PhD IN COMPUTER SCIENCE

MY MIND'S PLAYING TRICKS ON ME: UNDERSTANDING EVENT INTEGRATION IN RAPID STIMULUS STREAMS

BY MAHAN NAYEB GHANBAR HOSSEINI

SUPERVISED BY PROFESSOR HOWARD BOWMAN

75806 words over 286 pages

March 2022

Acknowledgements

This thesis would only consist of the current section if I would have individually thanked every person assisting me with it and with life in general. Since you are reading this, it is quite likely that you are one of them, so thank you. Still, I would like to highlight a few key people.

First, I am extremely grateful to Howard Bowman for being the kind of supervisor starting PhD students hope for. I particularly appreciate moments in which Howard encouraged me to pursue paths that were scary at first (and full of equations). The fact that I was successful with these paths has to a large extent been thanks to Howard's time, guidance and mentorship. I appreciate the School of Computing at the University of Kent for providing me with everything needed to complete this dissertation. I am grateful for collaborating with the bright minds of Srivas Chennu, Elkan Akyürek, Martin Eimer & Alon Zivony. I would like to thank Brad Wyble in particular for stimulating meetings, colourful Word documents, his patience with rerunning our simulations, and the paper that came out of all this. I also appreciate Howard, Brad, Srivas & Patrick Craston for developing and working on ST^2 models. Even though I might have shouted at the code at times, our model allowed me to foster my programming skills, develop a computational way of thinking and appreciate modelling as a whole. I would also like to thank my colleagues George Parish & William Jones for invaluable assistance with coding during my first years in Kent.

Further, I am grateful to my wonderful grandparents for their unconditional love. It made me. I thank my mother for life itself, inspiring conversations, and her endless support. I thank my father for teaching me to be my authentic self. I thank Human & Markus for allowing me to experience having brothers as well as an only child can. I thank Marek for being the kind soul that he is and for always being there. Special thanks go to my mates from Brühl for always ensuring that my feet are on the ground. I appreciate Zadaf for being a true friend, challenging my views and inspiring me to leave my zone of comfort. I thank Lilli for mutual growth and profound times. I appreciate Natti, Manar & Daniel for a wonderful home, for patience, and for always being empathetic. I am also deeply grateful to all the lovely people I met in Ehrenfeld, Groningen, Istanbul, Maastricht & the UK for making the last 10 years of study unforgettable. Special thanks go to Seve, Basti, Henk, Theresa, Rene, Ka, Stelle, Jakub, Amelia & Jesper.

One person that I will always particularly appreciate is Yvy Kreckel, who not only taught me the language that has allowed me to pursue my ambitions, but who believed in me to do so, being encouraging and supporting during a critical phase of my life.

And, finally, to Karl who has been watching me from afar, thank you for giving me strength whenever I need it the most, for understanding me in your unique way, and for many moments of inspiration and distraction. I will be forever grateful.

Abstract

Humans generally perceive the external world in a coherent manner. This perceptual coherence exists in most individuals despite the supposedly overwhelming number of individual bits of perceptual information available at any given moment. A substantial body of scientific work has been dedicated to understanding the cognitive reasons underlying the coherence of human perception. A key notion in this context was put forth almost a century ago by Gestalt psychologists (Von Ehrenfels, 1937; Wagemans et al., 2012), stating that humans group and categorise sensory input into objects because the human mind is dispositioned to perceive *patterns*. Grouping sensory input into distinct objects is achieved via *perceptual integration*. Integration can occur either temporally, spatially, as well as multimodally, allowing us to perceive the world as trees, humans, and melodies, thereby preventing sensory overload. The study of human perception often analyses difficult perceptual tasks in which processes, such as integration, regularly fail. The idea underlying this approach is that insight about the circumstances leading a system, such as that of human perception, to reach its limits and fail simultaneously informs about how the system functions when it functions well.

This thesis will study the limits of perceptual integration in the time domain. We will specifically analyse two cases in which visual stimuli presented in rapid sequence are integrated erroneously, leading to perceptions that were not presented as such in the physical world. The distractor intrusion phenomenon, in which multi-dimensional stimulus features are bound together wrongly, will be investigated first. The 2-feature Simultaneous Type/ Serial Token neural model (2f-ST^2 model) will be presented to account for distractor intrusions across a range of different empirical paradigms studied in humans. We will provide behavioural as well as virtual electrophysiological (EEG) results generated by the 2f-ST^2 model that qualitatively match those found with humans, providing evidence in favour of the 2f-ST^2 model's validity. Besides, we will provide a series of empirical analyses that further elucidate the cognitive mechanisms underlying distractor intrusions. In essence, these results suggest that whether integration occurs correctly or erroneously depends largely on the timing with which transient attentional enhancement (TAE) impacts relevant cognitive processes. Temporal event integration describes the cognitive process that binds two rapidly and successively presented visual stimuli into a single percept, given that there is a perceptually meaningful way of doing so. We analysed temporal event integration applying a machine learning approach to EEG data. Our results replicate previous findings on a whole-brain basis

and further suggest that temporal event integration occurs about 300 – 450 ms after stimuli were presented specifically, and, more generally, with characteristics that are in line with those of distractor intrusions and the dynamics proposed by the 2f-ST^2 model. We finally provide a series of simulation analyses that demonstrate the easily observable and dire methodological risk of *overhyping* when applying machine learning algorithms to neuroimaging datasets, such as those adopted by ourselves when investigating temporal event integration. Overhyping describes that choosing classification hyperparameters based on which generate the most desirable results (e.g., maximal classification accuracy) can threaten the external validity (i.e., generalisability) of an analysis if the necessary precautions are not taken. In this context we demonstrate that overhyping can lead to classifiers generating spurious above chance-level accuracies even though no signal was present in the data before providing effective safeguards that limit the risk of overhyping.

Table of Contents

Chapter 1 – Introduction

Humans perceive the world through the (basic) sensory modalities of light, sound, taste, temperature, pressure and smell. Each modality's sensory information has distinct physical properties as well as a corresponding sensory organ that receives and processes incoming information. Examples are light being electromagnetic radiation perceivable by the human eye or sound being pressure oscillations transmitted through a medium, such as air, perceivable by our ears. After the initial reception and processing of sensory information by our sensory organs, the information propagates to our brain, where each modality has their own specialised cortical areas that interpret the information received. The result of our brain's interpretation of the entirety of sensory information at any given moment constitutes our perception of that moment. The transformation of the totality of incoming sensory information to a unified, coherent percept is complex and affected by many factors, such as our experience, our motivation or intent, or our expectation of future sensory input. For example, sensory expectation could mean that we expect to see a lion quite soon after we heard it roar or hear thunder soon after we see lightning. In addition, the efficacy of our sensory organs and the brain areas responsible for processing sensory information not only varies between different humans (e.g., visual search strategies being based on individual preferences (Hogeboom & van Leeuwen, 1997)), but also over time for each individual person. Aging shapes perception substantially. In the case of hearing, for example, it has been established that humans hear high frequencies worse as they get older (Gordon-Salant, 2005). Due to the abundance of incoming sensory information and the complexity of the processes that generate a coherent percept, it can be argued that no single individual perceives the same sensory input *exactly* as another individual, despite the objective physical properties being identical. Even though humans likely never share their exact perception of any given moment or piece of information with one another, they do share a certain element of their perceptual processes: perceptual *coherence* (Handel, 2006).

Humans typically experience a sense of coherence when perceiving the world. This means that the sensory information they receive (bottom-up) typically matches what they expect (top-down) based on past experiences. Perceptual coherence is one of the features of the human perceptual system that allow us to process a multitude of single bits of sensory information every moment without feeling overwhelmed. Constancy is another feature of human perception, which allows us to recognize the same object as we receive information about it from different sensory inputs. For example, it allows us to understand that an animal

that is running towards us is not becoming larger, even though it is taking up more of our visual field. Another feature of human perception is perceptual contrast. Perceptual contrast states that the context in which we perceive an object has an impact on our perception of that object. Concretely, if a perceived object is extreme in a certain dimension, then neighbouring objects are perceived to be less extreme in that dimension. For example, a person standing next to a very tall person will be perceived as being shorter than if that same person would stand next to another person of similar height (Plous, 1993). Another contextual factor affecting perception is congruence. For example, van Leeuwen and Lachmann (2004) suggested that the (in)congruence of surrounding shapes affected the efficacy of processing shape stimuli, but not of processing letter stimuli. Consistent with the latter notion of congruence, Gestalt psychologists (Von Ehrenfels, 1937; Wagemans et al., 2012) proposed that we generally group and categorise sensory input into objects. According to Gestalt psychology, sensory information is grouped because the human mind is dispositioned to perceive *patterns*. Further, it is stated that this grouping of multiple sensory stimuli into a single object occurs according to a set of qualities: proximity (proximal stimuli are grouped together), similarity (physically similar stimuli are grouped together), closure (the tendency to perceive complete shapes, even though these might be incomplete), good continuation (overlap between objects is perceived as each object being uninterrupted), common fate (objects that are moving similarly are grouped together), and good form (objects with similar shape, colour or patterns are grouped together) (Von Ehrenfels, 1937; Wagemans et al., 2012).

Gestalt properties, stating the rules according to which human perception tends to group stimuli together, are a good example of the core perceptual process of *integration*. Perceptual integration allows us to perceive the world as buildings, people, and melodies, preventing sensory overload. Perceptual integration according to Gestalt properties can, however, be ambiguous at times. An example of perceptual ambiguity is provided in Figure 1 (adopted from Nicholls et al. (2018)), which shows one of the most widely known visual illusions. Titled *My Wife and My Mother-In-Law*, and drawn by cartoonist W. E. Hill in 1915, the drawing can either be seen as a young woman looking away from the viewer or a portrait of an old lady (note that if one should struggle with seeing both cases it helps to view the young lady's chin as the tip of the old lady's nose or vice versa). It should be stressed that cases of ambiguity concerning what is shown to our visual system, such as the case in Figure 1, occur despite the fact that the sensory input is fixed. That is, even though only a fixed set of

black and white areas and lines is presented, our visual system is able to see two different things solely based on how it groups together the provided visual stimuli into a single object.

Figure 1. My Wife and My Mother-In-Law drawing of cartoonist W. E. Hill, showing one of the most popular cases of a visual illusion based on perceptual ambiguity, in this case, emerging from visual stimuli being grouped together according to Gestalt properties. Despite the visual stimuli remaining fixed, our visual system is able to either see a young lady looking away from the viewer or a portrait of an old lady. Adopted from Nicholls et al. (2018).

Multimodal integration is another important phenomenon that enables perceptual coherence. Multimodal integration refers to the process with which information coming from different sensory modalities are integrated together, enabling us to perceive objects and events as single, coherent percepts. For example, somebody playing basketball knows that the bouncing sound they hear when bouncing the ball is due to the ball hitting the ground. Or, when observing a singer singing a song, we understand that the source of the auditory information (i.e., the melody) that reaches us is the singer whose mouth we can see move. Event integration describes the form of perceptual integration most relevant for this thesis. Event integration typically refers to *temporal* integration, i.e., separate events that occur at proximal, but separate times are integrated into the same event. This is in contrast to the ambiguous case of perceptual integration discussed in the context of Figure 1, since the visual illusion presented there is independent of any temporal aspect. However, our previous example of seeing lightning and anticipating thunder is a case of event integration, since *after*

hearing the anticipated thunder, we understand that its source was the *previously* seen lightning. We thus integrate the visual event of lightning with the auditory event of thunder into a single, coherent event.

The scientific study of perception regularly involves pushing our perceptual system to its limits. Pushing perception to a point at which its coherence breaks down due to presented stimuli being too complex, fast, or generally challenging, allows researchers to not only gain insight about *when* (i.e., under which conditions) the system breaks down, but, importantly, also about *how* (i.e., in which way) the system breaks down. Insight about how our perceptual system breaks down informs us about the system itself and, critically, about how it works when it functions correctly.

The Simultaneous Type/ Serial Token (STST) model is an influential theory of perception and attention, which has successfully explained a large set of perceptual phenomena and electrophysiological correlates, particularly those associated with the temporal deployment of attention (e.g., (Bowman et al., 2008; Bowman & Wyble, 2007; Chennu et al., 2009; Craston et al., 2008; Jones et al., 2020; Wyble et al., 2009, 2011)). An objective of this thesis is to attempt to provide further evidence for the STST theory by modelling temporal event integration findings with it.

In the context of visual perception, one classic way of pushing the system to its limits in order to study it is to use streams of rapidly presented visual stimuli. Rapid serial visual presentation (RSVP) experiments adopt such rapid stimulus streams, with visual stimuli being presented at the same spatial location at typical rates of 10-15 items / second. This thesis investigates perceptual event integration in the context of RSVP experiments with the particular aim of providing new insight about the neuronal mechanisms and processes underlying two cases in which our perceptual system reaches its limits in such experiments. The first case is temporal event integration, according to which two visual events (i.e., stimuli) that are presented in immediate succession, but are also combinable in a perceptually meaningful way according to Gestalt properties (Von Ehrenfels, 1937; Wagemans et al., 2012), are integrated into a single perceptual event. The second case of visual perception reaching its limits studied in this thesis is the distractor intrusion phenomenon. Distractor intrusions refer to the erroneous binding of multi-dimensional stimulus features into single percepts. We will model these phenomena using the STST framework, enabling us to acquire further evidence for the mechanisms included in the framework. More specifically, this will involve presenting an extension of the Simultaneous Type/ Serial Token computational model

(Bowman & Wyble, 2007) to account for the distractor intrusion phenomenon. This will be based upon the 2-feature Simultaneous Type, Serial Token (2f-ST^2), first introduced in Chennu et al. (2011). In this context, we will contrast our modelling work to the model provided by Botella et al. (2001), which posited that correct reports and distractor intrusion errors result from two separate processing routes. A main contribution of this thesis will be that a more parsimonious explanation will be provided with the 2f-ST^2 model, which will explain the key phenomena of related empirical paradigms without the necessity of treating correct reports and intrusion errors in a different manner.

Pursuing the aim of providing new insight about these two phenomena of perceptual event integration, we will apply machine learning algorithms to neuroimaging datasets (in addition to methodological work concerning a specific risk when doing so).

Core Hypotheses

Specifically, we will investigate the neuronal mechanisms and processes underlying event integration in rapid stimulus (i.e., RSVP) streams. For this, we will focus on the following core hypotheses, which will be referred back to throughout the thesis.

1 During rapid stimulus streams, the binding of multi-dimensional stimulus features into one percept depends on the timing of transient attentional enhancement (TAE).
2 The 2-feature Simultaneous Type, Serial Token (2f-ST^2) model accounts for a broad range of findings obtained with distractor intrusion experiments, suggesting that correct and intrusion reports are *qualitatively* the same, implying an interesting contrast to the dual-route model of Botella et al. (2001). The 2f-ST^2 model accounts for distractor intrusions via the following.
    a. The 2f-ST^2 model generates behavioural response distributions matching those obtained with various distractor intrusion experiments.
    b. The 2f-ST^2 model's blaster circuit's postsynaptic activation (i.e., its virtual N2pc) dynamics replicate electrophysiological findings obtained with the human N2pc ERP component.
    c. 2f-ST^2 virtual P3 dynamics replicate electrophysiological findings obtained with the human P3 ERP component.
3 Increased key feature salience induces less temporally variable TAE deployment in human cognition, which can further be computationally modelled with the 2f-ST^2 model.

4 The temporal integration of two combinable stimuli into a single perceptual event occurs with temporal and electrophysiological characteristics that are consistent with the distractor intrusion phenomenon and the 2f-ST^2 computational model.

5 Machine learning algorithms applied to neuroimaging datasets, particularly temporal generalisation analysis used to determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfitting to hyperparameters) classification results if appropriate preventive measures are not implemented.

Chapter 2 - Literature Review

The present literature review will provide an overview of the neuroscientific developments pertaining to selective attention and working-memory (WM) encoding. In particular, it will focus on rapid serial visual presentation (RSVP) paradigms and the attentional blink (AB) phenomenon, which was observed therein. This focus arises for two reasons. First, the two cases of event integration studied in this thesis, temporal event integration and distractor intrusions, both originated from RSVP paradigms. Moreover, the computational model we will present in Chapters 4-6 is an extension of a computational model that was initially developed to account for a variety of findings associated with the AB. Therefore, we will adopt the following structure in the present literature review. The review will commence with an introduction to RSVP paradigms and the AB in particular, providing a brief overview of the chronology with which the AB was observed, the core experimental findings that were central since its discovery as well as the theoretical as well as computational models that attempted to explain the cognitive mechanisms underlying it. We will subsequently introduce temporal event integration and the distractor intrusion phenomenon, which are the two cognitive phenomena that will be studied in the present thesis. The literature review will finally conclude with an introduction to a methodological approach central to the present thesis: multivariate pattern analyses (MVPA). In this context, we will focus on classification-based approaches to MVPA, particularly the temporal generalisation method (King & Dehaene, 2014), as the latter will be adopted in Chapters 7 and 8.

The Attentional Blink (AB)

Over the last decades, the AB has been studied extensively in the context of RSVP paradigms, which gave rise to multifaceted insight about human attentional processing,

conscious awareness, and WM encoding. Notably, it was almost 50 years ago when researchers first presented a sequence of visual stimuli in a rapid manner to study human memory (Potter & Levy, 1969). About two decades later, the earliest observations of an AB emerged from variants of an RSVP paradigm (Broadbent & Broadbent, 1987; Raymond et al., 1992). The AB's experimental paradigm was altered in countless different ways since, but its essence remained the same: The subject is presented with a stream of visual stimuli, shown one at a time. This stream predominantly contains distractor stimuli in addition to a few (classically two) target stimuli. The subject's task is to identify and report both target stimuli after each trial concludes. A classic example of this paradigm using distractor letters and target digits is illustrated in Figure 2a (adopted from Martens and Wyble (2010)). The 'blinking' of our attentional system, which gave the AB its name, is revealed when plotting the behavioural results of such an experiment, as can be seen in Figure 2b. Here, the average proportion of correct responses is plotted as a function of between-target lag for the first (T1) and the second (usually conditioned on correct T1 identification, i.e., T2|T1) target separately. Lag in this context describes the position of the second target with respect to the first, e.g., appearing right after it at lag 1.

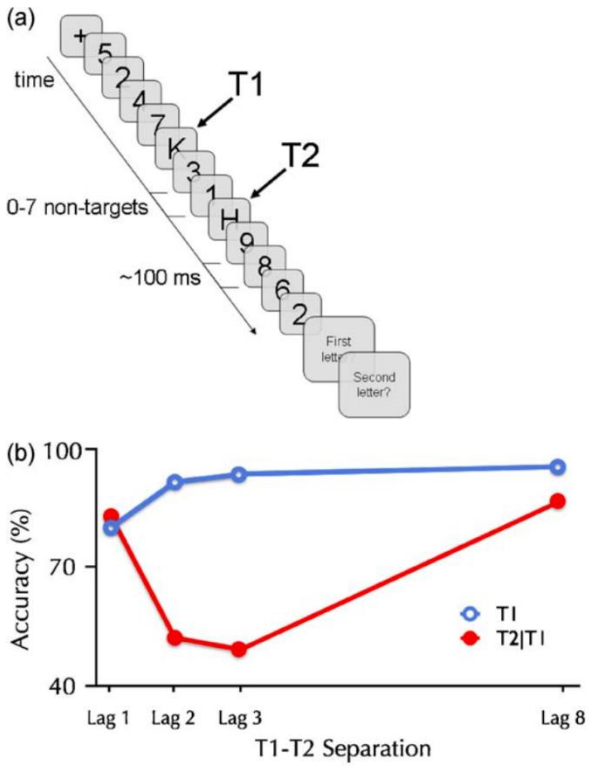


Figure 2. A traditional RSVP paradigm and AB effect. Panel (a) depicts an example trial. The trial unfolds from top to bottom, starting with a fixation stimulus and ending with response screens. The RSVP stream is comprised of distractor numbers and target letters. Panel (b) illustrates the attentional blink. The proportions of correctly identified first targets (T1) and second targets conditioned on correct T1 response (T2|T1) are plotted as a function of lag between targets. T2 performance is at T1 level at lag 1, drops strongly for later lags, being worst at lag 3 and slowly increases until almost being at T1 level again at lag 8. Adopted from Martens and Wyble (2010).

It is apparent from such plots that we are very skilled at identifying the first target, irrespective of when the second target follows thereafter. However, this is strikingly different for the second target. Identification of the second target in such paradigms is heavily dependent on its position in the RSVP stream with respect to the first target. Given that the first target was identified, it is easy for people to identify the second target as well if it follows the first immediately (lag 1). Importantly, T2 performance drops substantially when distractors are presented prior to the second target, which is referred to as the 'blink'. This blink is typically most pronounced in the case of two distractors being shown between the first and second target (lag 3). T2 performance at lags 2 and 4 is usually comparable, being slightly better than at lag 3. For later lags (5 – 7, 'post blink'), T2 performance gets increasingly better until it reaches the level of T1 identification at the latest lags (8) (Martens & Wyble, 2010).

After the earliest findings of the AB (Broadbent & Broadbent, 1987; Raymond et al., 1992), research was initially concentrated on understanding how it occurs, e.g. to what extent an unreportable T2 is still processed in the brain. In this context, behavioural (Maki et al., 1997; Potter et al., 2005; Visser et al., 2005) as well as neuroimaging (Luck et al., 1996; Marois et al., 2004; Nieuwenhuis et al., 2005; Sergent et al., 2005) evidence generated a consensus that regardless of whether a stimulus is reported or not, it is processed up to a conceptual level and it is a failure of consolidating T2 into WM in a retrievable form that causes the blink. Theories describing the latter (Chun & Potter, 1995; Dehaene et al., 2003; Isaak et al., 1999; Jolicœur & Dell'Acqua, 1999) defined this as central capacity limitations and were, for example, supported by priming studies (Martens et al., 2002; K. Shapiro et al., 1997), which have shown that a 'blinked' T2 representation can still affect related cognitive processes such as semantic processing. Theories of central capacity limitations were further supported by cases of the AB emerging across modalities (Arnell & Jenkins, 2004; Arnell & Jolicœur, 1999; Arnell & Larson, 2002), spatial locations (Duncan et al., 1994) and types of target identities (Raymond et al., 1992).

Neuroimaging studies have focussed on using electroencephalography (EEG) to study the AB as EEG signals carry high temporal resolution. Overall, EEG findings supported the conclusion gained from behavioural studies that the AB occurs in a later stage of processing and is due to a deficiency of encoding T2 into WM. This was mainly based on the analyses of event-related potentials (ERPs), which are EEG signals time-locked to an event of interest (e.g., T2 onset). These are then either averaged across trials (within-subject ERP) or, more

commonly, a second-level average is computed of all within-subject ERPs, which yields an across-subject ERP (also called the Grand Average). Moreover, numerous distinct ERP components (deflections from zero at a specific time-interval) are believed to correspond to certain cognitive processes. The investigation of four such ERP components has enabled scholars to understand the temporal unfolding of the AB. The P1 and N1 components, correlating to early perceptual processing and being measured until approximately 150 ms after stimulus onset, do not show any irregularity in response to an unreported, or 'blinked', T2 (Vogel et al., 1998). The first ERP component that is modulated by the AB is the N2pc, which occurs slightly later (≈ 200 ms post-stimulus) and reflects the allocation of attentional resources required for processing a target stimulus (Eimer, 1996; Kiss et al., 2008). In this context, a lag-dependent pattern was observed: T2s that follow T1s after a long lag do evoke N2pcs, whereas short-lag (likely to be blinked) T2s fail to do so (Martens & Wyble, 2010). The P3, one of the most intensely studied ERP components in neuroscience, indexes WM consolidation and is suppressed for T2s presented during the AB (Luck et al., 1996; Vogel et al., 1998), again supporting the claim that blinked T2s occur due to a failure of consolidation into WM. However, and in accordance with the priming studies introduced above, the N400 component, which (like the P3) occurs rather late and corresponds to semantic processing, is evoked in response to blinked T2s (Luck et al., 1996; Rolke et al., 2001; Vogel et al., 1998).

Models of the Attentional Blink

Since its discovery, several models were formulated in an attempt to explain the AB (for a review, see Dux and Marois (2009)). Early AB models, such as the Gating Theory (Raymond et al., 1992), the Interference Theory (K. L. Shapiro et al., 1994), bottleneck models (Chun & Potter, 1995), the temporary loss of control (TLC) hypothesis (Di Lollo et al., 2005), the delayed attentional reengagement account (Nieuwenstein, 2006) or hybrid models (Kawahara et al., 2006; Vogel et al., 1998) were informal. However, from the early 2000s, computational accounts have emerged. Examples of such are the gated auto-associator model (Chartier et al., 2004), the corollary discharge of attention movement model (Fragopanagos et al., 2005), the locus coeruleus-norepinephrine model (Nieuwenhuis et al., 2005), the boost and bounce theory (Olivers & Meeter, 2008), the (episodic) simultaneous type, serial token model (Bowman & Wyble, 2007; Wyble et al., 2009), the attention cascade model (Shih, 2008) and the threaded cognition model (Taatgen et al., 2009). The discussion at hand will explicitly introduce some key theories and models before providing an in-depth introduction to the Simultaneous Type, Serial Token (STST, (Bowman & Wyble, 2007)

model and its extension, the episodic STST (eSTST, (Wyble et al., 2009)). The reason for emphasizing the STST (Bowman & Wyble, 2007) and eSTST (Wyble et al., 2009) models in the current review is twofold. First, it has been argued elsewhere that these models, along with the attention cascade model (Shih, 2008), are able to account for the 'largest number of empirical findings, since they incorporate capacity-limited T1 processing' (Dux & Marois, 2009, p. 1697). More importantly, the current thesis aims will present an extension of the original STST model (Bowman & Wyble, 2007 to account for the distractor intrusion phenomenon in Chapters 4-6.

Global Workspace Models and the Global Neuronal Workspace (GNW) (Dehaene & Changeux, 2011)

Baars (1988, 1997, 2002) provided a Global Workspace Theory (GWT) to account for human consciousness. Baars (1997) compared the neural interactions that give rise to consciousness to a Theatre play. According to this initial GWT, selective attention shines a spotlight on those external stimuli and internal processes that access our conscious content, much like the actors who act out scenes on the Theatre's stage. Scenes that are performed on the stage are then broadcast globally throughout the cortex, similar to the Theatre's audience watching the play. Moreover, behind the scenes of the Theatre are those unconscious neural processes that despite shaping the scene that is being acted out on stage, themselves remain unseen (and, thus, unconscious). Baars' (1988, 1997, 2002) initial work opened the path for further studies and extensions to the GWT, such as those of Franklin and Graesser (1999), Shanahan (2006), and Bao et al. (2020). However, the arguably most influential extension of Baars' (1988, 1997, 2002) GWT was provided by Dehaene and colleagues (Dehaene et al., 1998, 2003; Dehaene & Changeux, 2011).

Dehaene et al. (1998) introduced their expansion of Baars' GWT (1988, 1997, 2002) in the context of effortful cognitive tasks, emphasizing that their model should not be viewed as exhaustively modelling consciousness (Dehaene et al., 1998). Dehaene et al.'s (1998) model introduces two computational spaces in the brain. The first computational space is comprised by interconnected and distant neurons with long axons that constitute a unique global workspace. The secondary computational space is comprised by modular processors that are specialised to some domain of cognition, such as attention, perception, motion, etc. (Dehaene et al., 1998). Dehaene and colleagues (Dehaene et al., 2003; Dehaene & Changeux, 2011) further extended the GWT, linking subjective conscious experience to measurable neuronal activity, and formulated the Global Neuronal Workspace (GNW) model. Contrasting

conscious versus unconscious processing, the authors (Dehaene et al., 2003; Dehaene & Changeux, 2011) postulated that conscious access to a piece of information is the result of the information being broadcast to multiple brain systems (i.e., the modular processors mentioned before) via neuronal networks that have long-ranging axons. The former were modelled as nodes corresponding to thalamocortical columns (interconnected thalamic and cortical neuronal networks), whereas the latter were proposed to be in prefrontal, parieto-temporal, and cingulate cortices (Dehaene et al., 2003; Dehaene & Changeux, 2011). Moreover, the modular brain systems, i.e., specialised neuronal networks, besides *receiving* information from the global workspace, also process respective pieces of information in a parallel manner. Which piece of information will be broadcast to the global workspace is determined by a winner-takes-all competition, which is modulated by top-down attentional mechanisms (Dehaene et al., 2003; Dehaene & Changeux, 2011). Applied to the AB, the GNW model therefore posits that the globally broadcast activation pattern initiated by the T1 (after winning the competition to conscious access) blocks entry of the T2 to the global workspace for about 200 ms. Whilst acknowledging that the GNW provides a compelling framework for how conscious contents enter subjective experience, the GNW has initially been deemed lacking as an explanation for the AB, particularly since the notion of the T2 being blocked access to the global workspace for ~200 ms after presentation of the T1 contradicts the finding of lag 1 sparing (Martens & Wyble, 2010). However, Simione et al. (2012) presented a neurodynamic model for Visual Selection and Awareness (ViSA), which can be considered an extension of earlier GW models (Dehaene et al., 2003; Dehaene & Changeux, 2011). The ViSA model is comprised by two perceptual processing modules and two access control modules and accounts for a wide range of empirical findings, such as the limited storage capacity of visuo-spatial WM, attentional cueing, and the AB (Simione et al., 2012). Importantly, the ViSA model demonstrates lag-1 sparing when replicating RSVP experiments provoking the AB, thus resolving this limitation of earlier GW models (Dehaene et al., 2003; Dehaene & Changeux, 2011).

