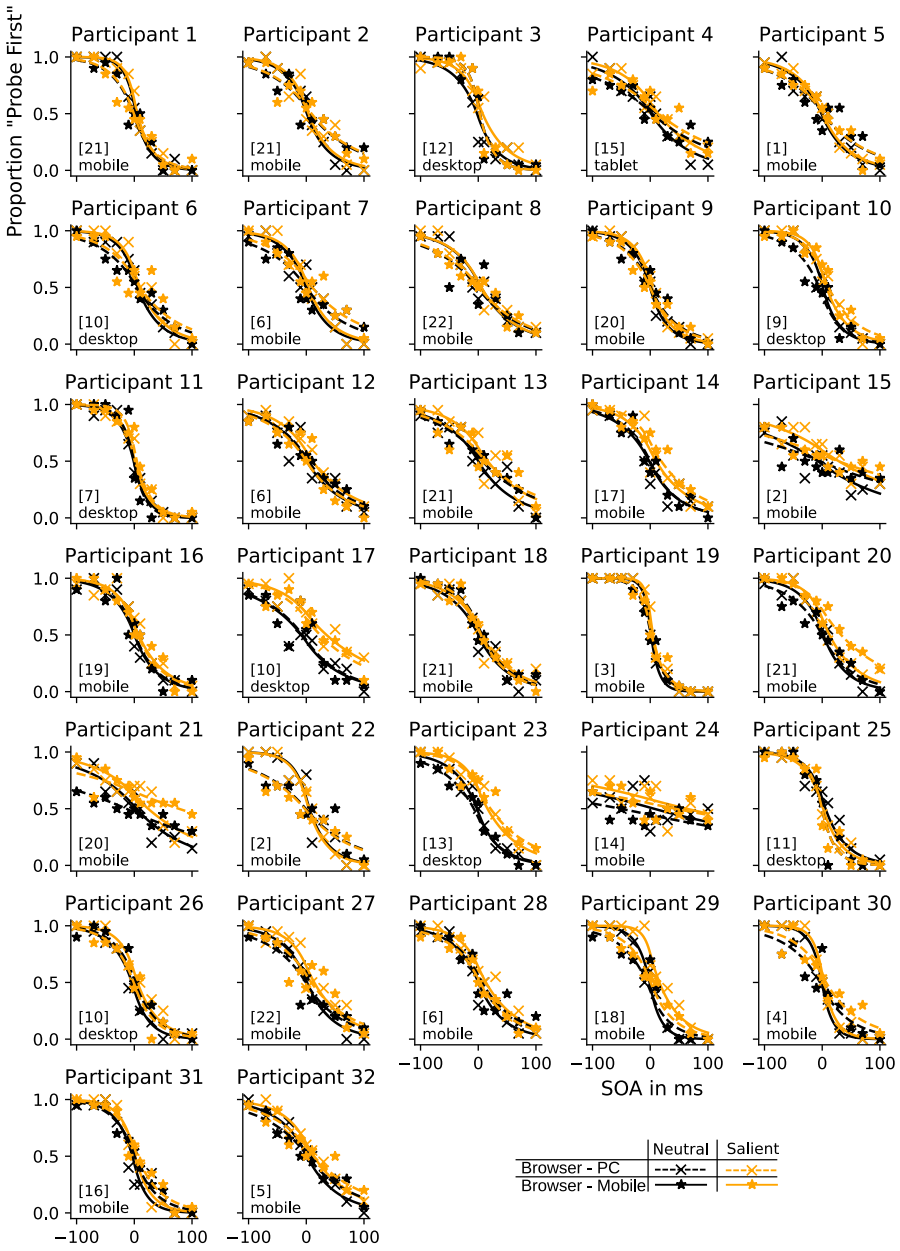

Figure 21: Results of Experiment 3: Individual data (points) and model estimates (curves; based on Equation 3 and the posterior modes of C and w_p^*) for the four conditions. The numbers in square brackets identify the device and browser used in the browser–mobile condition, referring to the list in Appendix A.3.