The Boost and Bounce Model (Olivers & Meeter, 2008)

The next computational model we consider is the Boost and Bounce Model presented by Olivers and Meeter (2008). According to this model, capacity limitations or a bottleneck property of the human attentional system is not central for explaining the AB. Instead, the Boost and Bounce Model (Olivers & Meeter, 2008) proposes a critical role of an attentional filter that depends on the task-relevance of stimuli. This filter enhances relevant and inhibits irrelevant information. Applied to the AB, the first, task-relevant, target causes this filter to

boost relevant cognitive processes, which leads to the target being encoded into WM. The *boost* was hypothesised to be provided by neuronal networks in the prefrontal cortex and the basal ganglia. Further, to ensure that task-irrelevant stimuli are not encoded into WM, the *boost* is followed by strong inhibition, i.e., a *bounce*, whenever a target-following distractor stimulus received an erroneous *boost* due to the target's presentation. This *bounce* thus prevents distractor stimuli from entering WM. T2s presented during the AB therefore fail to be encoded into WM because they are presented in the time-interval of the *bounce* inhibiting activation levels. Importantly, lag 1 sparing occurs because the T1 and the T2 are presented in immediate succession and no *bounce* has been initiated since no distractor was presented after the T1. Therefore, both target-stimuli are enhanced by the *boost* and successfully encoded into WM (Olivers & Meeter, 2008).

The Boost and Bounce Model (Olivers & Meeter, 2008) accounts for a number of empirical findings pertaining to the AB. Particularly the spreading the sparing finding (Olivers et al., 2007), in which task-performance remained high after the introduction of more than two subsequently presented RSVP target-stimuli, was challenging for capacity-limited theories of the AB and instead provided evidence in favour of the Boost and Bounce Model (Olivers & Meeter, 2008). However, it has been argued that it is still to be decided whether the model can explain the finding that distraction eliminates the AB without impacting T1 performance (Martens & Wyble, 2010). It has further been demonstrated that the AB can arise without a distractor following the T1 (Nieuwenstein et al., 2009), which can be seen as evidence against the Boost and Bounce Model (Olivers & Meeter, 2008), since this model would require a distractor to initiate the bounce.

The Threaded Cognition Model (Taatgen et al., 2009)

Taatgen et al. (2009) provided a computational model based on the threaded cognition theory of multi-tasking to explain the AB. According to the Threaded Cognition Model (Taatgen et al., 2009), the AB arises due to an overexertion of cognitive control. The main idea is that separate cognitive resources can be used in parallel for different tasks (threads). However, only one resource can be used at a time for a given thread. Importantly, threads compete for resources and, if not controlled by our cognitive system, the first thread that requires a resource will use that resource as soon as it is available. To prevent such unsystematic allocations of resources to threads, an overzealous attentional control mechanism regulates which resources are allocated to which threads. In the context of the AB, this overzealous attentional control mechanism uses a production rule that blocks target

detection during memory consolidation. The authors stress that the production rule itself is distinct from target detection and memory consolidation (Taatgen et al., 2009). According to the Threaded Cognition Model (Taatgen et al., 2009), the AB occurs because the T2 is presented at a time at which target detection is blocked because the T1 is being encoded into WM. This notion is conceptually similar to that provided by the Boost and Bounce Model (Olivers & Meeter, 2008) and was initially proposed within the Simultaneous Type/ Serial Token framework (Bowman & Wyble, 2007). However, the underlying mechanisms between the two models differ. Whereas the T2 is typically not encoded during the AB because of strong inhibition in the Boost and Bounce Model (Olivers & Meeter, 2008), 'blinked' T2s are due to the overzealous nature of the attentional control mechanism, blocking target detection during memory consolidation in the Threaded Cognition Model (Taatgen et al., 2009). Furthermore, the Threaded Cognition Model (Taatgen et al., 2009) states that, despite being useful in other contexts, the overzealous attentional control mechanism is unnecessary in RSVP experiments, since the T1 would be encoded into WM without such a mechanism, too. Importantly, the Threaded Cognition Model (Taatgen et al., 2009) explained empirical findings that demonstrated that the AB can be weakened by distraction, the reason being that the overzealous attentional control that leads to the AB is to some extent utilised elsewhere, e.g., for a secondary cognitive task (Martens & Wyble, 2010).

The Attention Cascade Model (Shih, 2008)

Shih (2008) presented the Attention Cascade Model, a mathematical model implementing multiple processing stages interacting with one another, to account for the AB. Each processing stage is formalised by a mathematical function and Shih (2008) stresses that these functions potentially might, but do not necessarily have to, reflect the activation of a population of neurons. Thus, the author abstains from providing a physiological description of how exactly the processing stages are implemented in the human brain (Shih, 2008). A given stimulus can be processed along two pathways in this model, one being mandatory and the other depending on the stimulus's bottom-up salience (Shih, 2008). The mandatory pathway activates the stimulus's long-term memory (LTM) trace, which leads to a preliminary representation of the stimulus being temporally available in a peripheral buffer. This notion is similar to the conceptual short-term memory presented in Potter (1993) and the item layer of the Simultaneous Type, Serial Token Model (Bowman & Wyble, 2007). The preliminary representation is transferred to the model's WM buffer. The extent of stimulus-representation (i.e., information) that is transferred depends on the onset and duration of an *attention window*. The attention window is part of the attention control mechanism, which constitutes

the model's most central part (Shih, 2008). The attention window can be initiated in one of two ways. The first way a stimulus can initiate the attention window is if its preliminary representation exceeds a criterion of top-down salience. This criterion is predetermined in the model's attention control mechanism and is contingent on stimulus characteristics, task demands and instructions. The second way in which a stimulus may trigger the attention window is if its bottom-up salience is sufficiently strong, meaning that the degree of physical difference between a given stimulus and the preceding stimuli is substantial. Further, the width of the attention window depends on an RSVP stream's presentation rate, task demands, and the window's inherent limitations (Shih, 2008). Stimuli representations subsequently enter the WM buffer and can be directly reported if they are stronger than a response threshold. If the latter should not be the case, stimuli representations enter a consolidation processor and from there a decision stage. Depending on whether representations are stronger than noise or not, either that stimulus's response is forwarded to the WM response buffer, and ultimately provided as the response, or a guess is made (Shih, 2008). Despite the fact that the Attention Cascade Model does not provide a physiological implementation of processes, it is able to account for a variety of empirical findings, such as lag 1 sparing, intrusion errors, stimulus competition, magnitude of the AB dip, target+1 blank, and stimulus salience (Shih, 2008). This model further, and importantly, is able to account for capacity-limited T1-processing, something that most other models (except the (e)STST models (Bowman & Wyble, 2007; Wyble et al., 2009)) are unable to (Dux & Marois, 2009).

The Locus Coeruleus Model (Nieuwenhuis et al., 2005)

Nieuwenhuis et al. (2005) provided a neurocomputational model to account for the AB that emphasises the significance of an attentional mechanism in the brainstem. Concretely, the authors presented the activation dynamics of a small brainstem nucleus, the locus coeruleus (LC), which was shown to regulate attentional deployment via noradrenergic projections to the cortex. Crucially, after the presentation of a visual target stimulus, LC neurons fire a strong burst, followed by a refractory period. This refractory period lasts several hundred milliseconds and is induced by noradrenaline/norepinephrine release. In their locus coeruleus norepinephrine (LC-NE) computational model, Nieuwenhuis et al. (2005) propose that an RSVP stream's T1 stimulus could initiate LC neurons to fire. Moreover, the blink itself, i.e., diminished T2 detection performance at intermediate lags, is argued to be due to the fact that the T2 is presented during the LC's refractory period, which commences about 200 ms after T1 onset. According to this model, the AB is due to the T2's task-relevance being unable to trigger a secondary activation-burst of LC neurons, since they are inhibited by

their refractory period. Critically, the model explains lag 1 sparing by the fact that T2 is presented *before* the LC's refractory period occurs and, thus, T2 processing benefits from the LC's burst provoked by T1 onset, leading to both stimuli being perceived (Nieuwenhuis et al., 2005). In general, the time-course of the LC's behaviour upon target detection, particularly due to its refractory period, and the AB show a striking overlap.

However, following their initial presentation of their computational model, the authors themselves tested a critical hypothesis about the relationship between the LC and the AB (Nieuwenhuis et al., 2007), failing to provide evidence in favour of a central prediction of their model. The tested hypothesis specifically predicted that if the LC's refractory period deteriorates T2 performance during the AB via noradrenergic mechanisms, then applying an adrenergic agent should visibly impact the neuronal mechanisms leading to the AB. To test this hypothesis, Nieuwenhuis et al. (2007) conducted an experiment in which clonidine, an adrenergic agent, or a placebo were administered and participants performed an AB as well as a visual search task. Even though clonidine was found to affect some cognitive mechanisms, such as increasing array search times or decreasing T1 performance, clonidine administration did not affect the AB's time-course (Nieuwenhuis et al., 2007). Moreover, Bowman et al. (2008) linked Nieuwenhuis et al.'s (2005) LC-NE model to the Simultaneous Type, Serial Token framework (Bowman & Wyble, 2007). In this context, several limitations of the LC-NE model have been suggested (Bowman et al., 2008), such as the model not generating blink attenuation with T1 +1 blank or spreading the sparing (Olivers et al., 2007). Nonetheless, the value of the findings with the LC-NE model (Nieuwenhuis et al., 2005) has been stressed, since they emphasised the role of temporal dynamics in attentional mechanisms, besides suggesting that salient cues initiate altered responses that likely are distinct from the mechanisms underlying the AB (Martens & Wyble, 2010).

The Simultaneous Type, Serial Token (STST) Model (Bowman & Wyble, 2007)

The STST model utilizes a types-token account to explain WM encoding of visual stimuli. Types correspond to features of items, whereas tokens carry information about their particular occurrence in time. The model architecture (Figure 3) can be summarized as follows. Stage 1 comprises the input and extraction of types. The central mechanism of the model is called tokenization and expresses the allocation of a token to a type in Stage 2, which results in successful WM encoding. Only one tokenization process can be active at a time. Selective attention impacts this process via a mechanism called the blaster. The blaster is triggered by salient items in Stage 1, which ultimately initiates tokenization. Importantly, if

a tokenization process is currently active, the blaster will be suppressed until it has finished. This is central for explaining the AB via the STST model as during the AB, T2 cannot be enhanced by the blaster due to T1's tokenization process still being active. Once it finishes, T2 often lacks sufficient activation to initiate its own tokenization, which results in an AB. Lag 1 sparing, which expresses superior performance in T2 accuracy at lag 1, occurs because T2 is presented while T1 is still being enhanced by the blaster, which leads to T2 obtaining sufficient activation to join T1's tokenization (Craston et al., 2008). This joint tokenisation of the T1 and the T2 is very relevant to the notion of event integration that is central to this thesis.

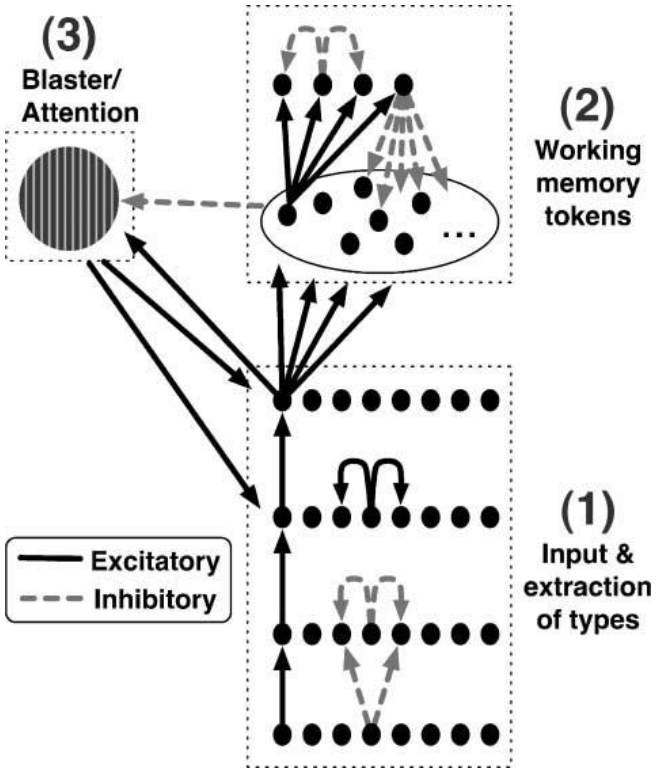


Figure 3. Architecture of the STST model. Stage 1: Input and extraction of types. Stage 2: WM encoding via tokenization. Stage 3: Temporal attention from the blaster. Adopted from Craston et al. (2008).

The STST model was extended with the episodic STST model (eSTST, Wyble et al., (2009)). The STST and eSTST models are much alike in terms of their overall architecture, both adopting Chun and Potter's (1995) framework that perception is a two-stage process encompassing object detection (via types) and the allocation of episodic information to those objects (via tokens). The explanation of lag 1 sparing changed rather drastically though, as the STST model's joint tokenization mechanism described above was abandoned in the eSTST model. It was replaced with T1 and T2's types competing simultaneously for the first token in the binding process. The winner of this competition is then bound to token one and the loser is encoded to token two (Wyble et al., 2009). This alternative definition was formulated in

reaction to the work of Olivers et al. (2007) introduced above, who showed that the AB can be spread to multiple targets as long as no distractors are shown between targets.

We will review the current state of neuroscientific literature pertaining to the first case of event integration in RSVP paradigms analysed by this thesis, the distractor intrusion phenomenon, next. In this context, human selective visual attention (as, for example, modelled via the blaster circuit in ST^2 models (Bowman & Wyble, 2007; Chennu et al., 2011; Wyble et al., 2009), plays an important role and will therefore be introduced in detail.

Distractor Intrusion Errors in RSVP Experiments

Selective visual attention describes the ability of our cognitive system to continuously select and focus on relevant information in our environment. Relevant information in this context can take the form of visual stimuli, for example. Such stimuli can be relevant in relation to a certain *task*. Selective attention plays a significant role in all kinds of tasks, as it allows us to focus our mind on task-relevant stimuli and to disregard irrelevant and distracting stimuli and events. An important concept in this context is that stimuli *compete* for our attention (Desimone & Duncan, 1995) and that competing stimuli may either be presented at the same time or sequentially in quick succession. The case of simultaneously presented competing stimuli has been studied extensively in *visual search experiments*. In such experiments, a target stimulus is presented simultaneously with multiple distractor stimuli and, hence, stimuli compete for our attentional capacity in the *spatial domain*. It has been shown that these tasks are easily solved when the target stimulus is very salient and different to the distractors, but these tasks become increasingly difficult to solve as target and distractor stimuli share more and more features (Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Wolfe, 2014). The difficulty that our attentional system faces in visual search tasks is that *the location* of the target stimulus is unknown before the search display is presented. This is less of a problem for very distinctive target stimuli as these can be identified easily once the search display is shown, which allows attentional resources to be deployed swiftly on to their location.

Selective Attention in the Time Domain and Distractor Intrusions

Visual search tasks study selective attention in the setting of competing stimuli being presented at *the same time, but at different spatial locations.* Another setting in which selective attention plays a crucial role arises when competing stimuli are presented at *the same spatial location, but at different times.* In the latter setting, selective attention is required to solve the issue of *temporal* uncertainty, which especially arises if stimuli are presented in

quick succession. Selective attention in the time domain is often studied with RSVP tasks, which were introduced in detail previously. RSVP tasks are particularly suitable for studying selective attention in the time domain, since attention is required to be placed on the correct (target) stimulus at the right time (when it is presented in the stream), because the *temporal position* of target stimuli is unknown prior to the onset of the RSVP stimulus-stream. RSVP experiments have studied selective attention in the time domain adopting numerous different experimental paradigms, varying parameters such as presentation rates (stimulus-onset-asynchronies, SOAs), the numbers of target stimuli within a trial (Olivers et al., 2007) or the stimuli themselves, using pictures (Potter et al., 2010; Potter & Fox, 2009) or even e-mail addresses (Harris et al., 2020) instead of the classically used target-letters and distractor-digits (Chun & Potter, 1995). RSVP experiments typically studied cases in which target- and distractor-stimuli involve distinctively different categories, as in the case of target-letters and distractor-digits.

An intriguing phenomenon arises, however, if target- and distractor-stimuli *share* the task-relevant, or response-, dimension, as is the case if, for example, target letters are embedded in a stream of distractor letters. In this case, it quickly becomes very challenging for selective attention to identify the target and filter out temporally adjacent distractor stimuli. Instead, subjects in such experiments regularly commit *distractor intrusion errors* and report the identity of a distractor stimulus presented in temporal proximity either before or after the target stimulus. In the context of RSVP experiments provoking distractor intrusion errors, the notion of key and response features is fundamental. Key features describe stimulus features that define targets as those meeting a task set. For example, in a coloured letters experiment, the instruction could be to report the identity of the red letter. Here, the key feature would be colour, as it is along this dimension that the differentiation between targets and distractors occurs. Response features, on the other hand, are those that need to be reported after each experimental trial, such as the red letter's identity in our previous example of coloured letters. The phenomenon of distractor intrusions has been demonstrated numerous times (Botella et al., 2001; Botella & Eriksen, 1992; Chun, 1997; Gathercole & Broadbent, 1984; Goodbourn & Holcombe, 2015; Intraub, 1985; Popple & Levi, 2007; Recht et al., 2019). Nonetheless, most theoretical and computational models of RSVP experiments (Bowman & Wyble, 2007; Dehaene et al., 2003; Fragopanagos et al., 2005; Olivers & Meeter, 2008; Shih, 2008; Taatgen et al., 2009; Wyble et al., 2009) do not account for the phenomenon of distractor intrusions. Nonetheless, three theoretical accounts are worth further

elaboration, as they provide valuable thoughts on how our attentional system might commit intrusion errors.

Theoretical Accounts of Distractor Intrusions

Vul and colleagues (Vul et al., 2009) empirically examined a number of alternative ways in which visual selective attention might operate. They specifically focussed on alternative accounts that would yield the pattern of across-trial variability often observed in selective attention experiments. Across-trial variability in this case describes the fact that repeatedly presenting the same experimental input to subjects does not lead to a deterministic pattern of the same responses being given. Distractor intrusions can hence be considered as a case of across-trial variability, as the same RSVP stream might lead to the correct report of a target stimulus in one trial, but to an intrusion error in another. The authors tested four possible modes in which attentional selection may be operating to produce across-trial variation: single-item selection, contiguous all-or-none selection, contiguous-graded selection and complex selection (Vul et al., 2009). These modes were tested in the time (Experiment 1) and space (Experiment 2) domain. The empirical approach implemented in Experiment 1 involved an RSVP experiment in which intrusion errors were possible and subjects were required to guess the identity of the target stimulus four times (the most likely, then the next most likely and so on) after each trial. Based on their data, Vul et al. (2009) proposed that selective attention operates via a contiguous-graded selection process. This account implies that once the correct key (also called selection) feature is detected, a time-window of selection is placed around the stimulus carrying the correct key feature (i.e., the target). Item representations of target as well as proximal distractors, which are presented during the postulated time-window of selection, are then temporarily stored in short-term memory. One of these items is subsequently sampled to be processed further, allowing its identification as well as report. Which item is selected for further processing (and, ultimately, a subject's response) is based on a process of sampling from a probability distribution. Importantly, the authors showed that the distribution of guess two reports was not modulated by which item was provided as guess one. This led to the conclusions that the temporal position of the selection window does not vary considerably across trials and that both guesses are sampled from the *same* probability distribution (Vul et al., 2009).

Botella and colleagues (Botella et al., 2001) provided another valuable theoretical account of intrusion errors in RSVP experiments. The authors formulated a theoretical model, which is based on Treisman and Gelade's (1980) feature integration theory (FIT). According

to FIT, intrusion errors should occur because of the inadequate focusing of attention. Botella et al. (2001) implemented this idea in a model consisting of four main elements: module K, module R, a focussing mechanism and a sophisticated guessing mechanism, respectively. Modules K and R act in parallel and extract key and response features, respectively. The focusing mechanism activates attentional processing once a task-relevant key feature has been detected, which the authors call t_c, the critical moment. This mechanism, however, fails at times, especially if presentation rates are very brief. In cases in which the focusing mechanism fails, the sophisticated guessing mechanism comes into play. Sophisticated guessing occurs via selecting the item with the highest module R activation at the critical moment, t_c. It is crucial to stress that in this framework correct reports occur *either* after successful completion of the focussing mechanism *or* after cases in which the sophisticated guessing mechanism happened to select the correct item in module R, as its respective representation had the highest activation value at time t_c. The latter case is described by the authors as *fortunate conjunctions* (conjunctions being synonymous to what we label intrusions), as correct reports were given after guessing. In contrast to correct reports, distractor intrusions are never reported after the focussing mechanism, but exclusively happen after sophisticated guessing (Botella et al., 2001). The authors furthermore provide empirical data to support their model. One observation that supported their dual-route model architecture was made after allowing subjects to report that the target was not presented in the response menu (Botella & Eriksen, 1992). If the target was indeed not presented as a possible response, subjects chose the "not in the menu" option 37% of the times, compared to 0.7% when it was included in the response screen. The baseline condition, which did not implement the "not in the menu" option, showed correct reports in 67% of trials. The authors hence concluded that those 67% of trials in which correct reports were given were the result of two kinds of processes. The first process is successful focussing, which happened 36.3% of times (37%-0.7%). This is because correctly identifying that the target stimulus was not presented in the response menu implies such high confidence that it must have been after successful focussing. The difference between the frequencies of correct reports without (67%) and with (36.3%) the "not in the menu" option furthermore suggests that a second processing route must exist, which accounts for the remaining 30.7% (67%-36.3%) of trials. The authors argue that this second kind of processing occurs via their sophisticated guessing mechanism (Botella et al., 2001). The authors additionally refer to reaction time data (Botella, 1992), which proposed a mixture of short and long trials involving correct reports. This again supports the

authors' dual-route model, as fast and slow correct reports are proposed to be the result of successful focussing and sophisticated guessing, respectively (Botella et al., 2001).

In contrast to the accounts of Vul et al. (2009) and Botella et al. (2001), which do not specifically consider temporal aspects of attentional processing, a temporal variability account was provided by Zivony and Eimer (2020b, 2020a). The authors initially postulated a theoretical account, which stated that representations of the target and the intruder item competed for encoding to working-memory (WM) and that *only* the winning item was encoded (Zivony & Eimer, 2020a). However, they since modified their encoding account based on a set of novel empirical findings (Zivony & Eimer, 2020b). Importantly, in both papers, they implemented experimental paradigms in which category-matching distractor items (i.e., intruder-items), which share the response dimension of the target item, were exclusively presented *immediately after* the target item on a subset of trials. The experimental paradigm of Zivony and Eimer's (2020a) Experiment 1 is presented in Figure 4, with time unfolding from top to bottom, and target stimuli being defined as digits that are surrounded by geometrical shapes. Since only two digits are presented in trials, the first always being the target, only correct reports or post-target (specifically, +1) intrusions can be made. Panel B shows the authors' (Zivony & Eimer, 2020a) control condition in which no intrusion errors could be made since the only digit presented was the target.

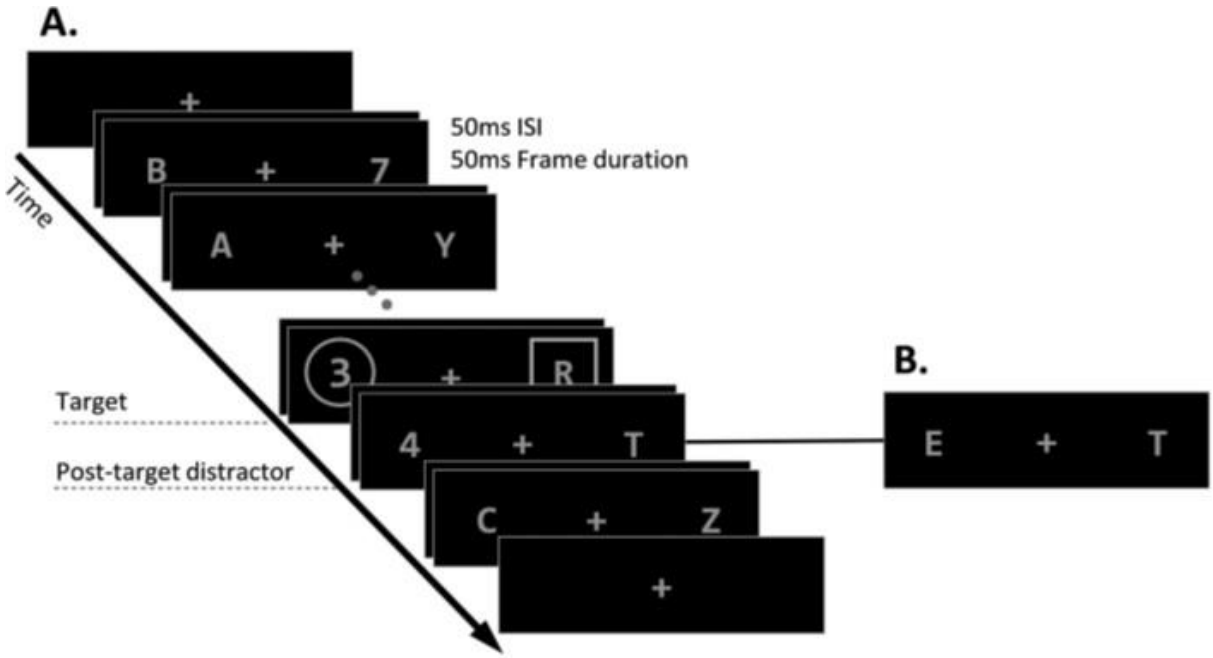


Figure 4. Example trial of Zivony and Eimer's (2020) Experiment 1, illustrating that in these experiments only post-target intrusions were possible, since only these matched the target stimulus's response dimension (i.e., both being digits).

The authors thus focused on post-target intrusion errors, rather than on pre-target ones. In their initial study (Zivony & Eimer, 2020a), a series of three dual-stream RSVP experiments was conducted to examine the temporal dynamics of the attentional system after post-target intrusion errors. The N2pc ERP-component was measured as an index of attention being allocated to visual objects carrying task-relevant features (Eimer, 1996; Woodman & Luck, 1999). Zivony and Eimer (2020a) showed that the N2pc's onset latency was significantly delayed in distractor intrusion trials compared to trials in which a correct response was provided. It was further demonstrated that this effect remained significant, irrespective of whether the key feature was an annulus (Experiment 1) or colour-marking (Experiment 2) and even when the location (i.e., which of the two RSVP-streams) of the target was cued before each trial (Experiment 3). The authors' (Zivony & Eimer, 2020a) N2pc components, showing latency differences between correct and intrusion conditions across different experiments, are presented in Figure 5.

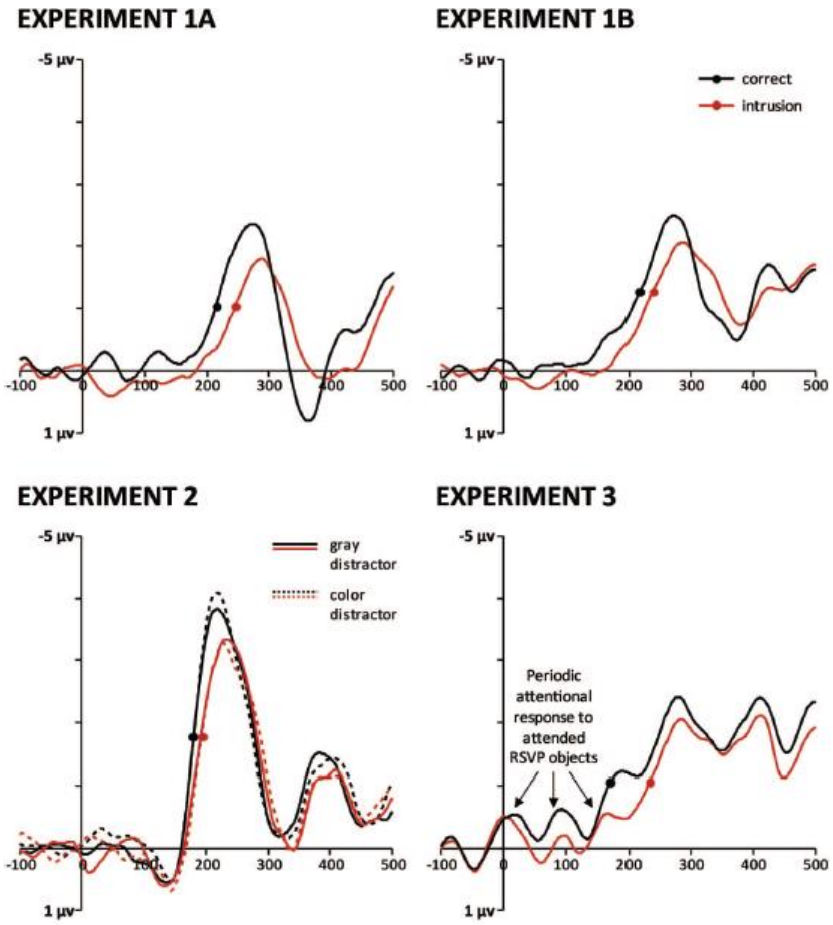


Figure 5. Zivony and Eimer's (2020) N2pc components. N2pcs showed latency differences, with N2pcs after intrusion reports being later than those after correct reports. This latency difference was found across three experiments, which either used a geometrical shape (Experiment 1) or colour (Experiment 2) as the target-defining key feature, or which cued in which of the two RSVP streams the target would appear (Experiment 3).

The authors formulated a temporal variability account in response to these results in which a transient attentional enhancement (TAE) mechanism is employed once a key-feature is detected in the RSVP stream, enhancing the strength of available stimulus-representations before gradually dissipating. Crucially, temporal variability of the TAE across trials can strongly affect visual awareness and perceptual reports. According to this temporal variability account, and as suggested by the author's N2pc results, post-target intrusion errors occur after a substantial delay to attentional engagement. This is because TAE enhancing item representations with such a delay leads to the intruder item winning the competition with the target item, hence being encoded into WM, and ultimately reported by the participants. However, the authors modified this encoding account, specifically changing the assumption that *only the winning item is encoded into WM*, after conducting a series of four experiments, again implementing dual-stream RSVP variants that carried the possibility of post-target intrusion errors (Zivony & Eimer, 2020b). In this, more recent, paper the authors specifically aimed at testing hypotheses about where in the cognitive processing pipeline the competition between the target and intruder item takes place (Zivony & Eimer, 2020b).

In their first experiment (see Figure 6), an approach similar to that of Vul et al. (2009) was adopted in which participants had to report the identity of the target twice (most likely then second most likely) using two response screens. The authors computed the conditional probabilities of the second response being the intruder given that the first report was the target (P(Intruder|Target)), as well as vice versa (P(Target|Intruder)). According to the encoding account, which specifically states that *only one item* is encoded into WM, both conditional probabilities should be at chance level. A post-encoding account was also tested, which proposes that the competition between the two items does occur *after WM encoding*, implying that both items are *always* encoded into WM. According to the post-encoding account, the conditional probabilities formulated above should be close to (baseline) performance observed in trials in which only the target and no category-matching intruder was presented. The authors found conditional probabilities that were in-between chance-level and baseline-performance, providing evidence *against* both the encoding as well as the post-encoding account. Instead, according to the authors, this result supports an *alternative encoding competition account*, according to which the competition between the two items *does not eliminate* the possibility that either one or both items will be encoded. Instead, the competition only *reduces the likelihood* that successful encoding of either one or both items occurs (Zivony & Eimer, 2020b). The remaining three experiments were conducted to examine alternative explanations of the results of Experiment 1 as well as to further probe

theoretical accounts of how and where the competition between item-representations occurs in human cognition.

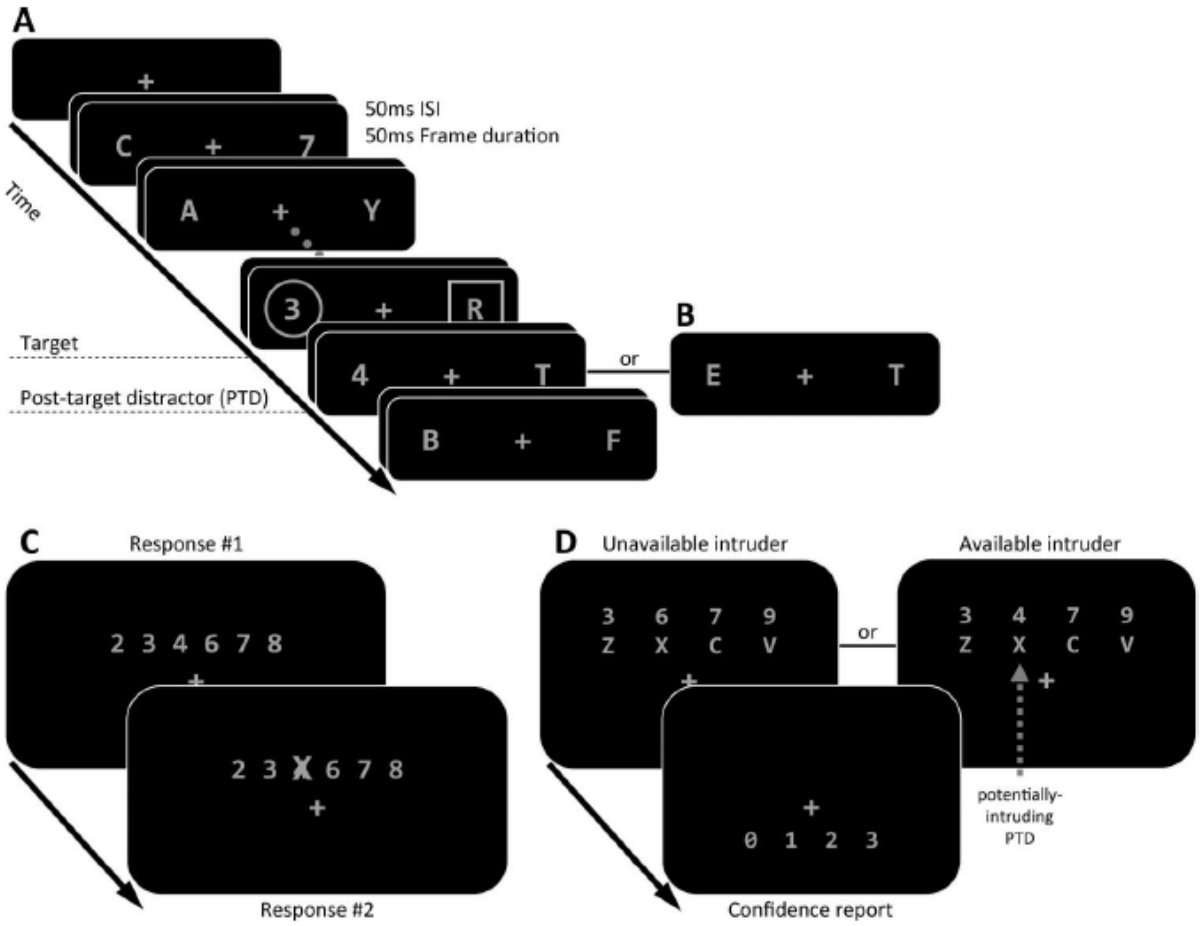


Figure 6. Experimental paradigm of Zivony and Eimer (2020). Panel A illustrates the RSVP stream, which was similar to the one presented in Figure 4. Again, post-target intruders were presented on a subset of trials (the digit '4' in Panel A) allowing for distractor intrusion errors, or not (the '4' being replaced by a category non-matching 'E' letter stimulus in Panel B), Panel C demonstrates the response screens used in Experiment 1 in which multiple responses were given. In the example of Panel C, the '4' stimulus was provided as the first response and was thus crossed out in the second response screen. Panel D illustrates the response screens of Experiment 2, in which only four responses were possible. The intruder stimulus was either presented in the response screen ('available intruder' top right screen in Panel D) or not ('unavailable intruder' top left screen in Panel D). Both screens were followed by a screen allowing for a report of confidence (Panel D).

Experiment 2 (see Figure 6) adopted a baseline condition in which no intruder was presented, and two intruder conditions, with the intruder either being presented in the response screen or not. The authors also recorded participants' confidence after each trial. This experiment provided evidence against the response threshold account, which states that participants are reluctant to report a second item represented in WM after providing the report of a first item (Zivony & Eimer, 2020b), as the accuracy of baseline trials was not equal to trials in which the intruder was presented in the stimulus-stream, but not displayed on the response screen. Furthermore, the confidence scores provided evidence against the encoding account, according to which confidence should be equal to encoding, as well as the

experiential blink account (Pincham et al., 2016). The latter account was derived from an attentional blink study (Pincham et al., 2016), which showed that at lag 1 (targets 1 and 2 being presented directly after one another), the subjective visibility of one of two targets remained low, even though report accuracy was high for both. Applied to experiments of distractor intrusions, this account would expect the target and intruder item to both be encoded into WM, with one of them being blocked from access to conscious awareness (Pincham et al., 2016). However, we would argue that transferring the findings of Pincham et al. (2016) to distractor intrusion paradigms has to be done with consideration of the fact that in contrast to integratable multi-dimensional stimulus-features that are bound together to single percepts in the case of intrusion errors, the classic AB target stimuli (target letters in streams of distractor digits) used by Pincham et al.'s (2016) experiments are not combinable into single percepts. For related work on integratable stimuli, see Simione et al. (2017), which demonstrated high accuracy and high subjective visibility for integrated percepts. In Experiment 3, a lure was added to the response screen to test some accounts according to which the competition between target and intruder items leads to *reduced precision* of both items' representations in WM. Lures were letters that were similar to target-letters (e.g., M & N or V & U). Even though lures were shown to reduce report accuracy, this did not depend on whether an intruder was present or not. This was further evidence against accounts such as the post-encoding account, which postulate reduced precision of item-representations in WM due to the competition between target and intruder items. However, this pattern of results was again compatible with the author's modified encoding account.

In their last experiment the same two-response procedure adopted in Experiment 1 was used (compare Figure 6). However, trials in Experiment 4 always included a post-target intruder item after the target and the authors measured the CDA ERP-component as a measure of WM-maintenance (Luria et al., 2016). The authors were interested in the CDA-contrast between trials in which only the target or the intruder item was reported (single-item) and trials in which participants reported both (two-item). According to the encoding account, CDA amplitude differences should be visible from the very beginning of the component and should be constant during the entire maintenance period (Zivony & Eimer, 2020b). The authors further tested the catch-and-release account with this experiment, according to which it is possible that participants encode both items in most trials but are only able to maintain one item long enough to be reported. Results showed smaller and later CDA components after single-item compared to two-item trials and larger mean CDA amplitude for the latter. The authors further divided the CDA time-window into early (400-500 ms) and late (500-800 ms)

and showed that the largest CDA differences between conditions was present in the early time-window. Also, single-item trials were not found to be different from zero during the early time-window. These results suggest that the number of items held in WM differed between single- and two-item trials and that this is the case as soon as the CDA emerges. Furthermore, no evidence in favour of the catch-and-release account was shown. Instead, this experiment again supported the authors modified encoding account.

To stress, Zivony and Eimer's (2020b) modified encoding account postulates that the target-intruder competition prevents either one of the two items being encoded into WM on a substantial proportion of trials. Still, this does not mean, as originally proposed by the authors (Zivony & Eimer, 2020a), that the competition *always blocks* at least one item from entering WM, it merely decreases the likelihood thereof. The authors state that the competition takes place at an earlier pre-encoding stage of processing and furthermore speculate that whether an item enters WM depends on whether an encoding threshold is reached or not (Zivony & Eimer, 2020b). Reaching this encoding threshold is again influenced by temporal variability underlying TAE impacting item representations, which was already crucial in their original proposition (Zivony & Eimer, 2020a). In addition, the authors reason that such an encoding threshold should involve some top-down flexibility, which varies based on task demands to ensure adaptability of our cognitive system. For example, if subjects expect to report one target, the encoding threshold should be set rather high to assure that no intruding distractor items are encoded instead. In contrast, if two targets should be reported, it would be adaptive to set the encoding threshold to a comparatively low value, which would allow both items of interest to be encoded. Initial empirical evidence for this idea was already provided in their fourth experiment (Zivony & Eimer, 2020b), which implemented the same paradigm as the first experiment with the exception that targets were followed by an intruder item in every trial. This presumably minor change was accompanied with a much higher likelihood of participants reporting *both items* compared to Experiment 1, suggesting a change in top-down control, supposedly via lowering of the encoding threshold.

A neural alternative: The 2-feature Simultaneous Type/ Serial Token Model (2f-ST^2)

In Chapters 4-6, we aim to augment this landscape of behavioural and neuroimaging findings as well as theoretical accounts, presenting further work on the 2-feature Simultaneous Type/ Serial Token Model (Chennu et al., 2011), which extends the original Simultaneous Type/ Serial Token model (Bowman et al., 2008; Bowman & Wyble, 2007) in order to account for and simulate a variety of the behavioural, as well as neuroimaging,

findings presented above. A preliminary version of this model was presented in Chennu, Bowman and Wyble (2011). In the current thesis, we will present a revision of the 2f-ST^2 model, which has enabled us to replicate Zivony and Eimer's (2020a, 2020b) experimental paradigm in which only two items share the response feature, in addition to Botella's (Botella, 1992; Botella et al., 1992, 2001; Botella & Eriksen, 1992) paradigm in which all stimuli of the RSVP stream carry the same kind of response feature. Compared to the different theoretical accounts presented above, our computational model is mostly in line with the account of Zivony and Eimer (Zivony & Eimer, 2020a, 2020b). This will be elaborated upon in depth in the general discussion.

The overall architecture and dynamics of the 2f-ST^2 model will be presented in-depth in Chapter 4's methods section. There we will not only introduce the model architecture and how different configurations of the model enable us to replicate several different experimental paradigms, but also provide a detailed list of all modifications made to the initial 2f-ST^2 model (Chennu et al., 2011).

Broadly, the 2f-ST^2 model extends the original model's (Bowman & Wyble, 2007) architecture (Figure 3), adopting separate key and response pathways in its Stage 1, which extract and simultaneously process items' key and response features. Likewise, two binding pools are implemented in the 2f-ST^2 model's Stage 2, allowing a working-memory token to be associated with a key (e.g., colour) and response (e.g., shape) type. Without going into too much detail, there are a few important ways in which the 2f-ST^2 model links to the computational model of Botella et al. (2001) and to the theoretical model proposed by Zivony and Eimer's (2020a, 2020b). First, the 2f-ST^2 model's key pathway extracts items' key features and hence conceptually resembles Botella et al.'s (2001) Module K. Task-relevant key types excite the 2f-ST^2's blaster circuit, which in turn enhances later Stage 1 activation levels across both pathways. This blaster enhancement mechanism therefore resembles the TAE processes described in Zivony and Eimer's (2020a, 2020b) accounts. Further, the 2f-ST^2 model implements lateral inhibition in the latest Stage 1 layer of its response pathway, which effectively induces a competition between response types to be bound to the currently active token and, thus, encoded into WM. This competition between response types can be seen as matching the target versus intruder competition described by Zivony and Eimer (2020b).

In summary, we will present further work on the 2-feature Simultaneous Type/ Serial Token (2f-ST^2, (Chennu et al., 2011)) computational model in Chapters 4-6, which is the first *neural* model that specifically accounts for distractor intrusion errors in RSVP experiments, to

probe Hypotheses 1 – 3 of this thesis. The 2f-ST^2 model accounts for distractor intrusion errors by modelling temporal variability underlying its TAE mechanism, the blaster. The theoretical account of Zivony and Eimer (2020b, 2020a) therefore is closely related to the 2f-ST^2 model.

We will address the second case of event integration in RSVP paradigms analysed by this thesis, temporal event integration, next. Temporal event integration describes the process by which multiple stimuli presented in sequence are merged together and reported as single percepts given that there is a perceptually meaningful way of doing so. Perceptually meaningful in this context concretely means according to Gestalt properties (Von Ehrenfels, 1937; Wagemans et al., 2012). Since integratable target stimuli in these experiments are always presented immediately after one another, i.e., at lag 1, it is worth briefly introducing studies that analysed lag 1 sparing in the Attentional Blink (i.e., the observation that at lag 1 participants report T2 accurately, given that they reported T1 correctly (measured as the conditional probability P(T2|T1), see Figure 2)).

Temporal Event Integration

The Attentional Blink's lag 1 sparing (Figure 2) was initially explained in terms of a 'sluggish gate' (Chun & Potter, 1995; K. L. Shapiro et al., 1994), which expresses that at lag 1 T2 is presented just in time to join T1's encoding process, similar to the idea of joint tokenization in the STST model (Bowman & Wyble, 2007). Potter et al. (2002) challenged this idea using target words embedded in RSVP streams of non-word distractors. Based on evidence suggesting a competition between T1 and T2 for attentional resources, they introduced a competition based account of lag 1 sparing (Potter et al., 2002). Hommel and Akyürek (2005) questioned both theories. They revealed a trade-off between target identity and order information during lag 1 sparing and specifically showed that lag 1 sparing can only be observed when people are allowed to ignore targets' order (Figure 7).

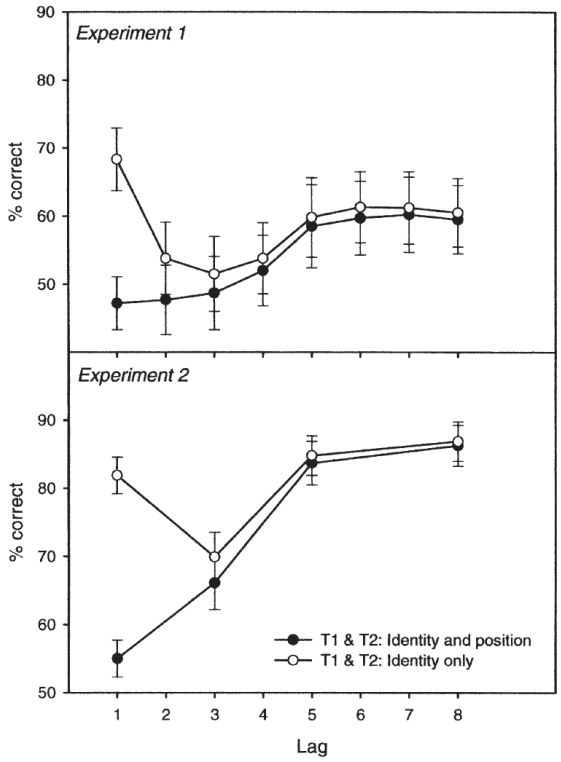


Figure 7. Subjects' average performance in two AB experiments plotted as a function of lag. It can be seen that lag 1 sparing depended on whether targets' order information could be ignored. Adopted from Hommel and Akyürek (2005).

This was later confirmed in one of the experiments of the original STST paper (Bowman & Wyble, 2007); see Figure 20 in that paper. Therefore, the question was raised why only performance on identity, but not order information, should benefit from a mechanism based on T1-T2 competition. This finding was instead rather indicative of an integrative account, such as expressed by the sluggish gate metaphor, for which T1 and T2 representations are merged into a single episode of memory. Nonetheless, Hommel and Akyürek (2005) found evidence for the competition based theory, too, by showing that targets' discriminability determines their competitive strength. Hence, they distanced themselves from treating these two accounts as being incompatible with each other and concluded that "there are reasons to assume that competition and integration accounts do not provide alternative interpretations of the same phenomenon but, rather, refer to the different possible outcomes of concurrent target processing" (Hommel & Akyürek, 2005, p. 1429).

Further investigating this issue, Akyürek et al. (2012) designed novel experimental paradigms using RSVP with target stimuli that can be combined in a perceptually meaningful way according to Gestalt properties (Von Ehrenfels, 1937; Wagemans et al., 2012). An example can be seen in Figure 8. In this experiment, combinations of corner lines were

presented as targets, which sometimes led to subjects perceiving their combination, depicted under ‘Int.’, as a single representation.

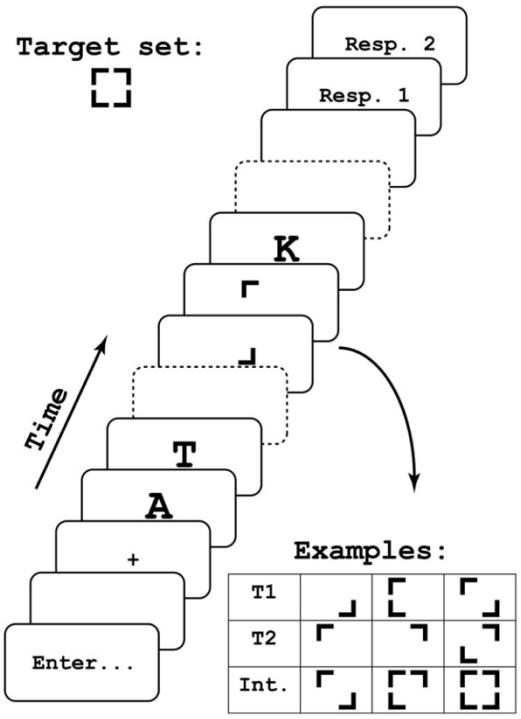


Figure 8. Experimental paradigm of Akyürek et al. (2012). The RSVP stream consisted of distractor letters and target corners that were combinable in a perceptually meaningful way. Examples of this integration process are shown in bottom right box under ‘Int.’. Adopted from Akyürek et al. (2012).

Akyürek et al.’s (2012) behavioural results consistently demonstrated across four experiments with different sets of mergeable target stimuli that when the possibility of temporal integration was present, their reports were three times as likely as ‘real’ order reversals. Moreover, once this was no longer the case, the number of ‘real’ order reversals tripled (Akyürek et al., 2012). These findings therefore provided rather strong evidence in favour of an integration-based (or, originally, sluggish-gate) theory of lag 1 sparing over accounts of T1-T2 competition. Nonetheless, it should be re-emphasized that these two explanations do not have to be mutually exclusive, but likely refer to different possible consequences of processing two target stimuli in parallel.

Most recent advancements in the field of temporal event integration during lag 1 sparing considered the dynamics of ERP components connected to attentional and WM processing (Akyürek et al., 2017). Chapter 7 analyses the data of Akyürek et al.’s (2017) study, which is why we will only briefly introduce it now but do so in detail later. The authors extended their experimental paradigm shown in Figure 8 by presenting two RSVP streams simultaneously, one of which contained target stimuli, and further varied between showing

only one and two targets per trial. One of their main findings was that the N2pc component was modulated solely based on how many target stimuli were *presented*. In contrast, components corresponding to WM (CDA & P3) depended on the number of targets *perceived*, being one for integrated two-target trials, irrespective of how many targets were actually presented. Importantly, it was found that temporal integration had an alleviating effect on CDA and P3 amplitude, implying that it is favourable for WM encoding to merge two targets into one representation. Hence, these findings can be viewed as further indication of a pivotal role of temporal event integration during lag 1 sparing.

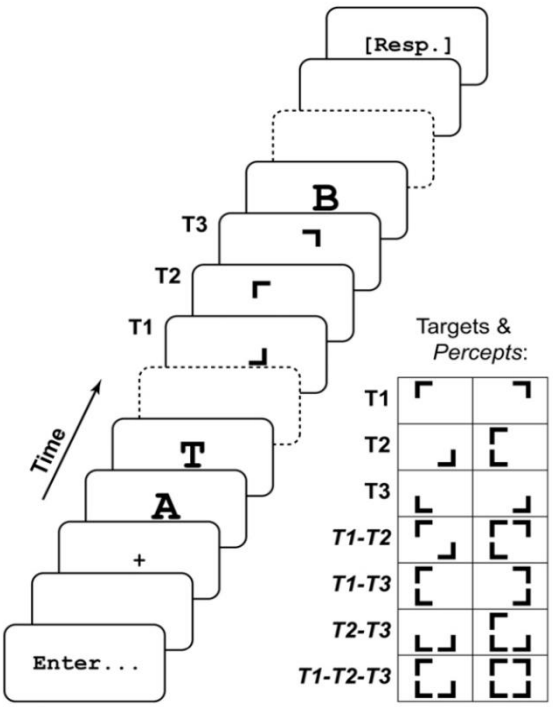


Figure 9. Experimental paradigm of Akyürek and Wolff's (2016) first experiment, which investigated temporal integration over an extended time-window and for three perceptually compatible target stimuli. An example three-target trial is shown on the left and individual target stimuli as well as their integrated combinations are shown on the right. Adopted from Akyürek and Wolff (2016).

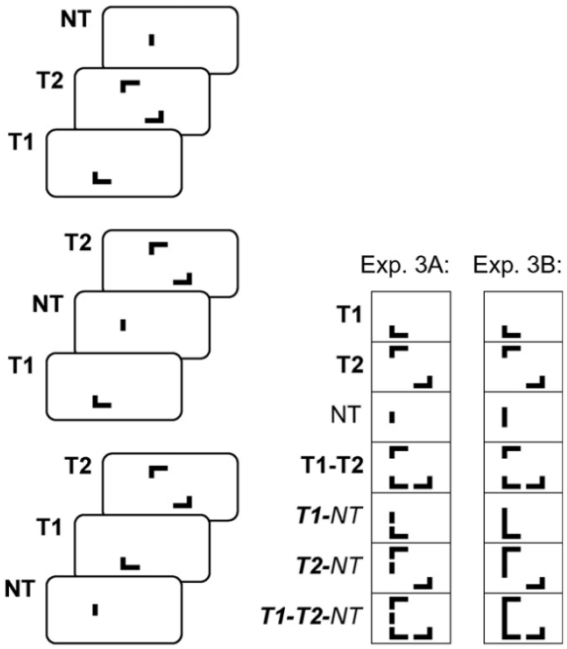


Figure 10. Illustration of target-nontarget combinations used in Akyürek and Wolff's (2016) experiments 3A and 3B. Possible sequences of T1, T2 and non-target (NT) are shown on the left. Examples of T1, T2 and NT stimuli and their respective integrated combinations are displayed on the right. Adopted from Akyürek and Wolff (2016).

Akyürek and Wolff (2016) recently proposed adjusting the eSTST model (Wyble et al., 2009) to account for temporal integration in addition to the spreading the sparing findings as a promising path towards a more complete AB model. This proposal was established on data that scrutinized temporal event integration over an extended time window using three mergeable target stimuli, thereby linking this phenomenon to an experimental setting similar to that adopted by Oliver et al. (2007). This work generated a number of interesting conclusions. First, the authors revealed that the integration of three successive targets (Figure 9) over a time window of 240 ms was as likely to occur as two-target integration in two-target trials. Further, instructing subjects to attend to the collective target features increased reports of integrated percepts, which implied an impact of endogenous top-down control on temporal integration. Finally, two experiments on perceptually combinable target-nontarget combinations (Figure 10) showed that integration in such a manner was rare compared to the target-target equivalent (Akyürek & Wolff, 2016). This therefore suggested exogenous bottom-up influence on temporal event integration. Importantly, target-nontarget integration was not only rare for three item trials but was likewise seldom observed when the non-target (NT) followed T1 immediately. The latter finding implies that bottom-up attentional selection (or in this case exclusion) operated rapidly enough to intervene even at such short intervals, which contradicts the original 'sluggish-gate' metaphor (Chun & Potter, 1995; K. L. Shapiro

et al., 1994), as this account would have predicted that a perceptually compatible NT would 'slip in' T1's encoding process.