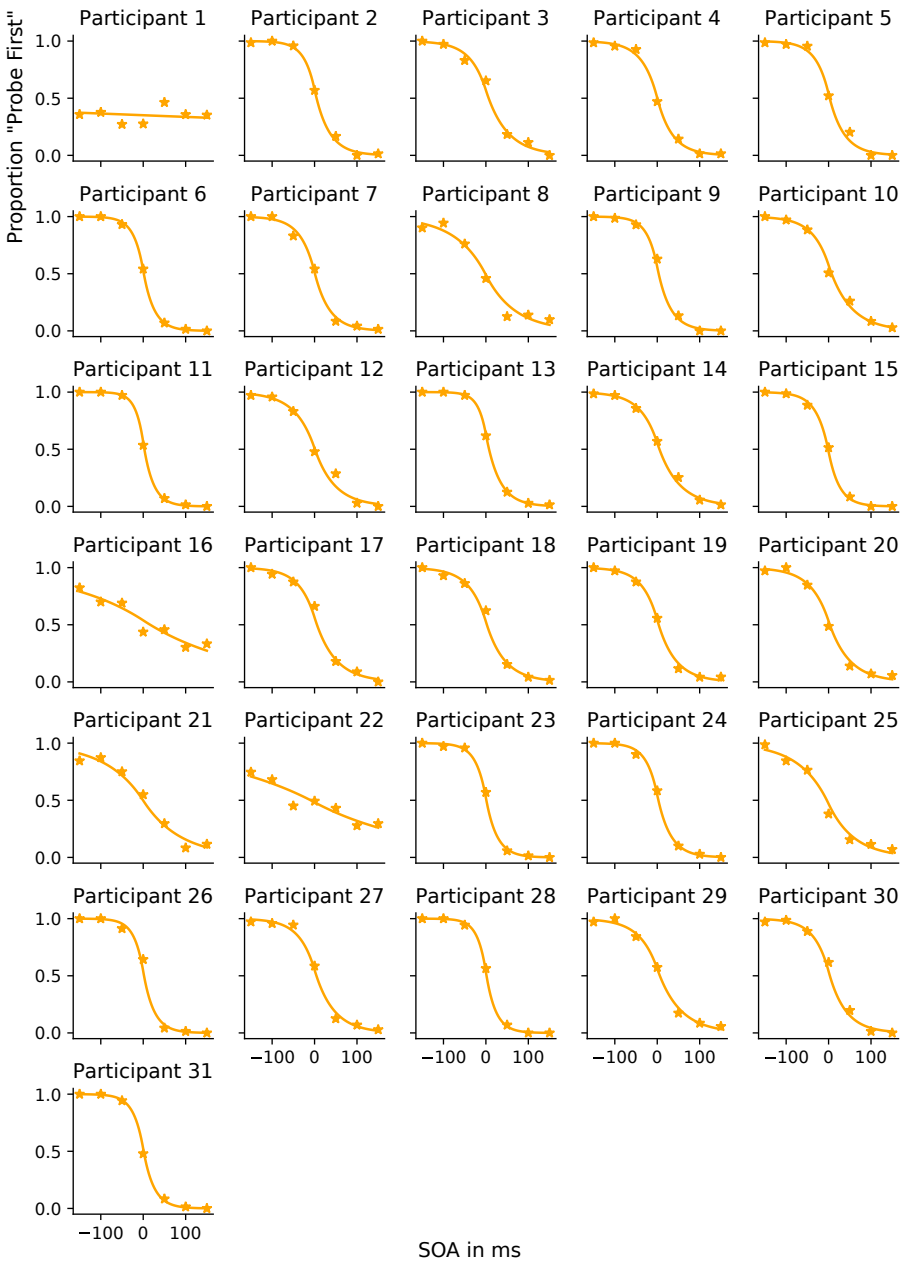

Figure 22: Results of Experiment 4, game condition: Individual data (points) and model estimates (curves; based on Equation 3 and the posterior modes of C and w_p^*).