It still is up for debate which exact role temporal event integration holds in the AB, especially in terms of lag 1 and multi-target sparing. Nonetheless, the evidence provided by Akyürek and colleagues (Akyürek et al., 2017; Akyürek & Wolff, 2016) suggests it is a critical cognitive process, with potential implications for the AB and, more generally, human attention and memory. In Chapter 7, we therefore aim at resolving the exact temporal as well as spatial unfolding of this process in human cognition. Doing the latter, we will adopt the (machine learning based) temporal generalisation method (King & Dehaene, 2014) to analyse Akyürek et al.'s (2017) EEG dataset. The temporal generalisation method is part of a group of methods that is increasing in popularity over the past years: multivariate pattern analyses (MVPA). We will therefore introduce MVPA methods in general, as well as the temporal generalisation method in particular, since it is particularly relevant for this thesis, next. We will adopt the temporal generalisation method in Chapter 8, too, to illustrate a methodological risk that is present when applying machine learning algorithms to neuroimaging datasets in general.

Multivariate Pattern Analysis (MVPA)

Neuroimaging research assesses patterns of brain activity most commonly via mass univariate, parametric statistical approaches. In this context, statistical inference is first performed via applying a general linear model (GLM) separately to, for example, all individual voxels recorded with functional magnetic resonance imaging (fMRI). This is then followed by assessing the statistical strength of each voxel's effect, for example via a t-test. The resulting p-values are subsequently corrected for the number of simultaneous tests performed, using topological inference over the whole cortex (Flandin & Friston, 2015; K. J. Friston et al., 1991; Kilner & Friston, 2010). An alternative to this statistical inference method that is being adopted more frequently over the past couple of years is called multivariate pattern analysis (MVPA). With MVPA one performs a single multivariate test across all voxels (fMRI) or electrodes (EEG). Prominent methods of MVPA are representational similarity analysis, RSA, (Kriegeskorte, 2008) and the temporal generalisation method (King & Dehaene, 2014). Before introducing the temporal generalisation method, we briefly introduce classification algorithms in general as they lay the foundation of the temporal generalisation method, which we will adopt in Chapters 7 & 8.

Classification algorithms

We will focus on support vector machines (SVM) for our discussion of classification algorithms because it is the algorithm we used predominantly in the current thesis. Although, the field of machine learning generated numerous different classification algorithms, such as Fisher's linear discriminant analysis (LDA), logistic regression and naïve bayes classifiers, neural networks, and decision trees. The overarching task of this family of algorithms is to differentiate, or classify, data based on specific features. This intrinsically generic function of classification algorithms made them popular for a wide range of applications. In neuroscience, classifiers are applied to neuroimaging data to decode brain patterns that correlate to a cognitive process of interest. To achieve the latter, classifiers need to be trained. This is done by presenting the classifier data features that are labelled according to the effect of interest. For neuroimaging, the feature of interest usually is brain response amplitude, measured by the BOLD response in fMRI or microvolts in EEG. Neuroimaging data is typically measured over a specific time-series, e.g., an experimental trial. For example, staying in the context of the AB, one could look at differences in brain activity during a trial based on whether a T2 was blinked or not. The classifier would be provided with numerous trials of both cases. Importantly, each trial needs to be labelled correctly depending on whether an attentional blink occurred or not. After processing all trials, the classifier fits an optimal hyperplane between the two conditions of interest to separate them in that dataspace, which concludes training. A trained classifier's efficacy in separating classes is finally estimated by first letting it classify data (i.e., patterns of brain activity) it does not know and then measuring how often it classified successfully (i.e., correctly predicted the blink).

Support vector machines (SVM)

Support vector machines (SVMs) operate in the way just described. However, there are a number of SVMs available, which differ in their kernels, or in *how* they fit the hyperplane between classes. Figure 11 illustrates a linear SVM algorithm that fit a hyperplane to separate classes.

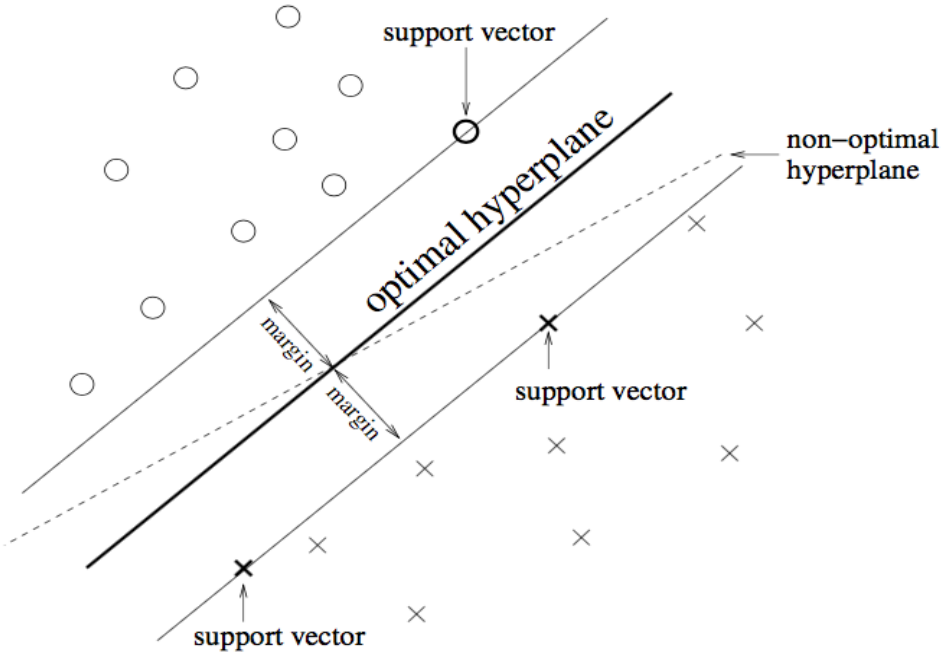


Figure 11. Two-class classification via an SVM. Circles and crosses correspond to data of two distinct classes. The optimal hyperplane is fitted between classes according to the maximum margin. Adopted from Lotte et al. (2007).

There are several different hyperplanes that would achieve this task, but the SVM always choses what is called the optimal hyperplane in Figure 11. Another name of this hyperplane is maximum-margin hyperplane, which hints at why it is the optimal choice for the SVM. The margins are displayed in Figure 11, too, and reflect the distance between the hyperplane and those data points nearest to that hyperplane, called support vectors. A hyperplane is considered optimal when its margins are maximal (Lotte et al., 2007). This is sometimes figuratively illustrated with maximizing the width of the 'street' around the separating hyperplane to the support vectors. Figure 11 provides an example non-optimal hyperplane, too, in which the top right cross and the bottom left circle would be the nearest support vectors. The hyperplane 'street' associated with those support vectors would be narrower compared to that of the optimal hyperplane and, hence, its margins would be smaller. Ensuring that the hyperplane's margins are maximal guarantees best classification performance of the linear SVM for any dataspace.

Moreover, since we will use polynomial as well as radial basis function (RBF) in addition to linear kernels when adopting SVM algorithms in Chapter 8, we introduce the former two kernels. Polynomial and RBF kernels (Prajapati & Patle, 2010) work in similar ways, both enabling the classification of overlapping datasets, meaning datasets that are not separable linearly. With polynomial kernels, a high-dimensional relationship between pairs of observations in the dataset is computed. This allows classification problems that are not

solvable with a linear kernel to be solvable without actually transforming the data into a high dimensional dataspace, as knowing the observations' relationship in that dataspace is sufficient. RBF kernels find support vectors in infinite dimensions. It behaves similar to a weighted nearest neighbour model (Cost & Salzberg, 2004), meaning that the nearest observations influence the algorithm's classification decisions the most when classifying new observations. Specifically, the influence (or weight) of neighbouring observations in the training dataset on classifying a new observation is a function of the squared distance of neighbours to the new observation. In essence, and again similar to polynomial kernels, RBF kernels compute the high-dimensional relationship between observation-pairs without the necessity of transforming data into these dimensions. What makes RBF kernels particularly powerful is that they provide the relationship between observation-pairs in infinite dimensions. Exactly how RBF kernels accomplish the latter is beyond the scope of this introduction. However, one very simplified explanation is that the RBF kernel is equal to a dot product that has coordinates for an infinite number of dimensions.

The Temporal Generalisation Method

King and Dehaene (2014) have introduced the temporal generalisation method, which uses classification models, such as SVMs, for analysing time-series neuroimaging data (Figure 12, adopted from (King & Dehaene, 2014)). As illustrated in Figure 12.1 & Figure 12.2, temporal generalisation trains distinct classifiers at every time-point of the experiment's time-series.

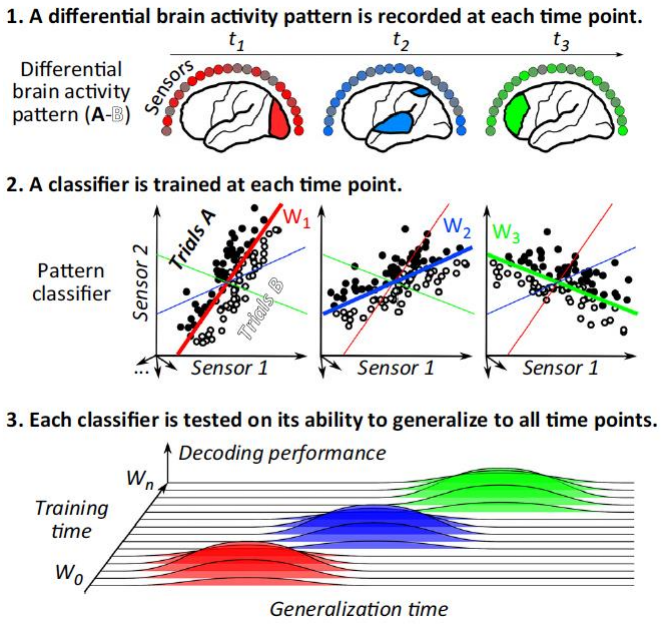


Figure 12. The temporal generalisation method. First, differential brain activity is recorded via neuroimaging. Then, separate classifiers are trained at each time-point of the time series. Finally, classification performance is measured as a function of training and testing time. Adopted from King and Dehaene (2014).

For example, an EEG experiment that sampled brain activity at 500 Hz and over a time-window of one second would require 500 classifiers to be trained. Training classifiers is performed as described in the previous paragraph. Once all classifiers have been trained on the EEG data recorded at *their* respective time-point, they are tested on their ability to differentiate the experimental conditions at *all* time-points of the time-series. Naturally, experimental conditions captured via class labels need to remain constant across training and testing. Finally, performance of classifiers is displayed in a matrix (or, 2D-map) as a function of training and testing time, which indicates the temporal dynamics of differential brain patterns observed based on the manipulation of interest (Figure 12.3 and Figure 13). A typical temporal generalisation (TG) map is displayed in Figure 13.

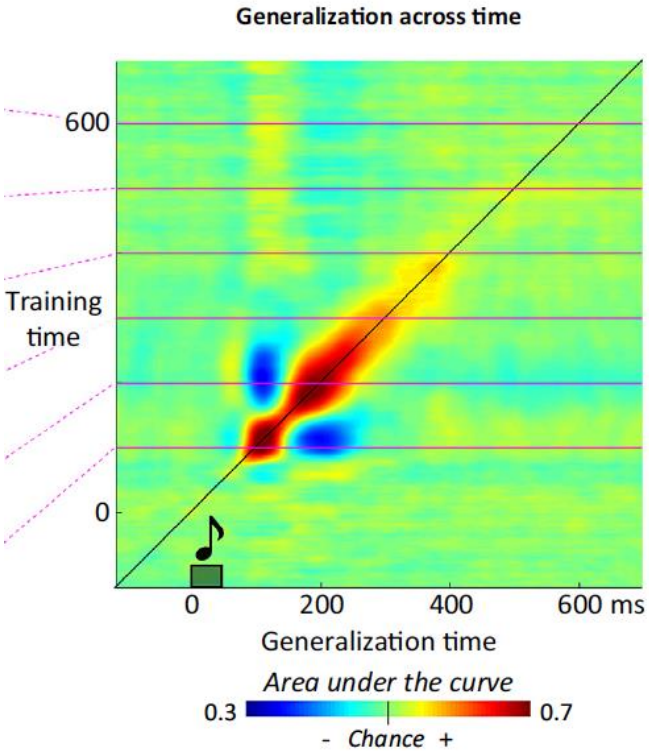


Figure 13. A temporal generalisation map showing classification performance as a function of training and testing (labelled here as generalization) time. Performance of classifiers is measured via area under the curve. Adopted from King and Dehaene (2014).

In Figure 13, 'diagonal decoding' is indicated by the black line and refers to the scenario when testing is equal to training time, revealing how effective a classifier was at categorizing unknown data recorded at the same time-point as the data that was used during the classifier's training. Diagonal classification can therefore be seen as clarifying whether a difference between conditions was present at a particular time-point. The pink lines in Figure 13 correlate to 'horizontal decoding' and demonstrate how well a given classifier was able to distinguish classes over the whole time-series. Testing classifiers at time-points other than

those at which they were trained makes this approach particularly interesting for neuroimaging, as it enables two novel types of observations. First, in addition to revealing differences in brain activity patterns at specific time-points, it facilitates inference about the extent of *temporal generalisation* of such difference patterns. For example, the diagonal classification pattern shown at roughly 200 ms in Figure 13 implies, due to its width, that the underlying difference in brain activity pattern extended over a time-window of approximately 150 – 250 ms. Second, the method further captures if such differences should re-occur later on in the time-series. The latter becomes apparent if one considers a hypothetical case of a diagonal classification pattern at 200 ms being accompanied by horizontal classification at 400 ms (i.e., training time = 200 ms & testing time = 400 ms). This not only implies that there was a difference at 200 and 400 ms, but that the same difference observed at 200 ms re-occurred 200 ms later in the trial. King and Dehaene (2014) provided additional hypothetical scenarios of brain differences such as of oscillating and sustained nature and specified how these would be reflected in a temporal generalisation map.

Applications of the temporal generalisation method have already yielded a number of valuable findings, for example about disorders of consciousness (King et al., 2013), the brain response to auditory novelty (King et al., 2014), visual object recognition (T. Carlson et al., 2013; T. A. Carlson et al., 2011; Cichy et al., 2014; Isik et al., 2014), dual-task interference (Marti et al., 2015) or the dynamics of temporal selection of targets in RSVP (Marti & Dehaene, 2017).

In summary, this literature review commenced with introducing RSVP experiments, particular those involving the AB as well as theoretical and computational models explaining the AB. We then introduced the two cases of event integration analysed by this thesis, distractor intrusion errors and temporal event integration, before finally introducing MVPA methods, particularly the temporal generalisation method. Before presenting the contribution of this thesis in the research Chapters 4-8, we will briefly introduce two statistical methods that will be adopted in Chapters 7 & 8, next.

Chapter 3 - Methods

Maximum Statistics Permutation Tests of TG Maps

In this section, we will introduce the two statistical tests we adopted to assess whether classification performance, measured as area-under-curve (AUC), in temporal generalisation (TG) maps was statistically significant. These tests will be used frequently in Chapters 7 & 8

and concretely assess whether the AUCs at hand differ substantially from chance level, e.g., from 50% for the case of two-class classification (we were only interested in two-class contrasts throughout the current thesis). Since the temporal generalisation method is still in its early years of development, a consensus in the literature has yet to be reached as to which statistical test shows the best fit for assessing such classification maps. Some scholars have made use of traditional, parametric tests, performing, for each obtained accuracy, a one-sample t-test versus chance level. Importantly, since an extraordinary number of tests are performed simultaneously, the null hypothesis of the effect of interest being absent in the population (henceforth called 'the null') will often be rejected erroneously. In statistics, this mistake is described as a 'Type I Error' and yields conclusions that are, in signal detection theory terms, 'false positives', declaring a statistically significant effect where there really is none. Under the null and for the standard alpha level of significance of five percent, every 20th test will on average lead to a wrongly rejected null hypothesis. Thus, the more tests one performs at the same time and if one takes the lowest p-value, the probability of Type I Error increases beyond five percent.

Example TG maps are presented later in the context of experimental trials that were recorded from -200 to 1000 ms around the stimulus of interest (Figure 85). These TG maps would include 360000 p-values after (1st level) statistical data analysis, which, on average, would lead to 18000 wrongly significant p-values under the null. Therefore, if one would not take any precautions about the interaction of number of simultaneous tests vs. number of Type I Errors, one would easily interpret some random, chance patterns of statistical significance in their TG maps as real effects and potentially formulate invalid conclusions based on them. Such precautions are corrections for multiple comparisons and the most prominent procedures are family-wise error and false discovery rate (Benjamini & Hochberg, 1995) corrections. Some of these corrections work directly on the p-values, increasing each value based upon how many tests were carried out in total, making it harder to reach statistical significance. However, even if one were to apply corrections for multiple comparisons to their p-values, there still is a fundamental pitfall in performing parametric tests, such as the Student's t-test on measures such as accuracies, as was pointed out by Allefeld, Görgen and Haynes (2016). The authors' central point of criticism revolved around the claim that traditional t-tests on single-subject accuracies or similar information-based measures, as they call them, are not valid. This is due to the fact that the populations' real value of such measures can never be below chance level. This, as the authors argue, changes the very meaning of the t-test, as it now tests the global null hypothesis, which states that

there is no effect in any subject. Thus, a significant effect against such a global null hypothesis implies that there are people for which there is an effect somewhere in the data. This is in stark contrast with the usual interpretation of significant p-values stemming from t-tests, namely that the given presence of information (or, effect) *generalises* to the population from which the sample was collected. The authors conclude their paper with an introduction of a permutation based prevalence inference test using the minimum statistic (Allefeld et al., 2016). It is worth mentioning that other scholars have advocated the use of computationally cheaper, non-parametric tests (King et al., 2013; Marti & Dehaene, 2017), such as the Mann-Whitney U-test for testing the significance of within-subject AUCs and the Wilcoxon signed-rank test (Mason & Graham, 2002) for between-subject analyses.

In the thesis at hand, we will advocate the approach of maximum statistics and specifically adopt two permutation tests, which will be introduced next. Permutation tests work by simulating the data under the null through resampling, using these simulations to generate a null distribution of the statistic of interest and subsequently testing the observed values against that distribution. A given condition with permutation tests is that the null needs to be simulated a very large number of times to ensure that one's estimate of the statistic's null distribution is reliable. Thus, it is common to compute tens of thousands of null simulations, which leads to the main disadvantage of permutation tests: computational costs. For the thesis at hand, we have programmed and applied two kinds of sign-swap permutation tests to our TG (or, AUC) maps, for which this disadvantage is absent. The conventional permutation approach for TG analyses would be to first randomly shuffle trials' class-labels and then run the classification analysis. Performing this, for example, ten thousand times would be computationally prohibitive. Instead, our sign-swap test simulated the null at the level of AUC maps, alleviating computational costs drastically. Furthermore, an importantly desirable characteristic of the sign-swap permutation test is that it automatically corrects for the multiple comparison problem demonstrated earlier. This rests upon the fact that statistical inference is, for example, based on AUC or cluster-size maxima, an approach known as maximum (or minimum) statistics. Therefore, the more tests are run simultaneously, the higher the maximum will be under the null, which makes it more difficult to reach significance. Both sign-swap permutation tests were coded manually in MATLAB.

Peak-Level Permutation Test of TG Maps

Our first permutation test assessed whether individual AUC values of the TG map were statistically significant. It commenced with the sign-swap, which is our method of

simulating the null, i.e., the absence of an experimental effect in the data. Specifically, we determined randomly and with equal likelihoods whether a given participant's AUC map was left as it was or whether their whole map was flipped around chance level by subtracting each AUC value from one. This was a valid way of simulating the null based on the fact that the null distribution of AUCs is symmetric around the chance level of 0.5, a matter known as the exchangeability assumption, probabilistic symmetry or de Finetti's theorem in mathematics and probability theory (De Finetti, 1937; Olav Kallenberg, 2005). Subsequently, the (across-participants) average null AUC map was computed, and the maximum AUC value of this map was placed in a distribution. Importantly, *maximum* in this case means the absolute maximum value, as this ensures a two-tailed statistical test. For example, if the most negative (i.e., below-chance) AUC value was further from chance-level (0.5) than the most positive (i.e., above-chance) AUC, we subtracted that negative AUC value from 1 to get the above-chance equivalent. This procedure was repeated ten thousand times, which resulted in a distribution of AUC maxima under the null. The observed group-level AUCs were then tested for significance against the corresponding null distribution of AUC maxima, again subtracting below-chance AUC values from one to ensure a two-tailed significance test. The p-values that this test generated were concretely interpretable as reflecting the proportion of AUC maxima under the null higher than a particular observed AUC. An observed AUC was labelled as being significant if the latter proportion was smaller than 0.05, corresponding to the classical significance threshold of five percent.

This permutation test enables peak-level inference, informing us about individual AUC values of the TG map. However, there are some alternatives to peak-level permutation tests, which arguably carry a higher degree of statistical power. These alternatives are based on the same overall idea of simulating the null. However, instead of assessing maximum AUC values under the null, these alternatives look at clusters. Such approaches are common in the field of neuroimaging, especially in the framework of statistical parametric mapping, SPM (K. J. Friston et al., 1994). In SPM analyses, clusters are either tested in terms of their extent (cluster-level inference) or number (set-level inference). It is known that peak-level tests provide the best spatial resolution at the cost of worst sensitivity, which is equivalent to statistical power. Set-level tests are located on the opposite side of this continuum, providing maximum sensitivity but minimal spatial resolution. The cluster-level approach lies between these two options, providing moderate spatial accuracy and sensitivity. Thus, it can be argued that our peak-level sign-swap permutation test likely involves what is called Type II Errors in

statistics, not labelling statistically significant AUCs as such. We therefore adopted a cluster-level sign-swap permutation test, too.

Cluster-Extent Test of TG Maps

Our cluster-extent (or, cluster-level) sign-swap permutation test was performed with the following procedure. For each group TG map, we first subtracted chance-level AUC (i.e., 0.5) from all its underlying single-subject TG maps. Each position (AUC) of the single-subject TG maps was then tested separately against a median of zero across subjects using a one-tailed (i.e., being greater than zero) or two-tailed (i.e., being different than zero) Wilcoxon signed-rank test, depending on whether we implemented a one- or two-tailed permutation test. This yielded a map of p-values, which was subsequently used to form clusters if neighbouring AUCs had p-values smaller than or equal to 0.05. We implemented a minimum size of 8 pixels for these p-value clusters. We stress that while this methodological choice is worth noting for the sake of transparency, it does not affect the statistical test itself, since the same criterion was adopted for true observed and permuted TG maps. We recorded the sizes of these clusters and proceeded to compute a series of permutations in order to determine which of the clusters were statistically significant compared to a null distribution (Nichols & Holmes, 2002; Pernet et al., 2015). We computed a total of 10000 permutations for this test. For each permutation, we performed a sign-swap at the level of single-subject maps to simulate the null. Specifically, instead of subtracting chance-level AUC (0.5) from all single-subject maps, as described above, we now randomly determine, with equal likelihoods and for each single-subject map separately, whether 0.5 was subtracted from it (map – 0.5) or whether the whole map was subtracted from 0.5 (0.5 – map). We again performed one- or two-tailed Wilcoxon signed-rank tests at each pixel of our permuted single-subject maps to give us maps of p-values and then formed clusters, as described before. After recording all clusters' sizes, we determined the biggest cluster and stored its size in a distribution. Repeating this process 10000 times results in a distribution of 10000 maximum cluster-sizes under the null against which our observed clusters are tested. Observed clusters were assigned p-values as the proportion of maximum (or biggest) clusters under the null found to be equal to or larger than them. This procedure was always performed separately for each group map tested (meaning permuted distributions of maximum cluster-sizes were never used more than once).

Chapter 4 – The 2-feature Simultaneous Type/ Serial Token (2f-ST^2) Computational Model accounting for Distractor Intrusion Errors

Introduction

Chapters 4-6 are dedicated to thoroughly investigate the first three hypotheses of this thesis, which are concerned with the distractor intrusion phenomenon. Chapters 4 & 5 will examine whether the binding of multi-dimensional stimulus features into one percept during rapid stimulus streams depends on the timing of transient attentional enhancement (TAE, Hypothesis 1) and whether the 2f-ST^2 model can account for findings obtained with distractor intrusion experiments (Hypothesis 2). In Chapter 4, we will present some of the EEG analyses as well as an elaborate introduction to the 2-feature Simultaneous Type, Serial Token (2f-ST^2) computational model and show how it elucidates the distractor intrusion phenomenon. Chapter 5 will subsequently introduce the *loss of responsiveness* phenomenon of the 2f-ST^2 model, which carried valuable insight about our model's behaviour and further provided intriguing implications for human cognition. In the sixth chapter we will focus on Hypothesis 3, investigating whether increased key feature salience leads to decreased temporal variability of TAE deployment in humans. In this context, previous literature as well as a novel empirical analysis will inspire a final modification to the 2f-ST^2's architecture, which, as we argue, means an even closer resemblance to the processes in human cognition that lead to distractor intrusion errors.

As introduced in detail previously, the distractor intrusion phenomenon (also called illusory conjunctions or conjunction errors, e.g., in the early work of Botella and colleagues (Botella, 1992; Botella et al., 2001)) describes the erroneous binding of (typically) two feature dimensions of two separate visual stimuli into a combined percept of one stimulus that has not been presented as such physically in the RSVP stream. Typical RSVP experiments in which intrusion errors can occur are those that present streams of coloured letters to participants. We provide a schematic illustration of how an extract of such a stream might look like in Figure 14. Figure 14, which is inspired by Figure 1 presented in Botella et al. (2011), depicts the central stimulus set of a typical RSVP stream of coloured letters, with digits next to stimulus frames indicating item-positions with respect to the target, the latter being at position 0. For example, the green 'P' stimulus in Figure 14 corresponds to the -2 item, since it is presented 2 frames prior to the target. Time unfolds from top to bottom. Participants' task would be to

report the red letter after the stream concluded, making the frame showing the red 'S' the target frame. A distractor intrusion error would be made if participants reported the identity of a neighbouring distractor stimulus instead of the 'S'. For example, if a participant reported that the red letter in this trial was the 'F', they would have made an intrusion error. Specifically, because the 'F' is the letter that follows the target 'S' immediately, we call this a +1-intrusion error. Likewise, if a participant would report the 'P' as being the target red letter, this would constitute a -2-intrusion error. In this context the concepts of key and response features are critical. Key features are those stimulus features that differentiate distractor from target stimuli. Key features can take different forms, as they depend on an experiment's task demands. In the example shown in Figure 14, the key feature is colour, as it is along this feature-dimension that the red 'S' stimulus is defined as the target. Response features are stimulus features that need to be reported after a trial. This again depends on task-demands. In the example stream shown in Figure 14, letter identity is the response feature, as participants are asked to report which letter they have seen as red. We provide a glossary defining the main terms relating to the distractor intrusion phenomenon and the 2f-ST^2 model later in Table 1.

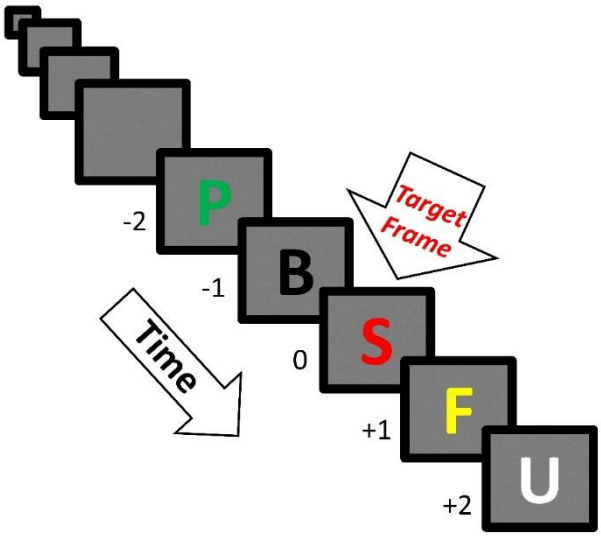


Figure 14. Example RSVP Stream that allows for distractor intrusion errors. Time unfolds from top to bottom. In this example, the task would be to report the red letter. Hence, the illustration depicts the central stimulus set surrounding the target frame that contains the red 'S' stimulus, with numbers next to stimulus-frames indicating respective item-positions with respect to the target (0). Intrusion errors are made if participants erroneously report a neighbouring distractor stimulus as being red. For example, a +1-intrusion error is made if the 'F', which immediately follows the red 'S' target frame, was reported as being red. Likewise, a -2-intrusion error is made if the 'P' was reported as being red.