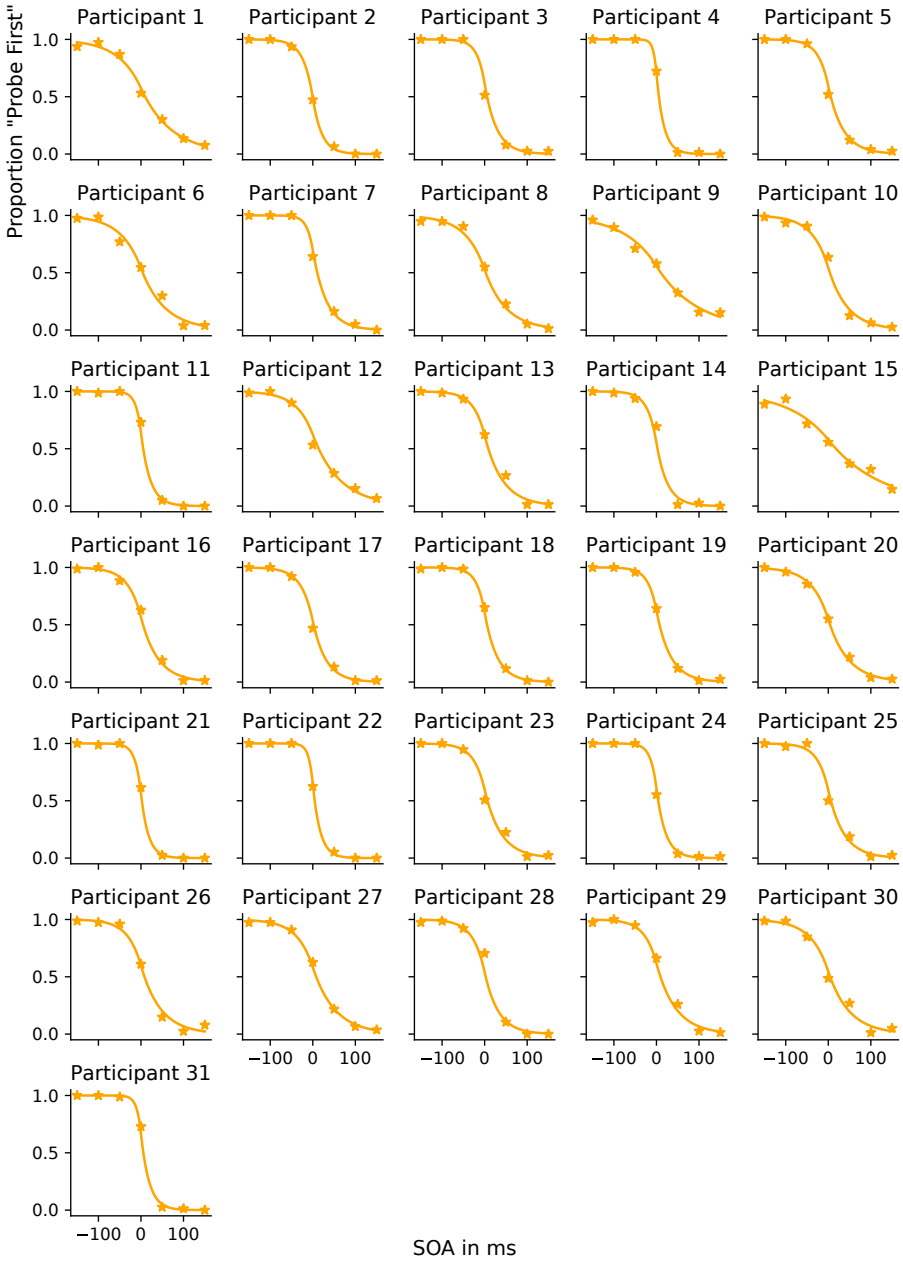

Figure 23: Results of Experiment 4, control condition: Individual data (points) and model estimates (curves; based on Equation 3 and the posterior modes of C and w_p^*).

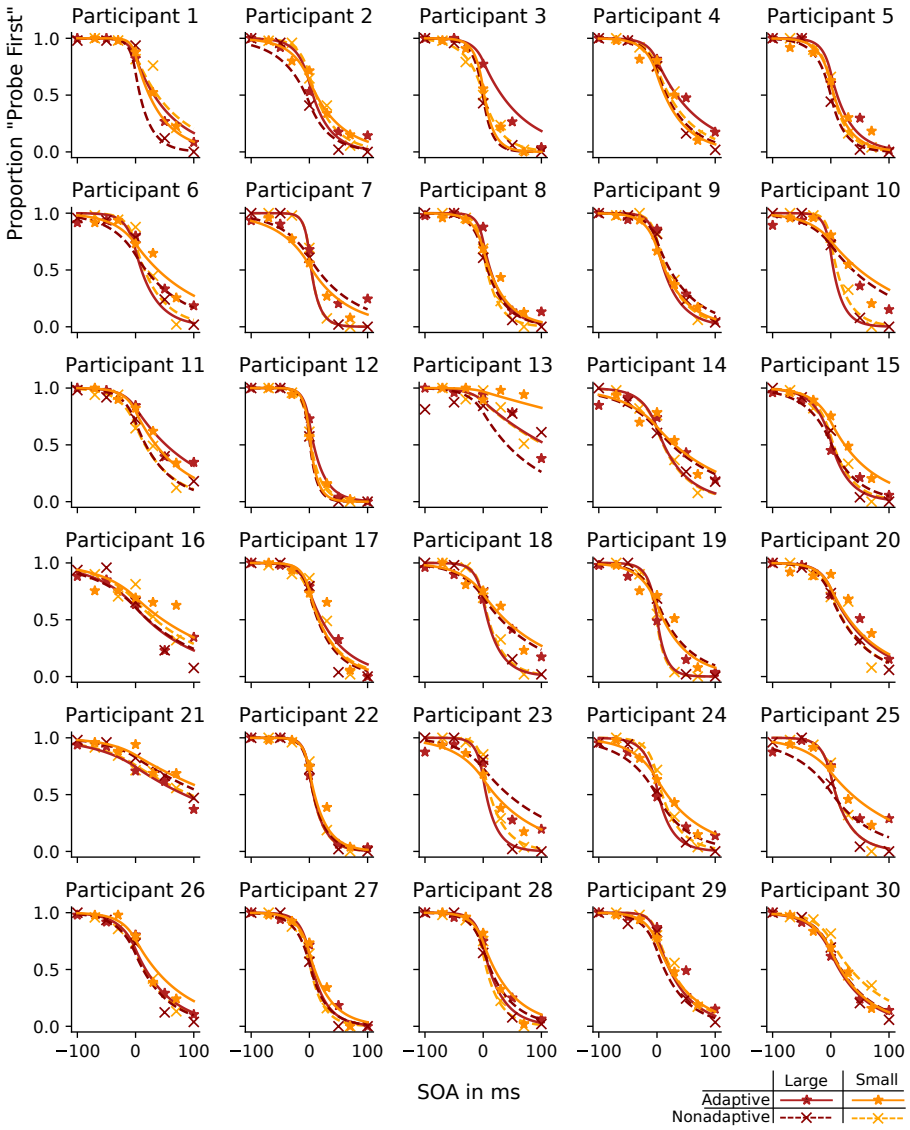

Figure 24: Results of Experiment 5: Individual data (points) and model estimates (curves; based on Equation 3 and the posterior modes of C and w_p^*).