The experiments presented in Botella et al.'s (2001) model paper are worth a further introduction, since we will replicate their empirical response distributions with the 2f-ST^2

model later in this chapter. Botella et al. (2001) manipulated key and response feature processing speeds, modulating target words' lexical frequency. An example stream of Botella et al.'s (2001) Experiment 1A is presented in Figure 15. Participants' task in this experiment was to report the colour of the stream's animal word. Botella et al. (2001) adopted high-frequency (HF) animal words (e.g., dog) and low-frequency (LF) ones (e.g., iguana), which modulated key feature processing speed, since HF animal words are detected faster as being animals (thus, the key feature is detected faster). Botella et al. (2001) observed a pre-target shift of intrusion responses going from the LF to the HF condition. The exact numbers corresponding to this pre-target shift as well as our explanation of it will be presented later in this chapter.

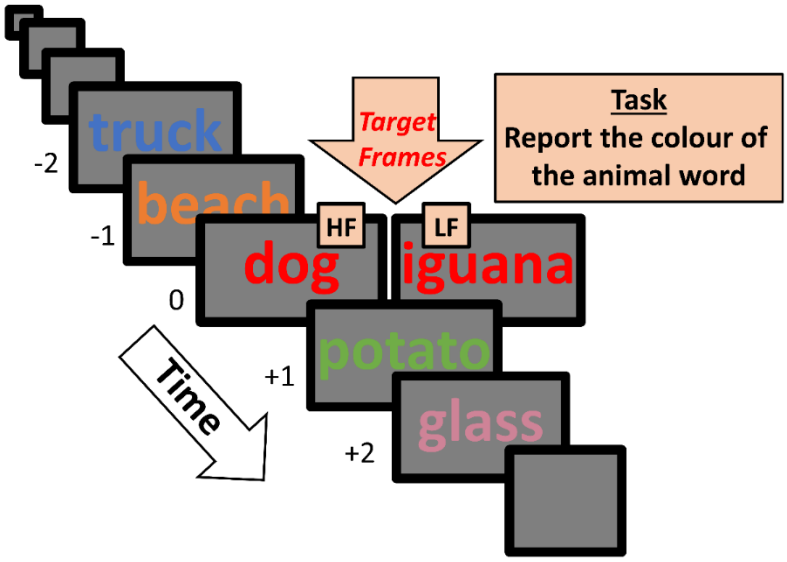


Figure 15. Experimental design of Botella et al.'s (2001) Experiment 1A. Plotting conventions are identical to those of Figure 14. Participants were asked to report the colour of the stream's animal word. Key feature processing speed was modulated, since high-frequency (HF) animal words (e.g., dog) are detected faster than low-frequency (LF) ones (e.g., iguana). Botella et al. (2001) observed a pre-target intrusion response shift going from the LF to the HF condition.

Figure 16 presents an example stream of Botella et al.'s (2001) Experiment 2. Participants were asked to report the uppercase word, which again were high- (SLOW in Figure 16, HF) or low-frequency words (PRATE in Figure 16, LF). This modulates response feature processing speeds empirically, since the key feature is fixed (uppercase-ness), while HF words are reported faster than LF ones. Botella et al. (2001) observed a post-target shift of intrusion errors in this experiment going from LF to HF conditions, which will again be presented numerically and accounted for with the 2f-ST^2 model later in this chapter.

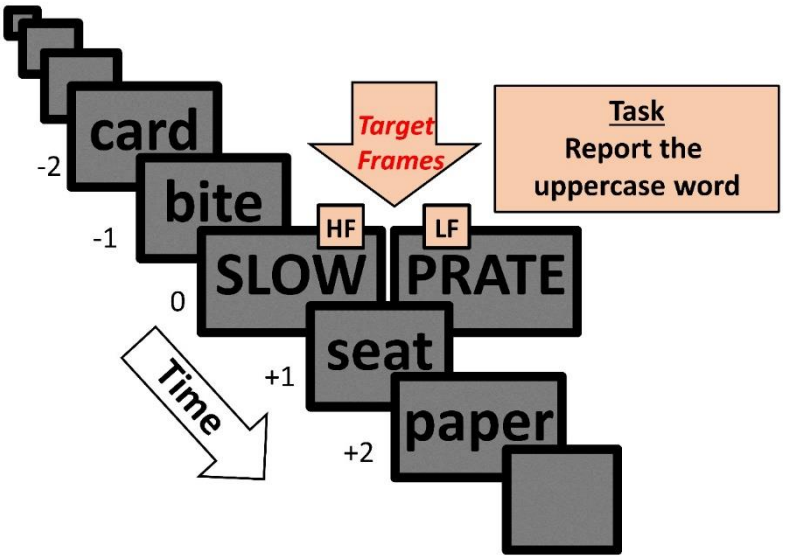


Figure 16. Experimental design of Botella et al.'s (2001) Experiment 2. Plotting conventions are identical to those of Figure 14. Participants were asked to report the uppercase word. Response feature processing speed was modulated, since high-frequency (HF) uppercase words (e.g., SLOW) are reported faster than low-frequency (LF) ones (e.g., PRATE). Botella et al. (2001) observed a post-target response shift in the HF condition.

The study of Zivony and Eimer (2020a), who used a refinement of Botella et al.'s (2001) experimental paradigm, is worth some further attention, as we will use datasets of this work in this chapter and the next two chapters. As introduced in detail previously, the authors (Zivony & Eimer, 2020a) conducted experiments that allowed for distractor intrusions and, in contrast to those of Botella et al. (2001), implemented dual RSVP streams so that the N2pc Event-Related Potential (ERP) component could be recorded. The most crucial difference between these two experimental paradigms, however, was that Zivony and Eimer (2020a) only allowed the target and the +1-intruder item, i.e., the distractor stimulus immediately following the target item (e.g., the yellow 'F' in Figure 14), to be potential responses. This is illustrated in Figure 17, which is reproduced from Zivony and Eimer (2020a) and which depicts an experimental trial in their Experiment 1 (we will introduce the author's Experiment 2 in Chapter 6). As shown in Figure 17, Zivony and Eimer (2020a) used shapes surrounding target digit stimuli in target frames to define task-relevance. Hence, the key feature was shape (i.e., an annulus surrounding the target '3' in Figure 17) and the response feature was *digit* identity (i.e., the shape of the target '3' and not of the intruder '4' that follows the '3' in Figure 17). In this experimental paradigm, only the target and the +1-intruder were potential responses, since only these two stimuli were digits and thus shared the response dimension of interest. Importantly, this is in contrast to the paradigms adopted by Botella and colleagues (Botella, 1992; Botella et al., 2001, 2011; Botella & Eriksen, 1992), e.g., as illustrated in

Figure 14, in which all stimuli could potentially have been the target as they were all letters and thus shared the response dimension.

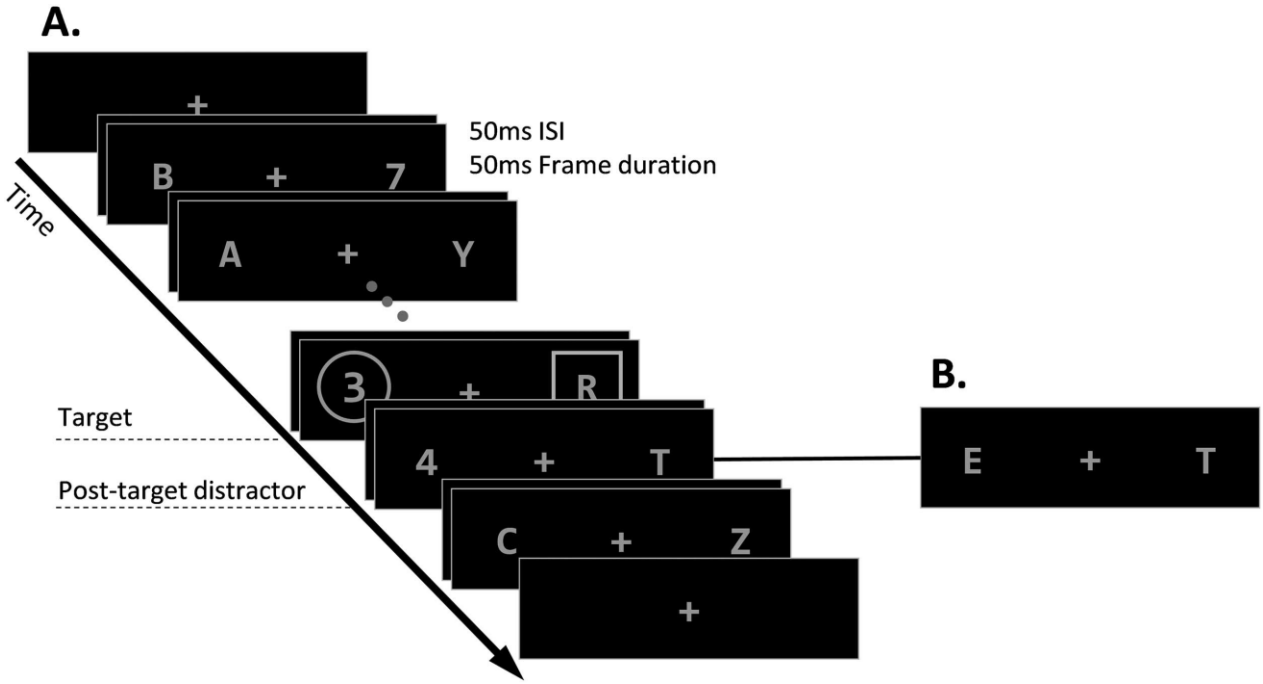


Figure 17. Illustration of the stimulus sequence in Zivony and Eimer's (2020a) Experiments 1. Participants had to report the target digit within one of two RSVP streams, defined by a predefined selection feature (e.g., circle/annulus). The target appeared at positions 5 to 8 within the stream and was followed by two additional frames. The posttarget frame contained a digit at the same location as the target on 75% of trials (A) and two letters on 25% of trials (B). ISI = interstimulus interval.

As discussed thoroughly in Chapter 2, there are several alternative accounts that attempt to explain the distractor intrusion phenomenon, such as those provided by Botella et al. (2001), Vul et al. (2009), and Zivony and Eimer (2020a, 2020b). In the current chapter, we will mainly refer to and compare the 2f-ST^2 computational model to Botella et al.'s (2001) model as well as to the theoretical accounts presented by Zivony and Eimer (2020a, 2020b). Importantly, the 2f-ST^2 computational model extends the ST^2 model (Bowman et al., 2008; Bowman & Wyble, 2007) and was previously presented in an initial version by Chennu (Chennu, 2009; Chennu et al., 2011). However, our current model supersedes it and in this thesis, when referring to the initial model (Chennu, 2009; Chennu et al., 2011), we will call it *initial 2f-ST^2*.

In this chapter, we will investigate the first two hypotheses of this thesis, i.e., that the binding of multi-dimensional stimulus features into single percepts depend on the timing of transient attentional enhancement (TAE, Hypothesis 1), and that the 2f-ST^2 model accounts for a broad range of findings obtained with distractor intrusion experiments (Hypothesis 2). To this end, we will first introduce how the current 2f-ST^2 model compares to the initial 2f-ST^2 (Chennu et al., 2011), describing the modifications that were made to the model, which

enabled it to account for the experimental paradigms adopted by Zivony and Eimer (2020a, 2020b) in addition to those adopted by Botella et al. (2001). In this context, we will replicate the main results of Chennu et al. (2011). The majority of these results will be presented very briefly, as an in-depth presentation of them has already been made (Chennu et al., 2011). However, it was important to demonstrate that our modifications did not weaken the model's capacity to simulate the experimental paradigms of Botella et al. (2001), for which the initial 2f-ST^2 was developed. We will therefore focus on providing metrics that illustrate how the current 2f-ST^2 model compares to the initial 2f-ST^2 with respect to simulating the main findings of Botella et al. (2001). We will furthermore provide novel virtual ERP patterns that explain seemingly counter intuitive reaction time (RT) patterns shown by Botella (1992), which were previously discussed by Chennu et al. (2011). Then the experimental paradigm adopted by Zivony and Eimer (2020a) will be investigated. In this context, we will introduce an ERP latency analysis that differs methodologically from the analyses provided by Zivony and Eimer (2020a), because, as we argue, our method is more appropriate for the subsequent comparison to the 2f-ST^2's virtual ERPs. Moreover, we will provide virtual response distributions as well as virtual ERP patterns to establish that the 2f-ST^2 model also qualitatively replicates the empirical results obtained by Zivony and Eimer (2020a). Finally, we will present an empirical analysis of the EEG dataset obtained in the context of Zivony and Eimer's (2020a) first experiment to probe a hypothesis that is implied by the 2f-ST^2's architecture, namely, that the N2pc and P3 ERP components are temporally correlated, which is implied by Hypothesis 1 of this thesis.

Methods

Computational Model

In general, this thesis will focus on presenting and discussing the modifications made to the initial 2f-ST^2 and a series of novel analyses that add to Chennu et al.'s (2009; 2011) work. We will thus exclude a detailed introduction of how the initial 2f-ST^2 superseded the ST^2 model. Still, we will replicate some key findings shown in the context of the initial 2f-ST^2 (Chennu, 2009; Chennu et al., 2011) to demonstrate that our updated model is still able to generate these, even after the implementation of our modifications.

In addition, and in contrast to the initial 2f-ST^2 (Chennu, 2009; Chennu et al., 2011), which was only run once, we ran the model a total of five times for all configurations, re-seeding the random number generator (rng) each time. The only results which were obtained after a single run (using the same rng-seed as the initial 2f-ST^2) are those in which we present individual neurons' traces (e.g., membrane potentials) in Chapter 5, as saving these increases computational effort drastically and was hence not feasible with five runs. We will mention single-run results explicitly wherever they occur. The 2f-ST^2 model's architecture (Figure 18) will be briefly introduced first.

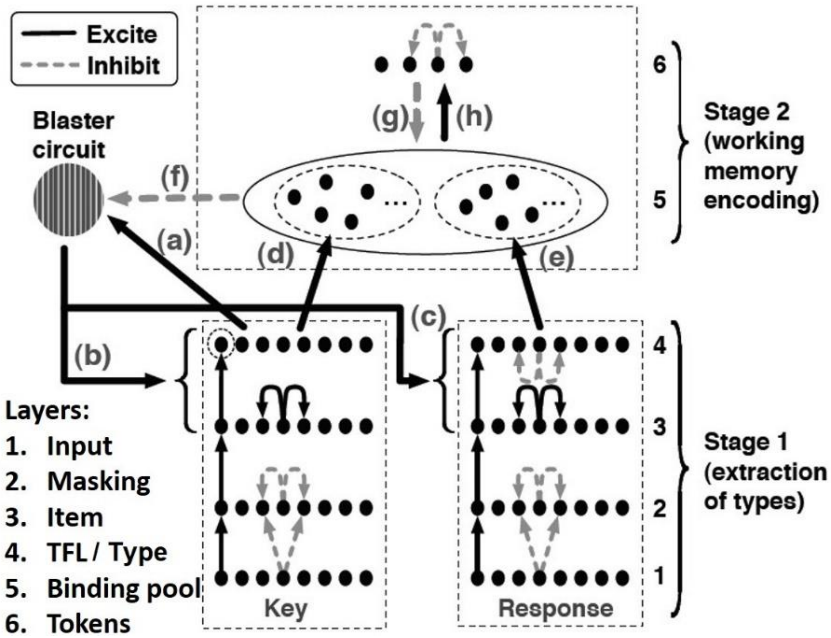


Figure 18. The 2f-ST^2 Model Architecture showing two Stage 1 pathways, Blaster Circuit, Binding Pool and Tokens. Stage 1 implements a key and a response pathway in which respective feature dimensions are extracted and processed. The key type meeting task-demands triggers the blaster circuit in the TFL/ type layer of the key pathway. The blaster boosts activation levels of both pathway's item and TFL/ type layers. The currently active token is finally bound to a pair of key and response types in Stage 2, meaning that the item was encoded in WM. Adopted from Chennu et al. (2011).

Architecture

The overall architecture and dynamics of the 2f-ST^2 model (Figure 18) mirror that of the original ST^2 model (Bowman & Wyble, 2007) closely, again implementing a parallel Stage 1 extracting and processing types, a sequential Stage 2 binding types to tokens, and a blaster circuit. The 2f-ST^2 model extends the original model's (Bowman & Wyble, 2007) architecture, adopting two pathways in its Stage 1, which enables it to extract two kinds of types, corresponding to different feature dimensions, for each item in the RSVP stream simultaneously. The key and response pathways extract and process items' key and response features, respectively. Stage 2 is extended in a similar fashion in the 2f-ST^2, now consisting of two binding pools, one for each feature dimension, which allows a working-memory token to be associated with a pair of types indicating the encoded item's key (e.g., colour) and response (e.g., shape) feature. To reiterate, types and tokens are fundamental concepts in the 2f-ST^2 model as well as in its predecessor models (Bowman & Wyble, 2007; Wyble et al., 2009). Types refer to the featural representation of a stimulus, encompassing all featural properties of it, whereas tokens inform about the episodic context in which a stimulus occurs, thus informing about the temporal sequence of stimuli.

The key pathway extracts items' key features and hence conceptually resembles Botella et al.'s (2001) Module K. Importantly, the RSVP stream's task demands (dashed circle around the left key TFL/ type node in Figure 18) are implemented in this pathway, specifically in its TFL/ type layer. The item that carries the task-relevant key feature (e.g., the colour red in Figure 14) activates the blaster circuit through the connection (a) of Figure 18. Therefore, distractors are modelled as items that lack the key-feature demanded by a given task (and, for example, correspond to black or yellow items) and hence do not activate the blaster circuit. Once the blaster is activated, and similar to the original ST^2 model (Bowman & Wyble, 2007), the activation of all nodes of the last two Stage 1 layers (item and TFL/ type) are enhanced. Importantly, in the 2f-ST^2 model the blaster enhances key and response item and TFL/ type layers simultaneously and to the same extent. This is illustrated as connections (b) & (c) in Figure 18. Additionally, the implementation of the blaster enhancing later Stage 1 activation levels resembles the TAE processes mentioned in Zivony and Eimer's (2020a, 2020b) accounts. Similar to the original ST^2 model (Bowman & Wyble, 2007), the blaster's enhancement means that the most strongly active node of each pathway's TFL/ type layer becomes sufficiently active, crossing an *effective* access threshold and initiating activation in their respective binding pool, illustrated as connections (d) and (e) in Figure 18.

Once respective binding pool nodes have been excited according to the latter process, successful binding between one key type, one response type and one token occurs. The response pathway dynamics depend on whether Botella et al.'s (2001) or Zivony and Eimer's (2020a, 2020b) experimental paradigm are to be modelled. These dynamics occur in the response TFL/ type layer, in which all response types can be bound to the token via their respective binding pool units when replicating Botella et al.'s (2001) experimental paradigm. However, when modelling Zivony and Eimer's (2020a, 2020b) paradigm, only the target and the +1-distractor can be active in this layer, as the remaining items resemble target category-nonmatching distractors. In both cases, lateral inhibition between category-matching items (i.e., all items in Botella et al.'s (2001) paradigm and the target & +1-distractor in Zivony and Eimer's (2020a, 2020b) paradigm) in the response TFL/ type layer generates a competition between types to excite their respective binding pool units. The target versus intruder competition examined by Zivony and Eimer (2020b) can be seen to be consistent with this aspect of the 2f-ST^2 model. Which response type is ultimately bound to the currently active token therefore depends on which TFL/ type response node was active the most at the time of the blaster enhancing late Stage 1 layers. This mechanism further enables the 2f-ST^2 model to commit to distractor intrusion errors, which occur whenever a distractor item's response type wins this pathway's competition (explained in detail in the next section).

Tokens in the 2f-ST^2 principally work in the same way as in the original ST^2 model (Bowman & Wyble, 2007). Specifically, the currently active token inhibits the other tokens' binding pool nodes whenever the model is encoding (Figure 18's connection (g)) and is excited by one key and one response binding pool node during binding (Figure 18's connection (h)). The successful binding of a given token to a pair of key and response types thus allows the 2f-ST^2 model to store episodic and ordered representations of items and hence constitutes WM encoding in the model.

Furthermore, it should be noted that, as in previous ST^2 models (Bowman & Wyble, 2007; Wyble et al., 2009), the 2f-ST^2 model implements gate-trace node-pairs. Gate (on) nodes (or neurons) excite trace (off) nodes after being externally excited themselves. Trace nodes in turn accrue the excitation they receive from gate nodes and, after crossing a threshold, inhibit the gate node strongly, suppressing all gate node activation. Trace nodes are for example important in informing that a certain event has terminated, for example that a given token has completed binding to a pair of key and response types.

We made a few minor changes to the 2f-ST^2's architecture compared to that of the initial 2f-ST^2. For a full list of how the original ST^2 model (Bowman & Wyble, 2007) was extended architecturally to generate the initial 2f-ST^2, see Chennu et al.'s work (Chennu, 2009; Chennu et al., 2011). Compared to the initial 2f-ST^2, the model architecture was modified as follows.

The item layer's bias was set to zero, as this distorted neurons' membrane potentials, especially when introducing a substantial time-delay to key pathway processing, which was required in some configurations of the model. Response pathway item layer neurons were furthermore not inhibited by their corresponding off-neurons (i.e., the item layer off-neurons originally included in Bowman and Wyble's (2007) ST^2-model or those in the initial 2f-ST^2 (Chennu et al., 2011)). We implemented this change to simplify the inhibitory dynamics of this pathway, which due to the lateral inhibition present in its TFL/ type were already rather complex. As in the initial 2f-ST^2 (which exclusively modelled such experiments), response pathway TFL/ type task demands were absent when modelling experiments in which all stimuli could plausibly have been targets (as in Botella's paradigms (Botella et al., 2001; Botella & Eriksen, 1992)). However, the current model implemented a task filter in the response pathway TFL/ type layer when modelling Zivony and Eimer's (2020a) experiment in which only two stimuli met task demands (i.e., the two digits in the stream that were adjacent in time in Figure 17). To this end, the 0 and +1 response types, which model the target stimulus and the intrusion distractor following it, met task demands and thus had different response task demand unit (TDU) values than the remaining (distractor) types. TDUs model task relevance of virtual stimuli by adding a bias to respective TFL/ type layer neurons. Across both pathways, distractor stimuli's TDUs bias TFL/ type layer processing negatively with a bias of -.3. Response pathway TDUs that met task demands (i.e., in the case of the 0- & +1-types when modelling Zivony and Eimer's (2020a) experiment) do *not* bias their TFL/ type layer neurons (bias = 0). Note that in the key pathway, task-relevant types had TDU biases of +.003, thus positively biasing their respective TFL/ type layer neuron. However, due to the lateral inhibition present in the response pathway TFL/ type layer, a lower bias value of task-relevant response TDUs was required.

For an overview of the 2f-ST^2 model's key terms and their definitions see the glossary in the end of this methods section (Table 1).

Dynamics & Configuration

We kept the model dynamics unchanged from the initial 2f-ST^2 (Chennu et al., 2011) and hence varied targets' key feature input-strengths across different simulation runs, while keeping all distractors' key feature input-strengths fixed throughout runs. Note that variability in target key feature input-strengths across runs induces variability in blaster firing latencies due to the model's architecture. Response feature input strengths varied similarly across runs for the target as well as its proximal distractors' (2 positions in either direction, i.e., -2 to +2) and remained fixed for all other distractors.

With respect to the model's configuration, there are three important delay parameters: τD, τK and τR. τD (random delay) is repeatedly sampled for each respective item in a virtual trials' RSVP stream and follows a gaussian distribution with a mean of zero and a standard deviation of 4 time-steps (~20 ms-equivalents). τK (key delay) and τR (response delay) add a positive or negative time-delay to the processing of the key and response pathway, respectively. Importantly, τK and τR are set when running the model and *remain fixed throughout a complete simulation run of the model.* Both parameters' 'default' value is 0 and changing them enables the replication of a variety of experiments involving manipulations that affect the speed of key or response feature processing in human cognition. A few important delay configurations will be introduced after an explanation of how intrusion errors occur in the 2f-ST^2 model, which will be provided next.

Intrusion Errors in the 2f-ST^2 model

We provide a schematic illustration of how the 2f-ST^2 model generates correct and intrusion responses in Figure 19. Time progresses from top to bottom panels in these plots and the left and right columns illustrate how correct and intrusion reports are generated, respectively. The model's key and response pathways depict individual key and response features, respectively, with key features being represented by black, red, and yellow circles and response features being represented by black rectangles labelled B, S, and F. Importantly, we used gradients to represent activation strength of an item's representation in the model's hierarchy at a given point in time (throughout a trial), with more transparent gradients meaning lower activation strengths. For example, the small black oval in the second row of Figure 19's left column is more solid towards its top and more transparent in its bottom area, which means that the black item's key feature has a rather strong activation in the key TFL/ type layer, but an already decaying activation in the previous (lower) key item layer.

The right column shows a +1-intrusion response, specifically, a trial in which the distractor stimulus presented *immediately after* the target was reported. Concretely, we show the example of coloured letters being presented in the (virtual) RSVP stream. In the illustrated case, a red S item represents the target (key-feature in the left pathway = red circle & response-feature in the right pathway = S-shape) and a yellow F represents the +1-intruder item (key-feature in the left pathway = yellow circle & response-feature in the right pathway = F-shape). The intrusion response shown here occurs due to delayed key-feature processing. As shown in the second row of Figure 19, this delay to key-feature processing means that the blaster is triggered at a later point in time preceding the intrusion response (right column). At this later time-point, the F-type of the response pathway is most active (illustrated by F rectangle being more solid or darker). This is in contrast to correct reports (left column), in which the correct S-type had the highest activation. The bottom two rows show how this difference in response-type activation strengths translates to either one response type (correct-S, left, or intruder-F, right) being bound to the token and hence encoded into WM. The third row illustrates how the blasters' enhancement strengthens the activation of *all types* (note how *all* gradients get more solid) in the last two stage 1 layers, leading the *strongest TFL/ type* (last stage 1 layer) unit to cross the threshold to binding pool entry. The last row illustrates how the corresponding binding pool units are bound to the token, generating a correct and intrusion response in left and right columns, respectively. As illustrated in the figure, these binding pool units have a conjunctive tuning, responding to the conjunction of a type and a token. Figure 19's bottom row further demonstrates that neurons in the response TFL/ type layer laterally inhibit each other. Naturally, the response TFL/ type neuron that has the strongest activation level when the blaster fires will laterally inhibit the other neurons in this layer more strongly, leading to the 'competition' to enter the response binding pool (and to bind to the currently active token) to have an unambiguous outcome (as illustrated by the rather large contrast in gradients between response TFL/ type neurons at this final stage).