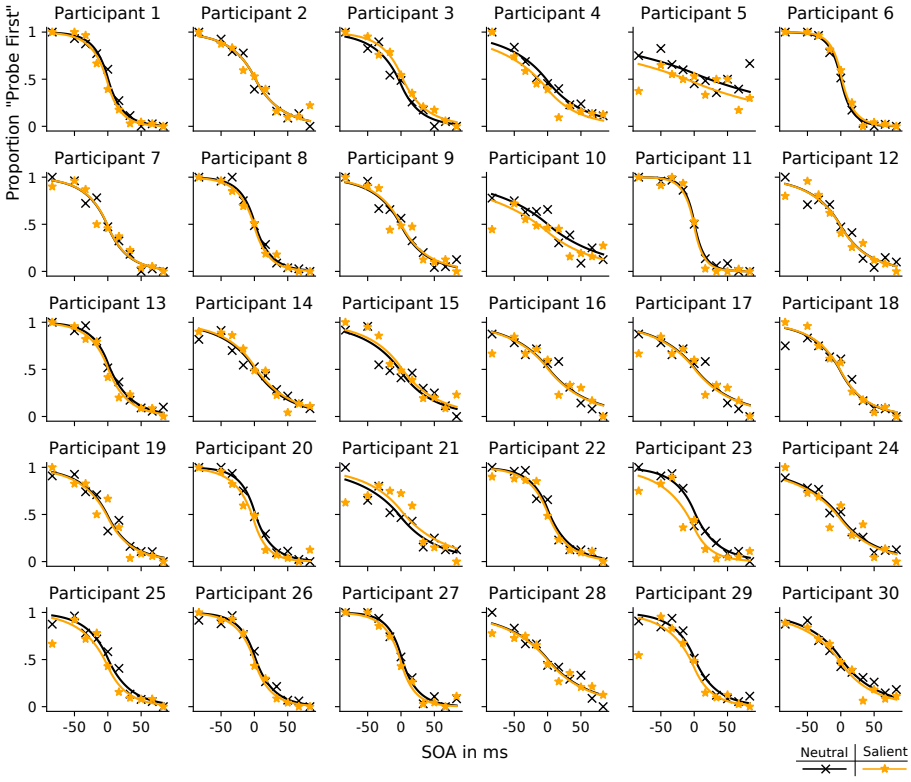

Figure 25: Results of Experiment 6: Individual data (points) and model estimates (curves; based on Equation 3 and the posterior modes of C and w_p^*) for the two conditions.

A.5 Sequence analyses for the correlations reported in Experiment 1 to 3

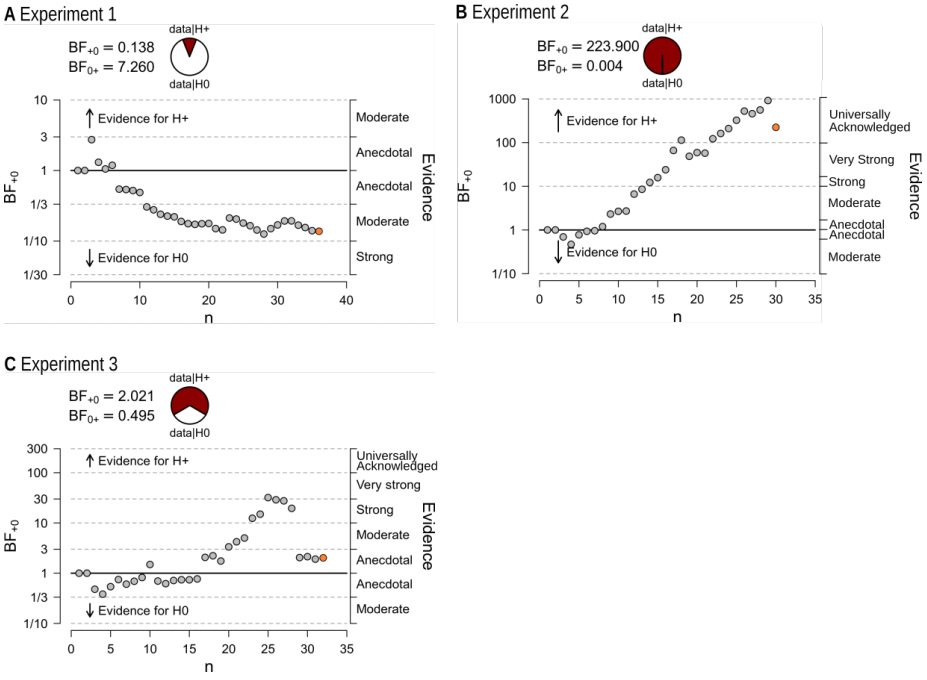

Figure 26: (a) to (c): Sequence analysis plots for the correlations of the w_p^* from the lab and mobile conditions for Experiments 1 to 3. The trajectory of evidence accumulation is shown for each participant added (n) to the analysis. The orange disk marks the final level of evidence.