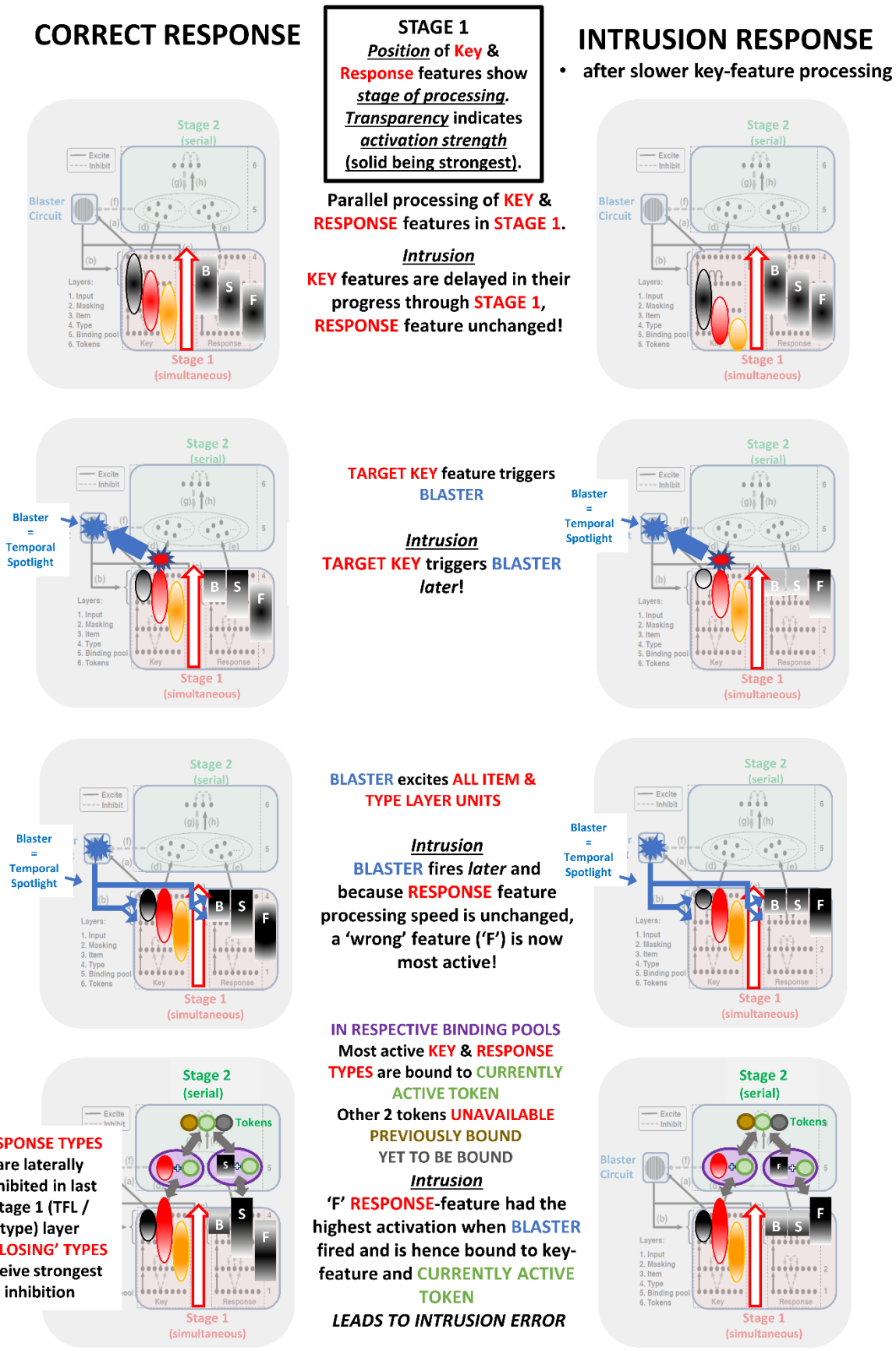


Figure 19. Schematic outline of how the 2f-ST^2 model generates correct and intrusion responses. The left and right columns show how correct and intrusion reports are generated, respectively. The top row illustrates parallel propagation through stage 1 layers, being typically delayed for the key pathway in (post-target) intrusion trials. The second row illustrates how this translates to the intruder (F) type being the most active at the time-point of blaster-firing. The third row illustrates how the blaster enhances *all types* in later stage 1 layers and how this leads to different types crossing the threshold to the response binding pool. The last row illustrates how as a result, different binding pool units are bound to the token, leading to a correct response in the left case and an intrusion response in the right case. Note that the last row also illustrates how response types are laterally inhibited by each other in the last stage 1 layer (TFL/ type). Lateral inhibition suppresses 'losing' (i.e., unbound) types.

The previous paragraphs, as well as Figure 19, illustrate the significant role of the timing of blaster enhancement with respect to the response that will be provided by the 2f-ST^2 model. The key and response types that are most active when the blaster fires will be bound to the currently active token. τK and τR, the fixed delays to key and response pathway processing, hence impact model dynamics by modulating the latter, i.e., by changing the distribution of activation strengths across response TFL/ type layer neurons, and particularly the most active neuron, at the time of blaster firing. τK accomplishes the latter by changing the latency of blaster firing itself, since a delay in key pathway processing means that the task-relevant key TFL/ type neuron excites the blaster later. τR on the other hand modulates the distribution of activation strengths across response TFL/ type neurons and the neuron being most active when the blaster fires. For example, a positive τR value means that response pathway processing commences *later*. Given a fixed blaster-firing latency, delayed response pathway processing would imply that *pre-target* intrusions are more likely. This is because, compared to a τR of 0, the virtual RSVP stream's response features had less time to propagate to the later Stage 1 layers of the response pathway and, thus, it is more likely that pre-target TFL/ type neurons are active strongly when the blaster fires. τK & τR are further instrumental for replicating empirical studies with the 2f-ST^2 model, since these parameters enable us to directly replicate key and response feature salience, which are often manipulated in experiments. This is because a very salient key or response feature is argued to be processed easily and swiftly by participants, which corresponds to a low (or negative) value in the respective pathway's delay. In contrast, stimulus-features that are difficult to notice (i.e., are not salient) will be processed slower, which corresponds to that feature being processed with a delay in the 2f-ST^2 model. Moreover, and due to the significance the timing of blaster firing has for the responses provided by the 2f-ST^2 model, we argue that each case in which the model should qualitatively replicate empirical studies not only provides evidence in favour of the model's architecture, but further provides evidence in favour of Hypothesis 1 of this thesis, i.e., that the binding of multi-dimensional stimulus features into single percepts during rapid stimulus streams depends on the timing of TAE deployment.

In the next section, we will introduce a couple of empirical studies that manipulated key and response feature salience and, further, explain how we configured τK & τR to resemble these studies.

Key and Response Feature Manipulations

To replicate the results presented with the initial 2f-ST^2 (Chennu et al., 2011), we implemented no task-filtering in the response TFL/ type layer. The initial 2f-ST^2 replicated a series of experiments conducted by Botella et al. (2001), who adopted different levels of word-frequency to modulate key or response feature processing speeds in human cognition. We introduced these experiments of Botella et al. (2001) earlier in this chapter (Figure 15 & Figure 16). In this context, LF and HF represent low- and high-frequency word (occurrence) conditions. In the model, τK and τR were both set to zero to replicate Botella et al.'s (2001) HF condition of Experiments 1A and 2 as well as the late key (LK) condition of Chennu et al.'s (2011) fortunate conjunctions experiment (introduced below). The LF conditions of Botella et al.'s (2001) Experiments 1A and 2 were replicated by adopting delay-pairs of $\tau K = 8$ / $\tau R = 0$ and $\tau K = 0$ / $\tau R = 2$ time-steps, respectively. The early key (EK) condition of Chennu et al.'s (2011) experiment was replicated using $\tau K = -8$ / $\tau R = 0$ time-steps. Note that one time-step of fixed delay in the model approximates 5 ms-equivalents.

In order to replicate the much slower human key feature processing speed of Zivony and Eimer's (2020a) Experiment 1, i.e., a dual-stream RSVP variant with an annulus as the key feature, we introduced a τK of 24 time-steps (~120 ms-equivalents) on top of the active task-filter in the response TFL/ type layer, which ensured that only two response types could progress to the binding pool. We also implemented a mechanism that ensures that the impact of our task-demand units (TDUs) on both pathways' TFL/ type layer neurons, which implements the layers' task-filter via a positive or negative bias, is only active *after* any delay in key pathway processing. This was done to prevent cases with a large key delay in which TDUs affected TFL/ type-layer neurons *before* the virtual RSVP-stream commenced.

Virtual Event-Related Components (vERPs) of the 2f-ST^2 model

The 2f-ST^2 model allows one to generate virtual ERP (vERP) components to compare to human ERPs. vERPs are based on the postsynaptic activation potentials of virtual neurons, which are the sum of a neuron's presynaptic activation multiplied by corresponding weights. The virtual N2pc (vN2pc) component is computed as the postsynaptic activation of the blaster's output neuron. The virtual P3 (vP3) is computed as a sum across four layers. Both (i.e., key and response) pathways' TFL/ type layers as well as both (i.e., key and response) binding units' gate projections are considered in the vP3 and the contribution of all four layers is equal (i.e., each one of the four layers contributes 25% to the final vP3 trace). In comparison with previous ST^2 modelling of virtual P3s (Chennu et al., 2009; Craston et al.,

2008), we excluded the item layers from computations of the P3. This reflects the TFL/ type layers and binding pool most naturally corresponding to the later cognitive processes related to working-memory encoding, which are typically argued to underlie human P3s (Akyürek et al. (2010), Polich (2007), although see Jones et al. (2020) and Pincham et al. (2016) for a linking of the (breakthrough) P3 to conscious perception. The item layer, on the other hand, reflects cognitive processes that are prior to encoding and hence were excluded. In addition, we stress that since virtual reaction times (RTs) in the 2f-ST^2 model are based on the currently active token's trace neuron, which is excited by both pathways' binding pool gate units, the 2f-ST^2 model's RTs will be highly correlated with the temporal dynamics of its virtual P3 component. Note that half of the 2f-ST^2 model's vP3 is comprised by binding pool gate unit activation. We will provide further remarks on modelling RTs in this way later on. In humans it has been argued that the processes underlying a response are serial and that P3 and RT latencies mark the completion of different stages of this processing pipeline (Ford et al., 1980). It can therefore be expected that P3 and RT latencies in humans should show some level of correlation, too. However, numerous factors are likely to affect this correlation, such as anatomical brain differences between subjects or variability in the temporal dynamics of motor processes due to interindividual differences or different task demands (e.g., the task requiring a motor, oral or visual response).

We implemented another modification to the model's virtual ERP components to resemble human ERP dynamics more accurately. This modification is a novelty and has not been adopted in any previous iteration of ST^2-vERPs (Chennu et al., 2009; Craston et al., 2008). In previous iterations, vERPs have been defined as the amplitude of postsynaptic potentials of virtual neurons across a given layer. In practice, this meant reading out virtual neurons' postsynaptic potentials and making sure that if multiple layers are summed (as in the vP3), the layers involved all contribute equally to the component (i.e., possible amplitude differences between layers of a given vERP component were adjusted for). However, the human EEG signal does not exactly, or exclusively, capture postsynaptic potentials. Human EEG measures, on a macro-scale, voltages arising from volume conduction, which induces charged electrons to reach the electrodes on the scalp. The currents that underlie this volume conduction originate from dipoles in the human brain that are the result of neural activity. These dipoles are in turn generated by micro-current sources at cell membranes (Nunez & Srinivasan, 2007). On this micro scale, the current is affected by sources of mixed signs, e.g., synchronous postsynaptic potentials, local inhibitory synapses as well as passive membrane (return) currents needed for current conservation (Nunez & Srinivasan, 2007). The current

flows through apical dendritic trees of cortical pyramidal cells (Michel & He, 2019) and, on an intermediate meso-scale, generates the dipole moment in a given cortical column, which reaches the scalp and is ultimately measured by the EEG electrodes attached to it. However, as described by the Forward Problem of EEG, the current does not propagate homogenously, as it propagates through different kinds of tissue (such as the scalp, skull, cerebrospinal fluid & brain), which all exhibit different conductivity levels and hence attenuate the current to different degrees (Michel & He, 2019).

We therefore computed the approximate derivative of our original vERPs (i.e., layer-wide postsynaptic potentials), using MATLAB's 'diff' function, since (modulo a multiplicative constant, the capacitance) current is the time derivative of the voltage, i.e., how it changes through time. We further 1-D smoothed the resulting time-series with a gaussian kernel. Smoothing was also computed in MATLAB using the 'gausswin' and 'filter' functions. We decided against zero-phase filtering (e.g., using MATLAB's filtfilt function), as it shifted vERP's peak-latencies more substantially. The utilized gaussian kernels were standardized to have an area-under-curve (AUC) equal to one prior to any smoothing. The denominator coefficient *a* of the filter-function was furthermore chosen in a way that led to amplitudes similar to those exhibited in our original vERPs (*a* being .05 and .025 for the vN2pc and vP3, respectively). Importantly, this coefficient was constant across different analyses of the same vERP component and only differed *between* components (but never between different response-conditions or model-configurations, for example).

These smoothed vERPs carry two important advantages. First, they resemble the current that is being picked up by the EEG-electrodes more closely and, simultaneously, capture important properties implied by the EEG signal being based on dipoles, e.g., that a polarity-swap (i.e., a negativity following a positivity or vice versa) implies a shift in the direction of ion-flow that is currently affecting the current the strongest. Nonetheless, we acknowledge the value of the previous approach of specifying vERPs as virtual postsynaptic potentials, as these vERPs reflect the internal processes and dynamics of the model most directly. For the reasons above, as well as for the sake of thoroughness and transparency in general, we will consistently present the original (unsmoothed) vERPs as well as the smoothed difference vERPs side-by-side.

Chennu et al.'s (2011) Experimental Methods

We will replicate the fortunate conjunctions experiment presented in Chennu et al. (2011) (alongside the main simulations that replicate Botella et al.'s (2001) major findings) to

demonstrate that our current 2f-ST^2 model is able to still replicate those experimental paradigms for which the initial 2f-ST^2 was developed. Chennu et al.'s (2011) experiment involved a sample size of 14 students who viewed RSVP streams of dark grey letters, digits and symbols that were surrounded by coloured squares. The experimental setup allowed for distractor intrusion errors of positions -2, -1, +1, and +2 and incorporated a forced-choice response screen, which in addition to the target and randomly selected distractor items also included the option 'None of the above'. For a complete description of this experiment, see the original paper (Chennu et al., 2011).

Zivony and Eimer's (2020a) Methods

Experimental Methods

We will simulate results of a dual-stream RSVP study: Experiments 1A & 1B (1B being a direct replication of 1A) of Zivony and Eimer (2020a). See the original paper for full details about their methodology. In their experiments (see Figure 17), participants were presented with two RSVP streams with lengths of 8 to 11 frames at equal distances from a fixation cross in the centre. Grey stimuli were presented in sequence on a black screen, with letters as distractors and digits as targets. The target digit was presented at positions 5 to 8 of the streams, differentiated by a surrounding annulus or square. Participants had to report the target as accurately as possible after each trial terminated. In target frames, a distractor letter was also presented in the other RSVP stream surrounded by either an annulus or square (which of the shapes identified the target digit was always pre-specified). The frame preceding the target frame always consisted of two letters (one in each stream) and earlier pre-target frames were equally likely to contain two letters or one letter and one digit (to ensure that attentional allocation was placed according to the annulus or square rather than alphanumerical category, i.e., subjects did not just search for the first digit in the stream). The frame that followed the target frame included another digit at the same location on 75% of trials. In the remaining 25%, a distractor letter was presented instead (Zivony & Eimer, 2020a). Hence, the annulus and the square were the key features in this setting and digit identity the response feature. Each frame was presented for 50 ms, followed by an inter-stimulus-interval (ISI) of 50 ms. Targets were equally likely to be presented in the left or right RSVP stream in each trial. The experimental paradigm is depicted in Figure 17.

EEG Methods

Event-related potentials (ERPs) were computed separately for trials in which participants reported the target-digit correctly (correct trials) and for reports of the post-target

digit-distractor stimulus (intrusion trials). Incorrect trials, i.e. with reports of neither the target nor the post-target digit-distractor, were excluded in ERP analyses. N2pcs, suggested to index attentional processing (Eimer, 1996; Kiss et al., 2008; Luck & Hillyard, 1994), were computed as the contralateral – ipsilateral difference wave between PO7 & PO8 electrodes w.r.t. the location of the target (e.g., PO8 – PO7 if the target was presented in the left RSVP stream). The P3 component, suggested to index WM consolidation (Polich, 2007), was defined as the ERP amplitude at the Pz electrode. Hence, we retained the original paper's (Zivony & Eimer, 2020a) EEG methodology in general, with the addition of a 25 Hz low-pass filter for P3s and a larger time-window of interest (because the original paper did not analyse P3 components). Moreover, we adopted a different method, dynamic time warping (DTW), to measure latency-differences in ERPs between correct and intrusion trials. We discuss this next.

Dynamic Time Warping as a Measure of Latency Differences

We analyse ERP components' latency differences between correct and intrusion conditions with a different method than Zivony & Eimer (2020a), as the authors were specifically interested in N2pc onset latency differences between correct and intrusion trials. Our present analysis, however, aims at replicating the *whole component's* latency characteristics with our model's vERPs, not just their onset's. We thus chose dynamic time warping (DTW), as this approach enables the latency of ERP components to be measured over a *region*. This contrasts with most commonly used latency measures, such as peak latency, fractional peak latency, and fractional area (Handy, 2005; Kiesel et al., 2008; Luck, 2014), which are computed with reference to a given point of the ERP wave. For discussion of the benefits of DTW compared to other EEG latency approaches, see Zoumpoulaki et al. (2015).

DTW measures the similarity between two time-series through aligning one time-series (called the *query*) to another (the *reference*). This alignment is *optimal*, meaning that a distance matrix is built from all points of the reference & query time-series and a warping path is chosen through this matrix such that the *minimal cumulative distance* is guaranteed. We provide an illustration in Figure 20, which shows our data's DTW analysis for an electrode that we are not interested in: Fz. We present the original (non-standardized) Grand Averages as well as the DTW warping path in Figure 20. When applied to ERPs, DTW provides a measure of latency, as well as succession (i.e., the property of one time series being earlier than the other). As proposed in Zoumpoulaki et al. (2015), we use the

standardized area between the warping path (blue line in Figure 20) and the main diagonal (red line in Figure 20), which we will henceforth call the DTW area-difference, for our statistical analysis of latency difference between components. This area measure indicates succession, as a positive value would imply that the reference time series (used for alignment, plotted on the y-axis in Figure 20) was *overall earlier* in time compared to the query time-series (on the x-axis in Figure 20). The DTW area-difference is plotted as a light green area in Figure 20. We furthermore compute the distance-distribution between x- & y-coordinates of the warping path. That is, each (x,y) coordinate on the warping path, has a horizontal distance to the main diagonal. The set of all such horizontal distances gives the distance-distribution. The median of this distance distribution allows the computation of components' latency difference in milliseconds by dividing the median by the sampling rate divided by 1000 (see Zoumpoulaki et al. (2015) for a formal comparison of the median to other options). We implemented a time-interval of interest of 150-400 ms for the N2pc (Eimer, 1996) and 250-800 ms for the P3 (Polich, 2007). DTW analyses were performed in MATLAB 2020b using the built-in dtw function.

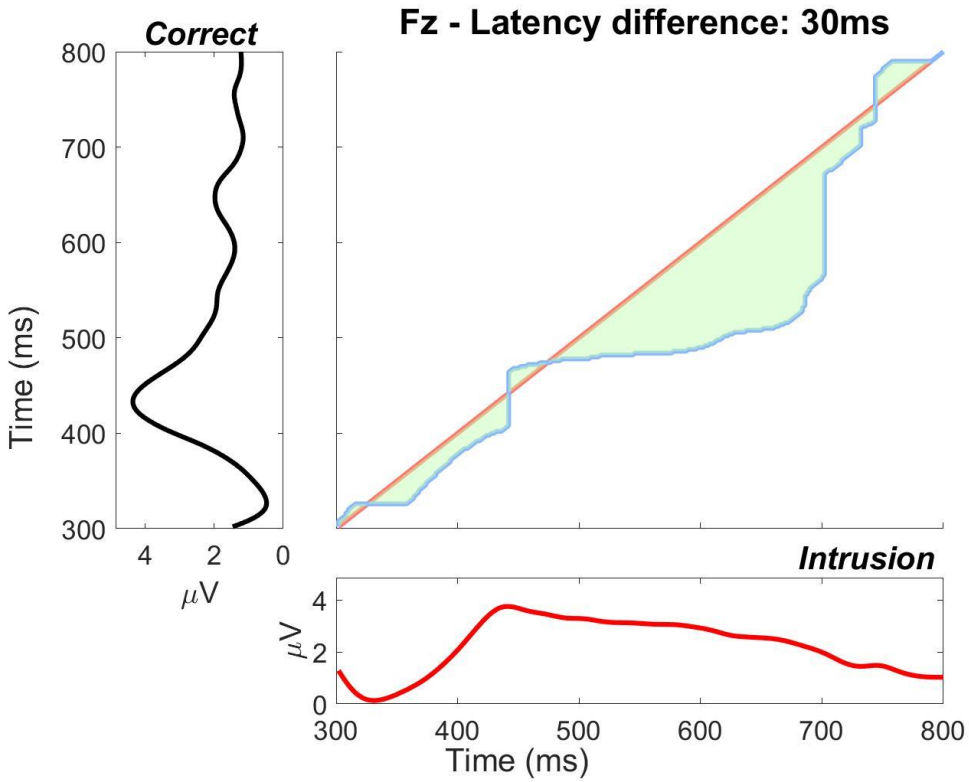


Figure 20. DTW Analysis at the Fz electrode. The reference time-series (correct trials' ERP, y-axis) is used to optimally align the query (intrusion trials' ERP, x-axis). The warping path is plotted as a blue line, the main diagonal (which would imply identical latency time-series if it were the warping path) is plotted in light red and the DTW area-difference used for statistical analysis is plotted as a green area. The warping path being located under the main diagonal implies that the reference time-series (correct) preceded the query (intrusion).

We assessed statistical significance of these DTW area-differences with a two-tailed permutation test. We considered a one-tailed test, due to the a-priori hypothesis that intrusion trials should lead to later ERP components than correct trials, which was based on the N2pc findings of Zivony & Eimer (2020a). However, we decided against a one-tailed test as that would have risked statistical double-dipping (Kriegeskorte et al., 2009), since the dataset upon which the a-priori hypothesis was based would be the same dataset as is analysed by us with DTW. We first implemented the standard paired t-test permutation procedure, on our subject-level data, where each subject has an ERP for correct and for intrusion. On each iteration of this permutation procedure, a "fair coin" is flipped for each subject; if this comes up heads, the ERPs for this subject are flipped between groups (correct to intrusion and intrusion to correct), if it comes up tails, the ERPs remain as they are. This generates a permuted data set. We then computed the permutation grand average ERP waves by taking the average wave across subjects for (permuted) correct- and intrusion-conditions separately. We subsequently performed the DTW analysis and computed ERP-component DTW area-differences as described above. We repeated this procedure 10000 times, which generated a distribution of DTW area-differences under the null. Finally, the p-value of our true observed DTW area-difference value was computed as the proportion of absolute (hence a two-tailed test) permuted (i.e., null) DTW area-differences *larger than* our true observed value. This approach is exactly as proposed previously by our lab (Zoumpoulaki et al., 2015), as we used the DTW area-difference value for all statistical analyses and the median of the DTW distance-distribution only to estimate components' latency differences in milliseconds.

Correlated Latencies between the N2pc and P3 component

Our model's architecture implies a specific inter-dependence between the N2pc and P3 ERP-components' latencies. This inter-dependence arises since activation levels propagate hierarchically through the model and, especially, due to the critical role of the blaster, which enhances activation levels of later stage 1 layers, which in turn propagate to processes in stage 2. Therefore, the timing and extent of blaster activation (vN2pc) largely determine the timing and extent of TFL/ type layers' activation levels as well as subsequent virtual working-memory encoding processes occurring in the binding pool (vP3). This characteristic of our model therefore implies a temporal correlation between the vN2pc and the vP3. For example, in a trial in which the blaster fires later, the blasters' enhancement will affect late stage 1 (item- & TFL/ type) layers later, which implies delayed stage 2 processes, such as the binding of types to tokens. Hence, latency-differences in our *virtual* N2pcs (blaster-output layer) are propagated to our vP3s (later layers), making the latencies of our two vERPs correlated.

Crucially, extrapolating this to human cognition, one expects human N2pc and P3 latencies to be correlated as well (which would provide evidence in favour of Hypothesis 1 of this thesis). This hypothesis was examined in two ways: 1) by showing that *both* the N2pc and the P3 are delayed in intrusion trials; and 2) with the following bootstrap analysis.

We stress that even though the correlation between N2pc and P3 latencies in human cognition is expected due to the hierarchical propagation of activity through the 2f-ST^2 model, this prediction is not unique to our computational model. The field of neuroscience has generated a wide-spread consensus about the cascaded nature of the brain's processing hierarchy (Mcclelland, 1979), i.e., that each cognitive process in the sensory pathway only starts when the previous process has completed. It can therefore be argued that correlated latencies of the N2pc and P3 component constitute common sense in the neurosciences, since selective attention processes (N2pc) precede WM-encoding (P3). Therefore, if some experimental trial should show an early N2pc, selective attention was operating early, allowing for WM-encoding to commence earlier, which should be reflected in an earlier latency of the P3. Still, we emphasise the value of the present analysis probing this expected correlation analysis empirically, due to the following two reasons. First, to our knowledge this analysis is the first of its kind, introducing a novel approach to assessing the extent of temporal correlation of any two ERP components. Second, this analysis assesses the extent of the cascaded nature of the brain's processing pipeline (Mcclelland, 1979) mentioned above. Critically, the present analysis is therefore able to potentially establish yet unknown factors that contribute to the extent of temporal correlation between two sequential and related cognitive processes. If, for example, one would show that the N2pc and P3 components are indeed correlated in time, but that this correlation is modulated by factors such as task demands, cognitive load, damage to certain cortical areas, or interindividual differences such as in age or brain anatomy, this would allow for novel kinds of insight about the cascaded nature of the brain.

The present analyses were conducted on standardized (i.e., z-scored) single-subject ERP-components and, identically to our DTW analyses, using a time-interval of interest of 150-400 ms for the N2pc and 250-800 ms for the P3. First, we randomly selected subjects with replacement 23 times, replicating the number of subjects in our other analyses. We then computed bootstrap across-subject Grand Average ERP waves for our N2pc- and P3-components separately. *Importantly, the same bootstrap sample of subject-replications was used for the N2pc and P3* (that is, if subject *i* appeared *k* times in the N2pc grand average,

they also appeared *k* times in the P3 grand average). We subsequently performed a DTW-analysis, akin to the one between correct and intrusion trials' (true observed) ERPs described in the previous paragraph, but now between pairs of true-observed and bootstrap Grand Average ERPs. That is, we are assessing the latency difference of each bootstrap sampled grand average to the central tendency estimate, which is the true observed grand average. This analysis was conducted separately for the N2pc and the P3 component. It therefore yielded one DTW area-difference measure (relative to grand-average) for the N2pc and one for the P3. We repeated this process 10000 times and z-scored the two distributions of DTW area-differences. Correlation coefficients were computed after Pearson as well as Spearman between the N2pc and P3 DTW area-difference distributions. A significant positive correlation would provide support for our hypothesis of a correlation between the N2pc and P3 components. This is because such a correlation would mean that if the bootstrap N2pc is earlier (or later) than the true observed N2pc, this shift in time translated to the P3 component. To stress, as just emphasized, the bootstrap samples were always matched between N2pc and P3 in each of our 10000 repetitions, pairing N2pc area-differences with P3 area-differences, enabling the correlations to be calculated. We performed this analysis for correct and intrusion trials separately to prevent the correlated latencies due to the bootstrap sampling being driven by the delayed N2pc and P3 in the intrusion relative to the correct condition, which we will show in our first dynamic time-warping analysis reflects coupling between N2pc and P3.

Before presenting the results of this chapter, we provide a glossary of key terms relating to the 2f-ST^2 model and the distractor intrusion phenomenon in Table 1.