A.6 Split-half reliability tests

Table 3–4 show results split-half reliability test. These were conducted by creating two non-overlapping random (but balanced concerning all conditions) sub-samples of the raw data of each experiment. These were then analyzed in the same way as in the main analysis. Point estimates for the parameters were obtained for each participant in each condition as the mode of the respective posterior. These point estimates were then correlated with a Bayesian version of a Pearson correlation (using JASP, JASP Team, 2019). The correlation coefficient ρ is an index of the reliability. Note that in the neutral conditions (Table 5), all variability in the w_p^* estimates must be due to chance and hence no correlation is expected for the parameter in this condition.

Experiment	C-value Lab			C-value "Wild" "Wild"		
	ρ	95 % HPD	BF+	ρ	95 % HPD	BF+
1	.62	[.39, .8]	2950	.8	[.67, .9]	6.45E+07
2	.89	[.79, .95]	2.68E+09	.8	[.64, .91]	2.07E+06
3	.84	[.71, .93]	8.85E+07	.9	[.81, .96]	5.62E+10
4	.59	[.33, .79]	393	.74	[.54, .87]	8.03E+4
5_{ns}				.54	[.26, .76]	94.8
5_{nl}				.63	[.38, .82]	1130
5_{as}				.59	[.32, .79]	294
5_{al}				.67	[.44, .84]	4160
6				.71	[.49, .86]	1.57E+04

Table 3: Split-half reliability tests for the C estimates. ρ: Pearson correlation coefficient; 95 % HPD: 95 % Highest Probability Density interval for the ρ estimate; BF+: Bayes Factor in favor of a positive correlation. For Experiment 5, the subscripts "a" and "n" refer to adaptive and nonadaptive, and "s" and "l" refers to small and large SOAs.

Experiment	w_p^*-value Salient Lab			w_p^*-value Salient "Wild"		
	ρ	95 % HPD	BF+	ρ	95 % HPD	BF+
1	.25	[.02, .52]	1.07	.13	[.01, .39]	0.26
2	.34	[.06, .63]	3.05	.52	[.23, .75]	60
3	.4	[.1, .66]	7.307	.42	[.12, .68]	11.1
4	.35	[.06, .62]	3.33	.32	[.04, .6]	2.21
5_{ns}				.71	[.49, .86]	1.41E+5
5_{nl}				.47	[.17, .72]	21.48
5_{as}				.69	[.47, .85]	8190
5_{al}				.77	[.59, .9]	3.58E+5
6				.2	[.01, .49]	0.53

Table 4: w_p^* estimates (salience condition). Split-half reliability tests for the C estimates. ρ: Pearson correlation coefficient; 95 % HPD: 95 % Highest Probability Density interval for the ρ estimate; BF+: Bayes Factor in favor of a positive correlation. For Experiment 5, the subscripts "a" and "n" refer to adaptive and nonadaptive, and "s" and "l" refers to small and large SOAs.

Experiment	w_p^*-value Neutral Lab			w_p^*-value Neutral "Wild"		
	ρ	95 % HPD	BF+	ρ	95 % HPD	BF+
1	0.05	[0, .2]	0.08	0.13	[0, .4]	0.26
2	0.13	[0, 4]	0.25	0.2	[.01, .5]	0.6
3	0.1	[0, .36]	0.19	0.2	[.06, .5]	0.66
6				.13	[.01, .4]	0.25

Table 5: w_p^* estimates (neutral condition) Split-half reliability tests for the C estimates. ρ: Pearson correlation coefficient; 95 % HPD: 95 % Highest Probability Density interval for the ρ estimate; BF+: Bayes Factor in favor of a positive correlation. In neutral conditions, no correlations are expected. Not all experiments contained neutral conditions.