Term	Definition
Average Position of Intrusion (API)	Measure informing about the locus of intrusion responses, computed as the sum across all intrusions' positions multiplied by the proportion of that given intrusion response with respect to the total frequency of intrusions. The API thus weighs +/-2 intrusions more heavily than +/-1 intrusions.
Blaster	Implementation of transient attentional enhancement processes in the 2f-ST^2 model. The blaster is excited by activation of the TFL/ type layer neuron in the key pathway that meets task-demands. The blaster in turn enhances activation levels across both pathways' item and TFL/ type layers.
Gate (On Node)	Gate (on) nodes are part of of gate-trace pairs. After receiving external excitation, gate nodes in turn excite trace (off) nodes. Trace nodes accrue excitation and, after crossing a threshold, suppress gate node activation.
Key Feature	Stimulus features that correspond to the feature dimension that differentiates distractor from target stimuli. The type of feature that constitutes key features depends on an experiment's task demands. For example, in an experiment that requires participants to report the stream's red letter, colour would be the key feature, since it is along this dimension that the target stimulus is defined.
Key Pathway	One of two simultaneous stage 1 pathways in the 2f-ST^2 model. In the key pathway, key feature types are processed in parallel, ultimately leading to a key type being bound to a token and a response type to the same token in stage 2.
Response Pathway	One of two simultaneous stage 1 pathways in the 2f-ST^2 model. In the response pathway, response feature types are processed in parallel, ultimately leading to a response type being bound to a token and a key type to the same token in stage 2.
Task Filtering in TFL/ Type Layers	Determines whether task-demands along a given feature dimension are present. The 2f-ST^2 model's key pathway TFL/ type layer always implements task-filtering, since the key feature differentiates target from distractor stimuli based on task demands. However, in the

	response pathway, task filtering is inactivate when replicating Botella et al.'s (2001) and active when replicating Zivony and Eimer's (2020a) experimental paradigms. This is because in Botella et al.'s (2001) experiments all stimuli share the same category along the response dimension (e.g., all being letters). However, in Zivony and Eimer's (2020a) experiments, only the target and the stimulus following the target immediately (+1 intruder) do so (e.g., the target & +1 being letters and the remaining distractors being digits).
Token	Episodic context in which an item occurs in an RSVP stream. Tokens provide instance-specific information, informing that a type occurred as well as when in time relative to other items a given type occurred.
Trace (Off Node)	Trace (off) nodes of gate-trace pairs. Trace nodes are activated & excited by gate (on) nodes. After accruing excitation and crossing a threshold, trace nodes suppress gate node activation.
Type	Featural representation of an RSVP stream's item. Types include all featural properties of items. For example, a red S letter-stimulus would include a type for its colour, its shape, as well as semantic features (the item being a letter and not a digit).
τD – Random Delay	τD is a random delay added to each item in the virtual RSVP stream. It is repeatedly sampled for each respective item and follows a gaussian distribution with a mean of zero and a standard deviation of 4 time-steps (~20 ms-equivalents).
τK – Key Pathway Delay	τK is a constant delay added to key pathway processing. τK can take positive as well as negative values, the latter speeding up key pathway processing of the 2f-ST^2 model. τK remains fixed throughout a complete run of the model and most critically impacts the latency with which the blaster fires. This is because τK modulates the timing with which the task-relevant key TFL/ type layer neuron is excited by its corresponding item layer neuron.
τR – Response Pathway Delay	τR is a constant delay added to response pathway processing. τR can take positive as well as negative values, the latter speeding up response pathway processing of the 2f-ST^2 model. τR remains fixed throughout a complete run of the model. τR affects the 2f-ST^2 model's responses by affecting the probabilities of different response

	TFL/ type neurons being most active at the time of blaster firing and thus being ultimately bound to the currently active token.
vERP – Original	Original vERPs of the 2f-ST^2 model indicate the postsynaptic activation potentials, which corresponds to their presynaptic activation potentials multiplied by their respective weights.
vERP – Smoothed Difference	Smoothed difference vERPs were obtained by computing the approximate derivative of original vERPs before 1-D smoothing them with a gaussian kernel. Kernel parameters were chosen in a way that ensured comparable amplitudes between original and smoothed difference vERPs
vN2pc	The vN2pc is equal to the postsynaptic activation potential of the 2f-ST^2 model's blaster output layer.
vP3	The vP3 is equal to the postsynaptic activation potential of the 2f-ST^2 model's TFL/ type and binding units' gate projections across both pathways. All four layers contribute equally to the vP3. Compared to previous iterations of the ST^2 model (Chennu et al., 2009; Craston et al., 2008), we decided to exclude the item layers in our vP3.

Table 1. Glossary defining all important terms relating to the 2f-ST^2 computational model as well as the distractor intrusion phenomenon. Terms are presented in alphabetical order.

Results

Replicating Behavioural Results without Response Task Filtering

In this section, we will demonstrate that our current 2f-ST^2 model still replicates the main results presented with the initial 2f-ST^2 model (Chennu et al., 2011). In all analyses and model-configurations of this section, there was no task filtering in the response pathway. This reflects the fact that in these Botella et al. (2001) tasks, all response features were in principle, able to be bound to the currently available token. That is, there were no task constraints applied to the response feature. It should moreover be stressed that these results were obtained after a single run of the model, using the same rng seed as in the initial 2f-ST^2 to ensure comparability. Further, this section will repeatedly use a measure called the Average Position of Intrusions (API) to demonstrate the qualitative fit between the 2f-ST^2 model and empirical data. The API measure is computed as the sum across all intrusions' positions with respect to the target multiplied by the proportion of a given intrusion across all intrusion trials ($API = \sum_{i=-2_{i\neq 0}}^{i=+2} i * \frac{n_i}{\sum_{i=-2_{i\neq 0}}^{i=+2} n_i}$). Thus, it is the expectation with respect to the probability distribution

of responses across intrusion positions. Importantly, the API measure weighs +-2 intrusions more heavily than +-1 intrusions. Botella et al. (2001) argued that this property more accurately reflects the position of intrusions, as -2 intrusions are more pre-target than -1 intrusions, for example. Thus, the API is a metric that quantifies the loci of response distributions and thereby simultaneously informs us about shifts thereof that may either result from empirical manipulations in human experiments or different values of τK and τR in the 2f-ST^2 model.

The 2f-ST^2 model replicated the response distributions of Experiments 1A and 2 of Botella et al. (2001), presented in Figure 21, Panels A and B, respectively. As presented previously in Figure 15 & Figure 16, Botella et al. (2001) varied word frequency in these experiments to manipulate key and response feature processing speeds in Experiments 1A and 2, respectively. Key feature processing speed was manipulated (Experiment 1A) using streams of coloured word stimuli and asking participants to report the colour of the animal word. In this case, high frequency (HF) animal words (e.g., dog) corresponded to a more salient key feature than low frequency (LF) ones (e.g., iguana). Response feature processing speed was manipulated similarly, now asking participants to report the only uppercase word in a stream of grey word stimuli. HF words (e.g., slow) now corresponded to more salient response features than LF ones (e.g., prate). Note that Figure 21 was adopted from Figures 2 & 3 of Chennu et al. (2011) and updated to reflect the current 2f-ST^2 model's response distributions. The 2f-ST^2 model replicates the pre-target shift of intrusion responses going from LF to HF conditions shown by Botella et al.'s (2001) Experiment 1A. This pre-target shift was accompanied with empirical API values of -.306 (HF) and .043 (LF) (Botella et al., 2001). Keeping τR fixed at 0 time-steps, we simulated the HF and LF conditions of this experiment using τK values of 0 and 8 (~40 ms) time-steps. For this pair of delay-configurations, the initial 2f-ST^2 model generated API values of -0.11 and 0.78 for the HF and LF conditions (Chennu et al., 2011), respectively. The current version of the 2f-ST^2 model replicated this pattern, yielding API values of -0.22 (HF) and 0.61 (LF). Experiment 2 of Botella et al. (2001) generated API values of .131 and -.034 for the HF and LF conditions, respectively, meaning a post-target intrusion response shift going from LF to HF conditions. This experiment was replicated with the 2f-ST^2 model keeping τK fixed at 0 and simulating the response-feature salience manipulation using τRs of 0 and 2 time-steps for the HF and LF conditions. This led to respective API values of -0.11 and -0.31 with the initial 2f-ST^2 model (Chennu et al., 2011) and of -0.22 and -0.39 with the current 2f-ST^2. The 2f-ST^2 model thus

replicates the post-target intrusion response shift going from LF to HF conditions observed by Botella et al. (2001) in Experiment 2.

These results therefore indicate that a more salient key feature is accompanied with a response shift towards more pre-target intrusions, whereas more salient response features yield post-target intrusion responses more frequently. While Botella et al. (2001) argued for the validity of his dual-route computational model based on these findings, for us these findings provide initial evidence for Hypothesis 1 of this thesis, i.e., that binding multi-dimensional stimulus features into single percepts depends on the timing of transient attentional enhancement (TAE).

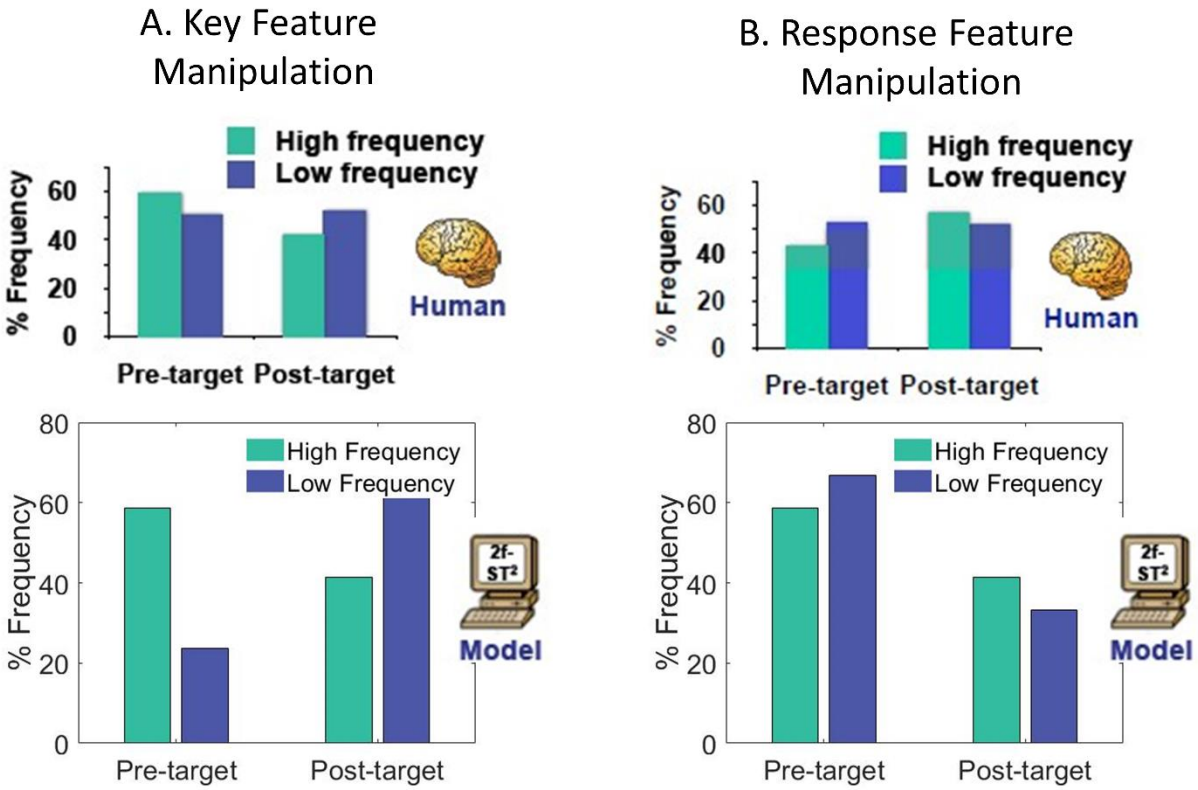


Figure 21. The 2f-ST^2 model replicates the response distributions found in Botella et al.'s (2001) Experiments 1A and 2. Botella et al. (2001) modulated word frequency to obtain an isolated modulation of key feature processing speeds in Experiment 1A (Panel A), which was accompanied with an increase in pre-target intrusion responses for more salient key features. The 2f-ST^2 model replicates this pattern (Panel A). Botella et al. (2001) modulated word frequency to obtain an isolated modulation of response feature processing speeds in Experiment 2 (Panel B). This was accompanied with an increase in post-target intrusion responses for more salient response features. Note that we adopted this figure from Chennu et al. (2011) and updated the two bottom panels to reflect the current version of the 2f-ST^2 model.

Finally, the Fortunate Conjunction Experiment of Chennu et al. (2011) manipulated key feature processing speed using letter and symbol targets. We present the response distributions shown in humans empirically by Chennu et al. (2011) and generated by the 2f-ST^2 model in Figure 22's Panels A & B, respectively. Note that Figure 22 is adopted from Chennu et al. (2011) with Panel B being updated to reflect the current 2f-ST^2 model's

behaviour. Chennu et al.'s (2011) experiment generated API values of 0.18 and 0.31 for letter and symbol targets, respectively. These response conditions were replicated with the early (EK) and late key (LK) conditions in the initial 2f-ST^2 model, which implemented τKs of -8 and 0 time-steps to replicate each respective condition, whilst τR was kept fixed in both conditions at 0. This led to API values of -0.86 and -0.11 in the initial 2f-ST^2 (Chennu et al., 2011) for EK and LK conditions. The corresponding values in the current 2f-ST^2 were -0.9 and -0.22.

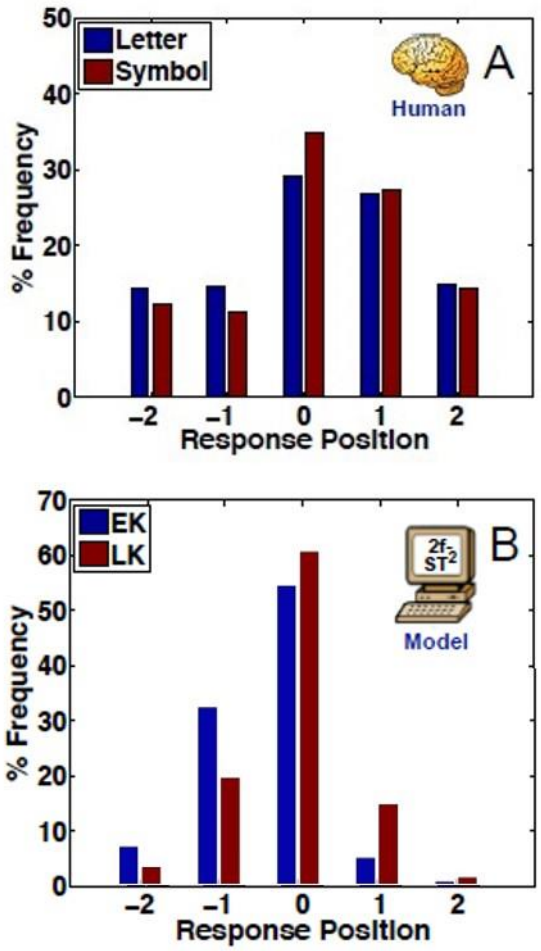


Figure 22. Chennu et al.'s (2011) Fortunate Conjunction Experiment (Panel A) replicated with the 2f-ST^2 model (Panel B). Letter and symbol experimental conditions were replicated with *τ*Ks of 0 & 8 time-steps (corresponding to early key (EK) & late key (LK)), respectively. Adopted from Chennu et al. (2011) and updated to reflect the current 2f-ST^2 model's behaviour in Panel B.

Overall, these results demonstrate that even though we adjusted the initial 2f-ST^2 model in a few architectural ways, the general pattern of results remained the same and the model still successfully replicates Experiments 1A and 2 of Botella et al. (2001), as well as the Fortunate Conjunction Experiment conducted by Chennu et al. (2011). These results can therefore be seen as evidence for Hypothesis 2a, i.e., that the 2f-ST^2 model generates

behavioural response distributions matching those obtained with various distractor intrusion experiments.

Counterintuitive Reaction Time (RT) Pattern of Botella (1992)

One very central finding presented with the initial 2f-ST^2 (Chennu et al., 2011) is worth further elaboration, since we will present novel vERP patterns related to this finding next. The key finding shown by Chennu et al. (2011) was that the initial 2f-ST^2 was able to replicate a rather counterintuitive pattern of Reaction Times (RT), which was shown by Botella (1992). In their study (Botella, 1992), correct reports were responded to the fastest, followed by pre-target errors and then post-target errors. This finding is striking and counterintuitive, as one would expect pre-target errors to be made quicker than correct reports, because the item that was ultimately reported by the participant was presented *earlier in the RSVP sequence than the correct item was.* Thus, the results of Botella (1992) suggested that the pattern of RTs in distractor intrusion experiments does not follow the sequence of item presentation in the corresponding RSVP streams. Chennu et al. (2011) defined RTs in the initial 2f-ST^2 as the point in time at which the currently available token completes binding to the key as well as response neuron. An index of this event is provided by the currently active token's trace neuron (Bowman & Wyble, 2007). Specifically, the model provides its response in a given trial if the token's trace neuron reaches 75% of its maximum amplitude. This event was thus used to define reaction times in the 2f-ST^2 model.

We acknowledge that this definition of reaction times in the 2f-ST^2 model approximates processes and aspects leading to human reaction times. Human reaction times in response to cognitive tasks have been extensively studied by the field of mental chronometry (Medina et al., 2015). Neuroscientific studies have demonstrated that a response to a stimulus results from a processing hierarchy in the brain in which WM-encoding follows initial low-level perceptual processes. Importantly, this is followed by higher-level cognitive processes in prefrontal cortex exciting neurons in the motor cortex which in turn initiate movement of the body parts that are required for a given response (Posner, 2005). In the example of Botella's (1992) experiment, participants had to move their finger to provide a response. Therefore, WM-encoding of the perceived stimulus in a given stream was naturally not sufficient for the response to occur. In addition, the neuronal population in motor cortices controlling movement of that finger were excited and activation propagated to neuromuscular junctions, ultimately leading to chemical processes in the finger's muscle cells which made it contract. It is clear that in the 2f-ST^2 model the processes required for a response that follow WM-

encoding were not considered. Our computational model is therefore limited in this sense, i.e., not considering the motor processes and variability (e.g., interindividual or temporal) thereof when providing virtual reaction times. Figure 23 demonstrates that the current version of the 2f-ST^2 model nonetheless replicates the counterintuitive pattern of RTs shown by Botella (1992). We acknowledge the quantitative differences between the RT values obtained by Botella (1992) and those generated by 2f-ST^2, with our model yielding responses that are over 100 ms-equivalents later than the human data. However, we stress that we kept the parameter space of the original ST^2 model, which was recognized to have delayed responses compared to human RSVP data in general (Bowman et al., 2008; Bowman & Wyble, 2007). We further stress that throughout this work, we focus on obtaining a qualitative fit with the 2f-ST^2 model to human behavioural as well as neuroimaging patterns.

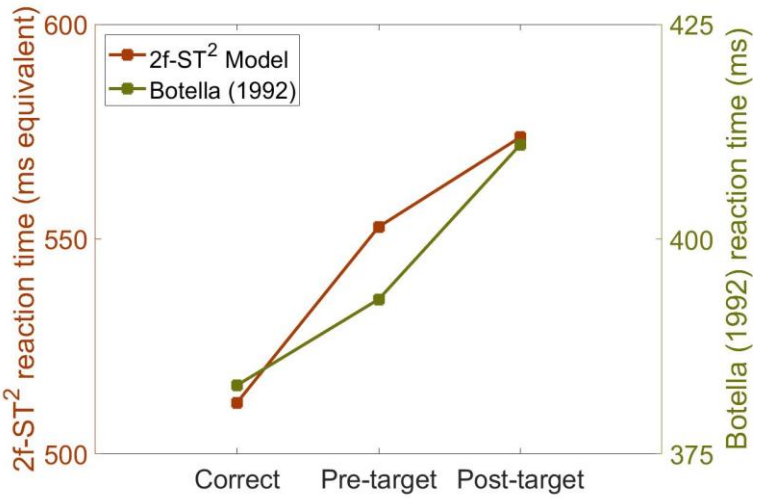


Figure 23. Reaction Time Patterns of Botella (1992) replicated by 2f-ST^2. Note that y-axes are different.

2f-ST^2's virtual ERPs explaining RT patterns

The counterintuitive pattern of reaction times presented in Figure 23 can be understood well by looking at our model's virtual ERP components. Figure 24 shows the virtual N2pc and P3 components, which were computed as described in the *Virtual Event-Related Components (vERPs) of the 2f-ST^2 model* paragraph. Green, black, and red traces correspond to pre-target intrusion, correct response, and post-target intrusion trials, respectively. Vertical coloured lines in all vERP plots indicate the average latency of blaster-firing for each respective response condition, with their exact numerical values being provided in plots' legends. Blaster-firing latencies were defined as the point in time at which 50% of the area under the vN2pc (measuring the blaster's output) curve was reached. The vP3 pattern should be especially considered when thinking about reaction times (RT), as we model

RTs as the time-point in which the binding of a type to a token completes in the binding pool, a process that is generated by the layers that constitute the vP3. As displayed in Figure 24B, correct trials show an earlier vP3 than pre- as well as post-target intrusion trials, which qualitatively fits the reaction time patterns well. This can be explained by the following.

In correct trials, the blaster fires close in time to the activation peak of a strongly active target response TFL/ type neuron. The target response TFL/ type neuron consequently wins the competition with distractor response TFL/ type neurons, which exists due to the lateral inhibition present in this layer, quickly. However, in intrusion trials, the target's key and response features are generally weaker. The target's key feature being weaker means a later firing of the blaster, which is shown as later intrusion vN2pcs in Figure 24A, since the vN2pc is measured as the blaster's output. Pre- and post-target intrusions occur if respective response TFL/ type neurons are stronger than the target's response TFL/ type neuron. Importantly, in these cases the blaster fires temporally further from the activation peak of the subsequently bound distractor type's response TFL/ type neuron. This response TFL/ type neuron therefore requires more time to overcome the lateral inhibition of this layer and win the competition to the binding pool. The additional time this process requires in intrusion trials leads to delayed vP3s (which reflect activation levels of the response TFL/ type layer & the response binding pool) as well as increased mean RTs.

We call this phenomenon a *loss of responsiveness* of these weak types, which we will elaborate upon later, as this same phenomenon plays a central role in the vERP dynamics after implementing the task-filter in the response pathway to replicate Zivony and Eimer's (2020a) findings.

Further, the qualitative fit between Botella's (1992) and the 2f-ST^2's RT data as well as the vERP-based explanations thereof provide further evidence in favour of Hypothesis 1, i.e., that TAE timing determines the binding of multi-dimensional stimulus features into single percepts, as well as Hypothesis 2, i.e., that the 2f-ST^2 accounts for a broad range of findings obtained with distractor intrusion experiments.

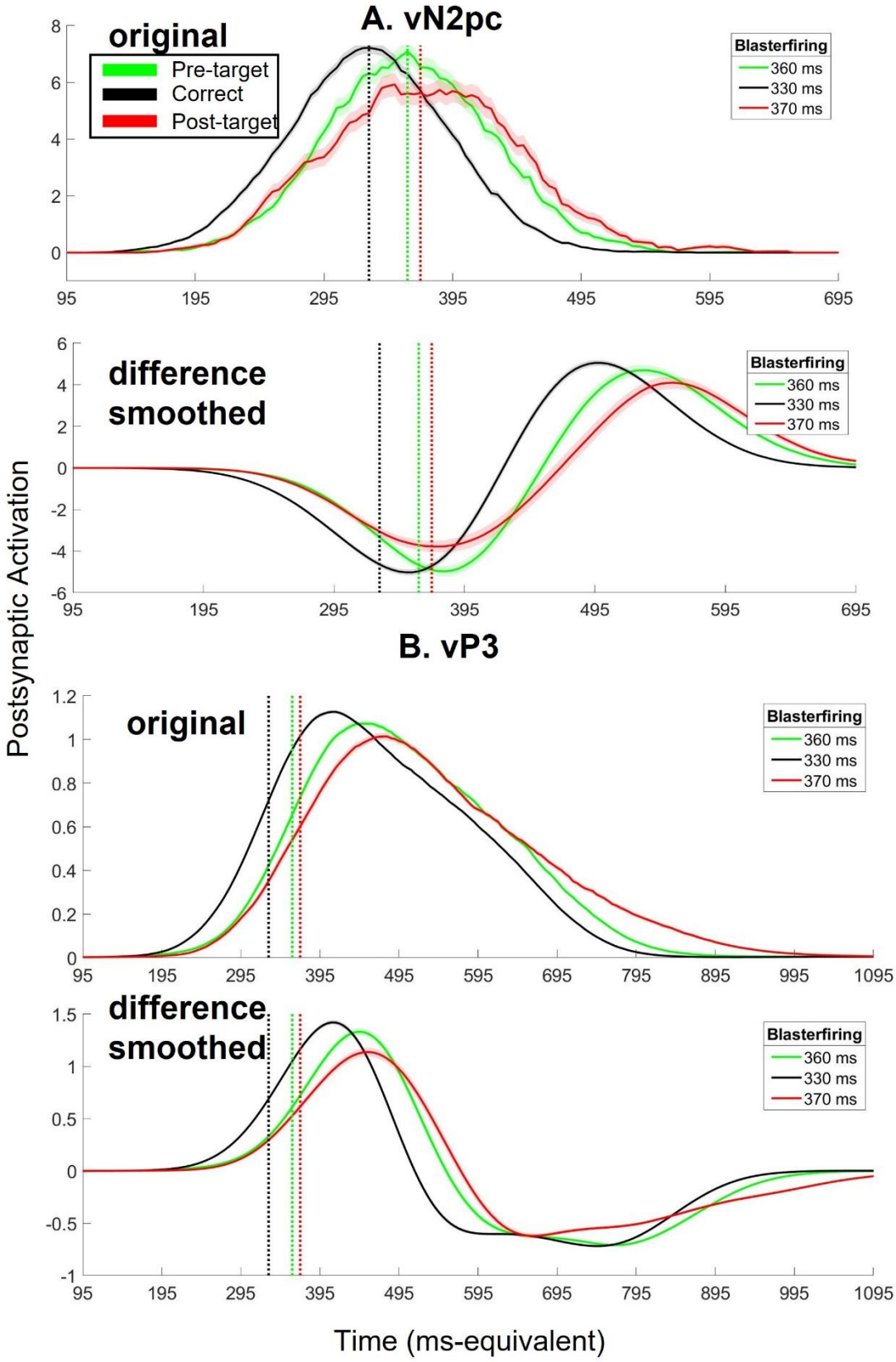


Figure 24. 2f-ST^{2}'s vERPs without response pathway task-filtering and no processing delays to either pathway explaining RT findings of Figure 23. Green, black, and red traces correspond to pre-target, correct, and post-target response conditions, respectively. Note that intrusion conditions collapse across +/- 1/2 intrusions, hence, the green trace corresponds to -2 as well as -1 intrusion trials, for example. We present the vN2pc in Panel A and the vP3 in Panel B, plotting the original (model-output) vERPs as well as the difference smoothed vERPs. Vertical lines indicate average blaster-firing (with their numerical values being shown in the legend on the right).

Zivony & Eimer's (2020a) Experiments 1A & 1B Human Event Related Potentials (ERPs)

In Figure 25, we present Grand Average waves of all 23 single-subject ERPs of Zivony and Eimer's (2020a) Experiments 1A and 1B. We have introduced these experiments in the introduction of this chapter (Figure 17). To sum, in both experiments (1B being a direct replication of 1A), dual RSVP streams were presented, and participants were asked to report the digit target that was surrounded by an annulus. In streams of distractor letters, Zivony and Eimer (2020a) only presented either one or two digit stimuli (either the target or the target as well as the immediately following digit (+1 intruder)). This experimental design thus only allows for the possibility of +1 intruders, since only the distractor following the target immediately shared the target's response dimension (i.e., both being digits).

In both ERP components of this experiment (note that we combined Experiments 1A & 1B), the N2pc (Figure 25A) as well as the P3 (Figure 25B) showed latency differences, with intrusion trials showing later ERPs than correct trials. Furthermore, the N2pc (Figure 25A) has a higher amplitude after correct trials (more negative for a negative going effect), which was already noted by Zivony & Eimer (2020a). Peak amplitudes of P3 components are comparable but occur earlier after correct trials (Figure 25B). In the next sections, we will test these latency differences for statistical significance using dynamic time warping, test whether the N2pc and P3 are temporally correlated and finally replicate them with our model's virtual counterparts.

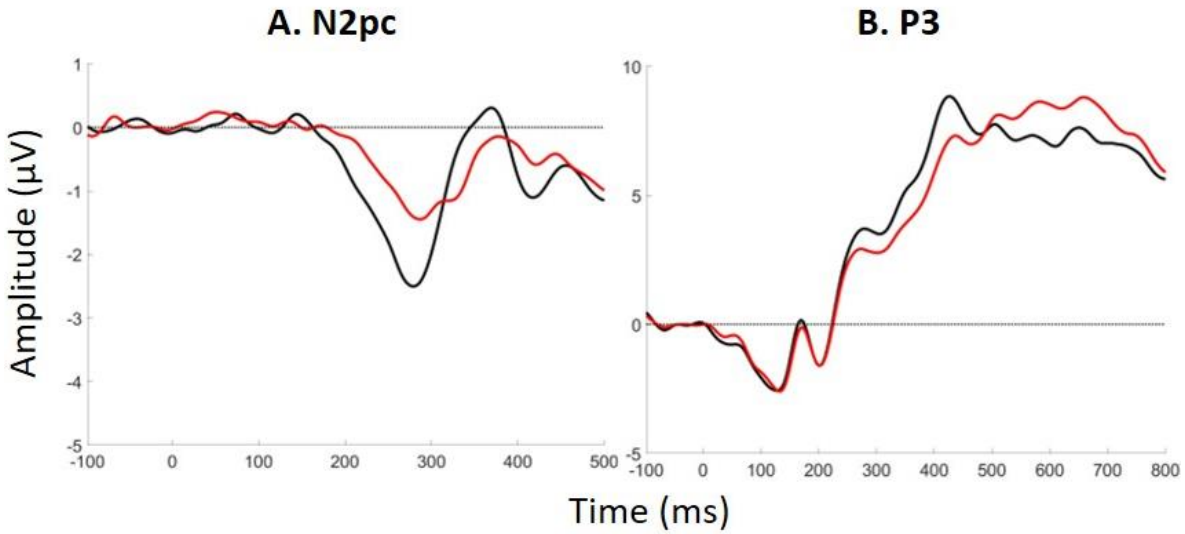


Figure 25. Human ERP data of Zivony & Eimer's (2020) experiment 1. Black and red lines indicate ERPs of correct and intrusion conditions, respectively. We combined the dataset of the authors' (Zivony & Eimer, 2020a) Experiments 1A and 1B, as 1B was a direct replication of 1A.

Dynamic Time Warping (DTW) Latency Difference Analysis of Zivony and Eimer's (2020a) Paradigm

Replicating the N2pc latency differences

As it is robust against high frequency noise, which particularly effects measures of latency focussed on individual points, we used dynamic time warping (DTW) to replicate Zivony and Eimer's (2020a) N2pc latency differences between correct and intrusion responses (Figure 26 & Figure 27). We furthermore used the same approach to examine the same latency contrast for the P3 component (measured at Pz, Figure 28 & Figure 29). Figure 26 shows the DTW warping path that was found by the algorithm to ensure optimal alignment (i.e., minimal Euclidian distance). We present the DTW reference signal, the N2pc of correct trials, in black on the y-axis, and the query signal, the N2pc of intrusion trials, in red on the x-axis. We computed the latency difference in milliseconds based on the median of the warping path's distance-distribution between x- & y-coordinates, which for the N2pc was 18 ms. This is in line with Zivony and Eimer's (2020a) 50% average peak amplitude criterion which yielded latency differences of 30 and 20 ms in Experiments 1A and 1B, respectively (note we combined these two experiments into our analysis, as the original Experiment 1B was a direct replication of 1A). It should be noted that the present latency difference of 18 ms is also closely in line with other work by these authors (Zivony & Eimer, 2020b), where intrusion trials implied an N2pc component that was 19 ms later than correct trials. The permutation distribution of DTW area-differences under the null is shown in Figure 27. Our two-tailed permutation test supported our hypothesis that intrusion trials had a later N2pc component than correct trials (p = .0013).

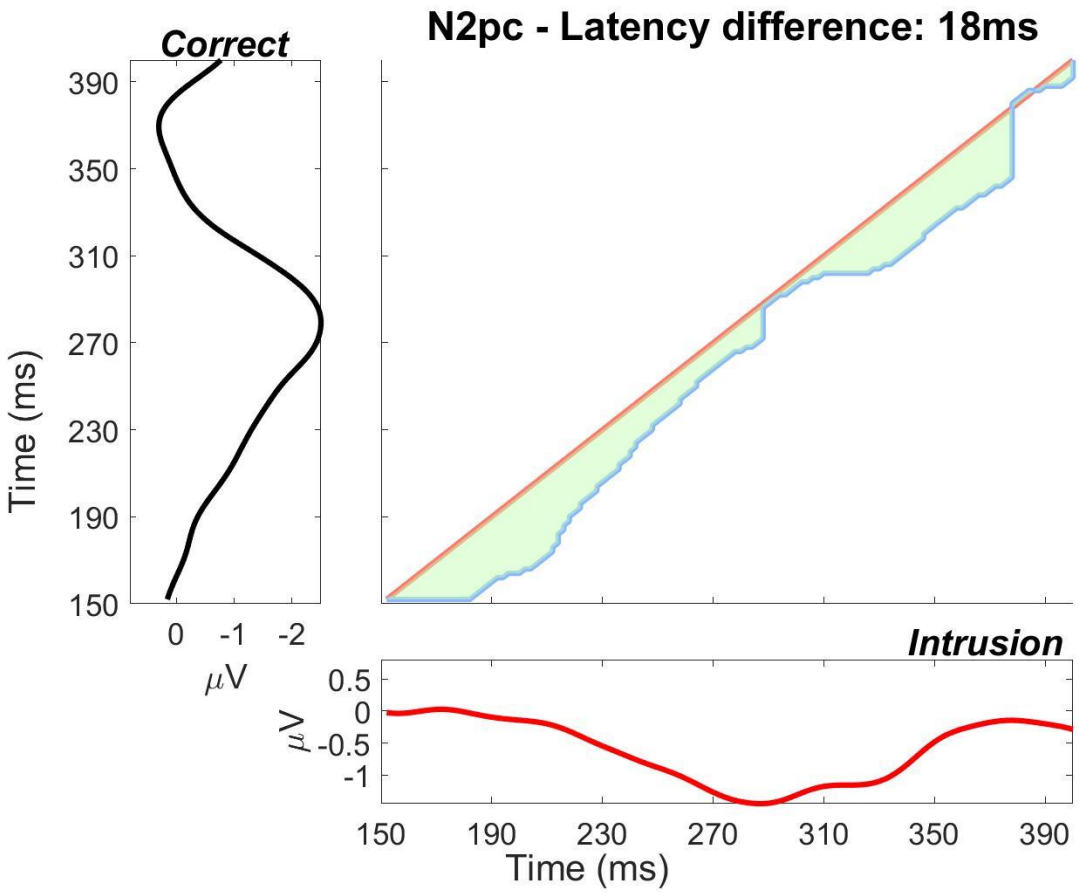


Figure 26. N2pc DTW warping path after optimal alignment as well as original correct (black) and intrusion (red) N2pcs. DTW analysis yielded a latency difference of 18 ms between correct and intrusion conditions. Note that because the warping path (blue) is located under the main diagonal (light-red), the reference time-series (correct) preceded the query (reference). These results are overall in line with those shown by Zivony and Eimer (2020a).

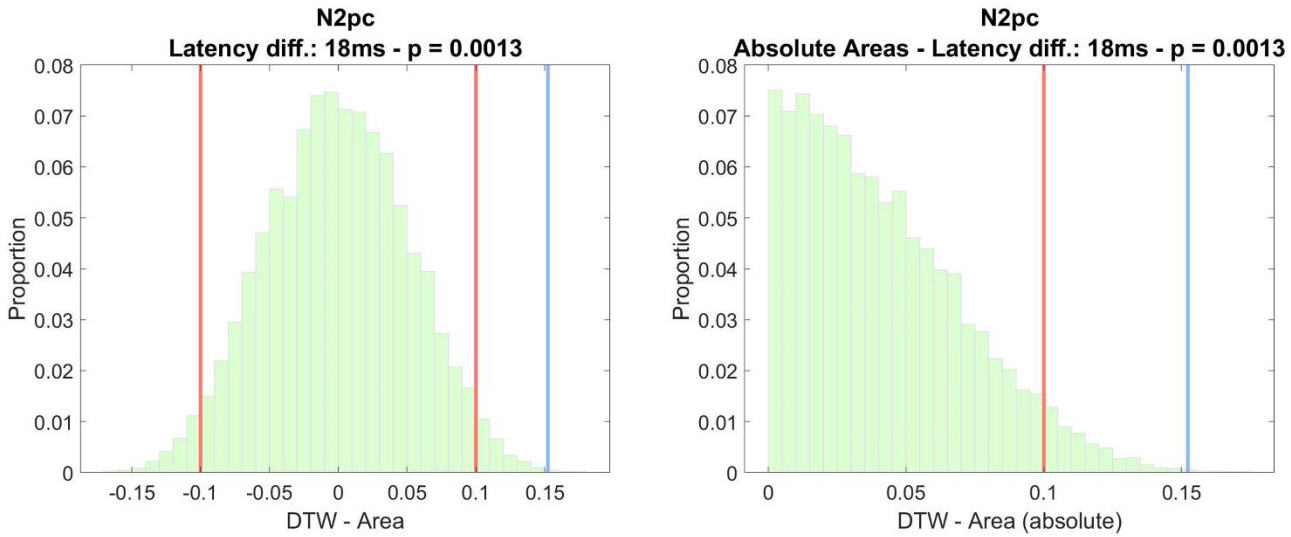


Figure 27. Distributions of permuted (null) DTW latency differences between correct and intrusion trials' N2pcs (left) and their absolute values (right) used for two-tailed significance testing. The red vertical line indicates significance threshold of our two-tailed test with an Alpha of 5%. The blue vertical line shows our observed DTW latency difference.

P3 latency differences at Pz

We further analysed latency differences between correct and intrusion trials using DTW for Zivony & Eimer's (2020a) P3 component at the Pz electrode (Figure 28 & Figure 29). We present the DTW warping path and original ERP components in Figure 28, and the permutation distribution of latency-differences in Figure 29. Again, intrusion trials showed a later P3 component than correct trials, with the latency difference being 73 ms, which was highly statistically significant after running our two-tailed permutation test (p = .0003). Therefore, we replicate and extend the results of Zivony and Eimer (2020a), suggesting that not only the deployment of selective attention is delayed in intrusion trials (N2pc being 18 ms later), but that this latency-difference is also evident for the P3 component (being 73 ms later after intrusions), suggesting later WM-encoding for intrusions.

If we put the findings of this section with those of the previous section, we obtain evidence for a key prediction of ST^2 models. As previously discussed, ST^2 asserts that transient attentional enhancement (TAE) and encoding into working memory are coupled. That is, when the blaster fires late, encoding will be late. This suggests a temporal coupling between the N2pc and the P3. This is exactly what we see here: the N2pc and the P3 are *both* earlier in correct trials. That is, when the blaster fires earlier (N2pc earlier), encoding is also earlier (P3 earlier).

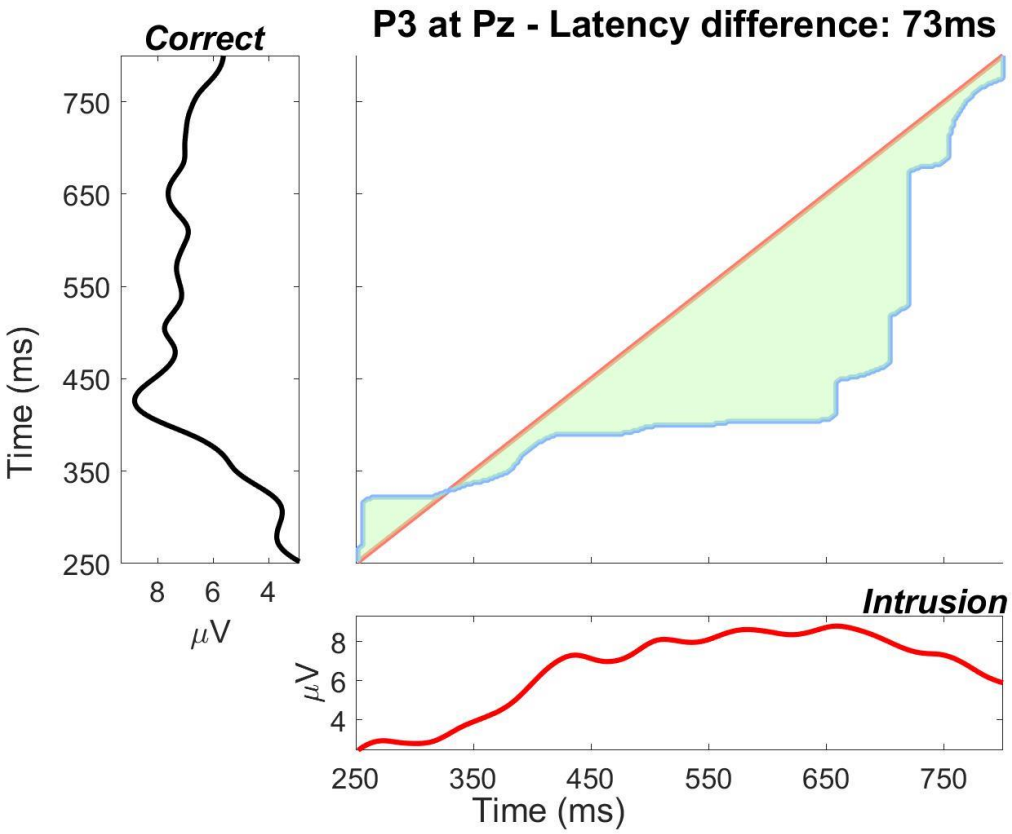


Figure 28. P3 DTW warping path after optimal alignment as well as original correct (black) and intrusion (red) P3. The P3 component was measured at the Pz electrode. Plotting conventions are identical to those of Figure 26. These DTW results again indicate that the intrusion condition's P3 was later than that of correct trials.

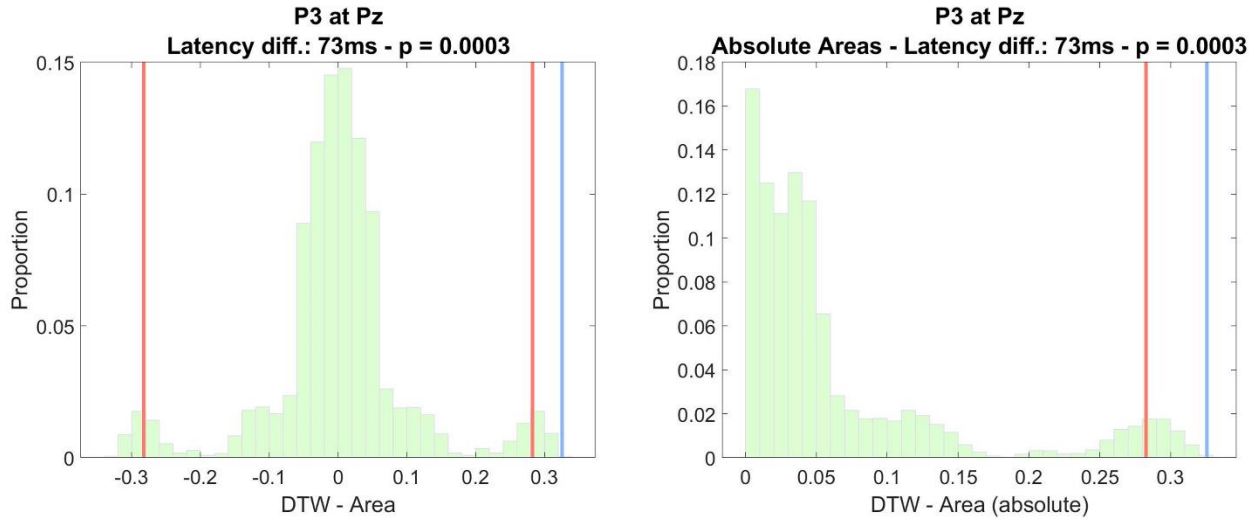


Figure 29. Distributions of permuted (null) DTW area differences between correct and intrusion trials' P3 components (left) and their absolute values (right) used for two-tailed significance testing. The red vertical line indicates significant threshold of our two-tailed test with an Alpha of 5%. The blue vertical line shows our observed DTW area difference.

Correlation between human N2pc & P3 latencies

In the previous section, we provided evidence that human N2pc and P3 components are temporally correlated, i.e., later P3 components follow comparably later N2pcs. That was based upon differences across experimental conditions, i.e., Correct vs Intrusion. However, this does not definitively ensure that this coupling obtains beyond experimental conditions, i.e., due to the intrinsic variability in latencies that "spontaneously" arise. This section responds to this aspect by showing that N2pc and P3 latencies are coupled even within conditions. Figure 30 presents the results of the bootstrap analysis we conducted to probe this hypothesis, which importantly, was applied to correct and intrusion conditions separately. Figure 30 displays the scatterplots of DTW area-difference-pairs with the line of best linear fit as well as two marginal distributions per scatterplot. Both analyses yielded a positive and statistically significant correlation between N2pc & P3 DTW-latency after using the same bootstrap samples for the two components in each of the 10000 bootstrap repetitions. The correlation values after Pearson were $r = .34$, for correct and $r = .16$, for intrusion trials. We provide the rank correlation after Spearman as well to account for the possibility of non-linear relationships between the DTW-latency values that were correlated (indeed, we do observe some loss of normality in marginal distributions, see Kurtosis and Skewness measures, suggesting heteroskedasticity). Spearman correlations were $r = .40$ and $r = .14$ for correct and intrusion trials, respectively. All four correlations had p-values smaller than .0001 and were hence highly significant. However, we emphasize that p-values obtained in resampling analyses of the kind shown here, should be interpreted with care. This is because the degrees of freedom are determined by the programmer (9998 in this case) and as discussed in Friston (2012), the fallacy of classical inference states that once the sample size is sufficiently large, p-values become trivial, as the smallest effects suffice for significance. Crucially, this does not mean that analyses with many degrees of freedom are inherently flawed, but that one should focus on measures of standardized effect sizes, such as correlation coefficients, when interpreting their results (Lorca-Puls et al., 2018). These positive correlations provide support for our model's architecture and specifically for the hypothesis that attentional deployment and working-memory encoding (or in model terms, blaster-firing and binding types to tokens in stage 2) in humans is correlated. In particular, although these correlations are small, they reflect a clear hypothesis that arises from our model and is important for verifying it.

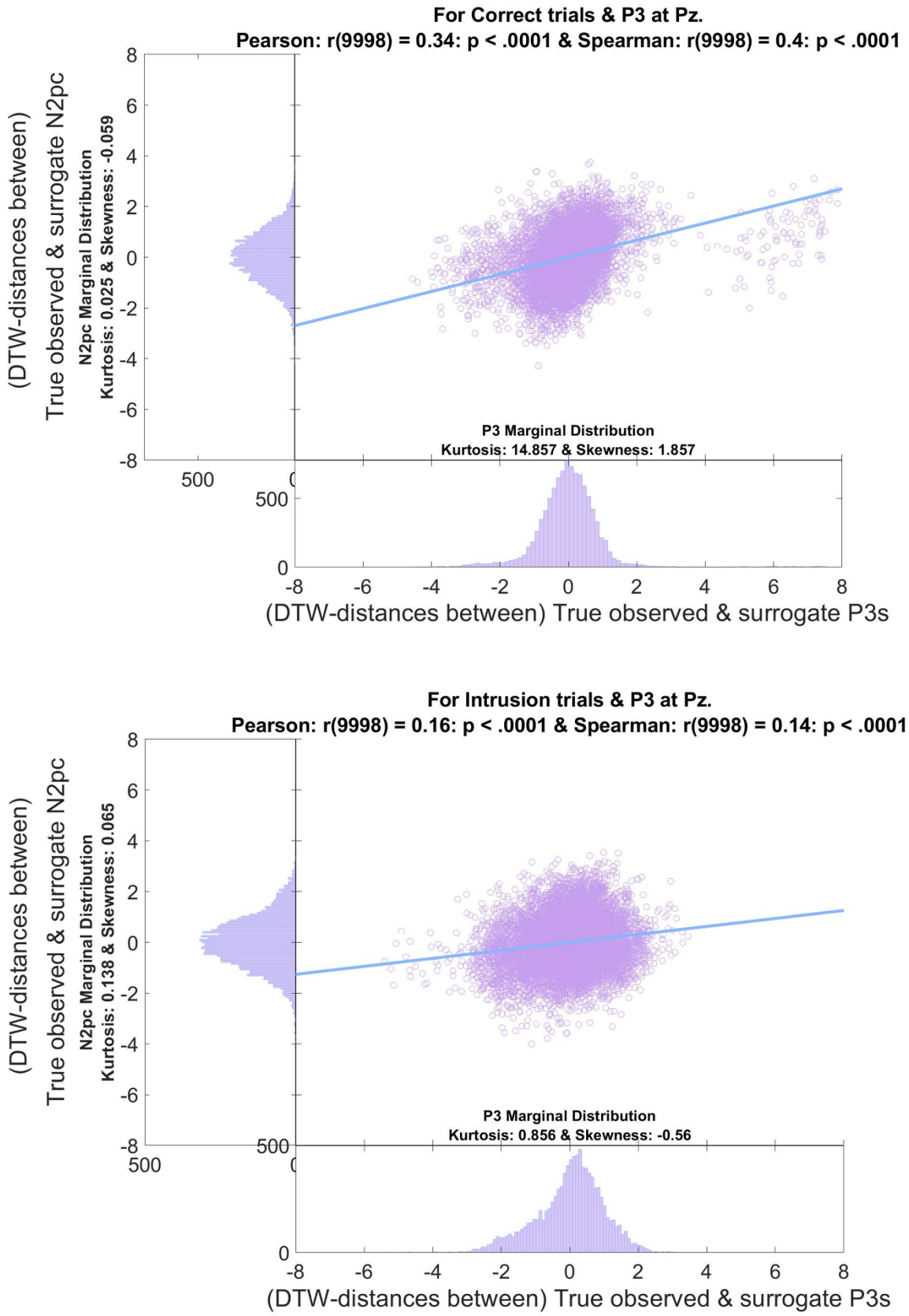


Figure 30. Bootstrap analysis of correlated N2pc and P3 latencies. Top and bottom panels show scatterplots of bootstrapped pairs of DTW area-differences and the line of best linear fit for correct and intrusion trials, respectively. We furthermore present the marginal distributions of true-observed & surrogate ERP-components on their respective axes in both panels.

It is also striking that the correlation between N2pc and P3 is considerably stronger for correct trials then for intrusion trials. Intuitively this makes sense, suggesting that attentional enhancement (N2pc) and working memory encoding (P3) are more strongly coupled when the system is performing accurately, with attentional enhancement very directly opening the gate to conscious perception and encoding. In contrast, intrusion errors may be marked by weaker (as confirmed by the reduced amplitude N2pc for intrusion trials, see Figure 25) and perhaps more temporally variable attentional enhancement that does not so directly initiate conscious perception/encoding, perhaps indicating that conscious perception/ encoding is more impacted by noise than volitional attention when intrusion errors are generated. This resonates with the findings in Chennu et al. (2009) that temporal variability increases when attention is not available.

This finding may also be informative with regards to the debate over models. In particular, Botella et al.'s (2001) model postulated two pathways, one that involved accurate direction of attention and generated correct responses, and the other that was impacted by noise and generated intrusion errors (Botella et al., 2001). Our finding that attentional enhancement is more strongly coupled with conscious perception/ encoding during correct responses resonates with this idea. Although, we do not believe that the finding contradicts the notion that correct responses are just fortunate conjunctions. Indeed, it would be consistent with the idea that "good fortune" could be indexed by the extent to which transient attention and conscious perception/ encoding are coupled, but where this "extent" is a continuum, rather than dichotomous, as suggested by Botella et al.'s (2001) model.

In summary, the results of the current and previous sections further support Hypothesis 1 of this thesis, i.e., that the binding of multi-dimensional stimulus features into single percepts depends on the timing of TAE.

Replicating Zivony & Eimer's (2020a) Behavioural & Electrophysiological Results: the 24 time-step (120 ms-equivalents) Key Delay

Importantly, in all analyses and model-configurations of this section, the response filter at the TFL/ type layer was active, meaning that only the target response feature (the 0-neuron) and the intrusion stimulus directly following it (the +1-neuron) were able to be bound to the currently available token. All other response neurons were fully suppressed by the task filter. This is consistent with the change from the classic Botella (Botella et al., 2001) tasks, in which there is no (target-nontarget) category change in the response feature, and the Zivony & Eimer (2020a) task, in which there is. We furthermore added a 24 time-steps (120 ms-

equivalent) delay to processing of the key pathway in these analyses. This delay simulates the fact that the key features used in Zivony and Eimer (2020a) (a square or an annulus surrounding the target) is less salient and hence detected less quickly by the attentional system than a colour-change, which was the key feature implemented in Botella et al.'s (2001) paradigms. Additionally, Botella et al. (2001) style colour-cuing directly marks the response stimulus, rather than surrounding it, potentially avoiding a movement of attention from the key-feature cue to the response-feature item. Experiment 2, implementing colour as the key feature, of Zivony and Eimer (2020a) can be seen as a demonstration of this, as more correct responses as well as a sharper N2pc were observed compared to Experiment 1. This interaction of a more salient key feature leading to a shift in the response distributions as well as different N2pcs will be elaborated upon further as well as simulated by the 2f-ST^2 directly in Chapter 6.

Accuracy

We replicated Zivony and Eimer's (2020a) response accuracy pattern closely (Figure 31) with 2f-ST^2, yielding a distribution of 46% correct and 54% intrusion responses when implementing a 120 ms-equivalent delay in key pathway processing. This compares well to Zivony and Eimer's (2020a) findings of 34.95% correct, 57.65% intrusion and 7.4% wrong responses (these values are the average accuracy values across those of the original study's Experiments 1A (36.1% correct, 57% intrusion, 6.9% misses) and its direct replication 1B (33.8% correct, 58.3% intrusion, 7.9% misses (Zivony & Eimer, 2020a)). The apparent discrepancy in the proportion of correct trials between our model (46%) and human data (34.95%) is because, for reasons of consistency throughout this thesis and with previous 2f-ST^2 work (Chennu et al., 2011), we excluded trials in which wrong responses (misses) were given by our model.

The qualitative fit between Zivony and Eimer's (2020a) and the 2f-ST^2's response distributions (Figure 31) provide further evidence in favour of Hypothesis 2a, i.e., that the 2f-ST^2 model generates behavioural response distributions matching those obtained with various distractor intrusion experiments.

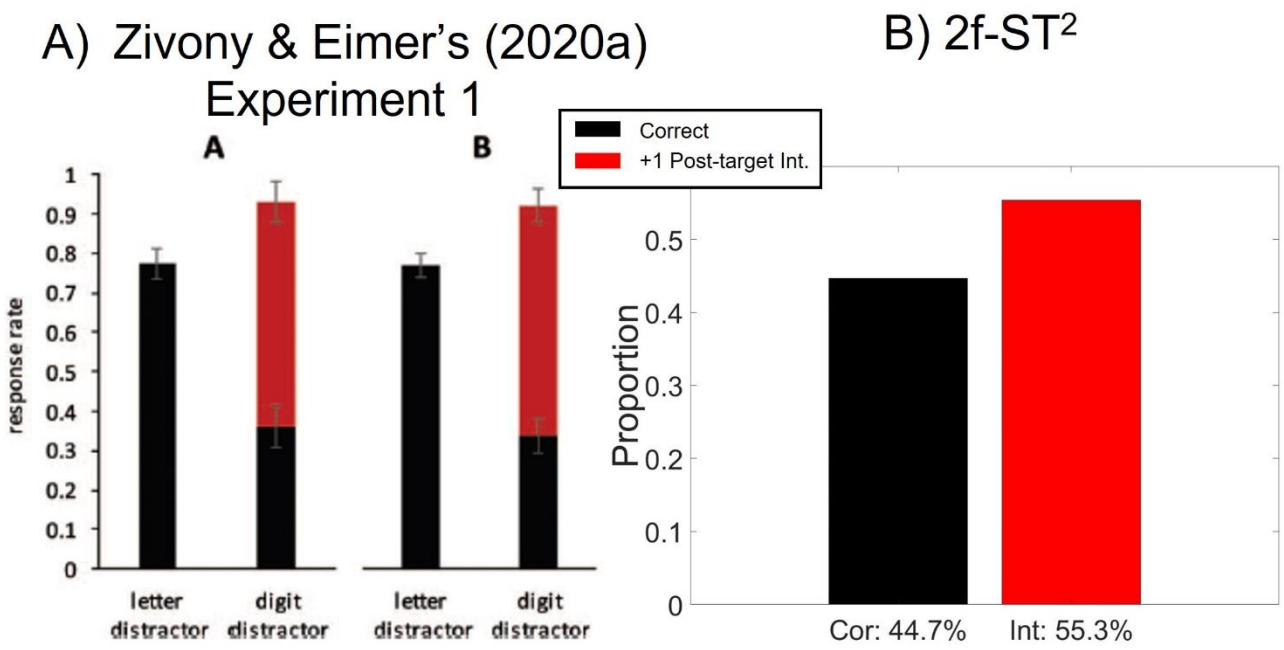


Figure 31. Response distributions of Zivony & Eimer's (2020) experiment 1 (A: Original Study, B: Replication) (left) and 2f-ST^2 (right) with task-filtering in the response pathway and a τK of 24 time-steps (120 ms-equivalents). Black and red bars correspond to correct and intrusion trials, respectively. Letter distractor in Panel A correspond to trials in which a distractor letter was presented immediately after the target digit and, thus, no intrusion errors were possible. Digit distractor refers to trials in which a digit distractor was presented immediately after the target digit, hence allowing for intrusion errors to occur.

vERPs

Figure 32 displays the virtual N2pc (Figure 32A) and P3 (Figure 32B) components of the 2f-ST^2 model with an active task-filter in the response TFL/ type layer and a key-pathway processing delay of 24 time-steps (120 ms-equivalent). Vertical lines in Figure 32 display average blaster-firing latencies of each response condition (computed as the latency of 50% area under the blaster output (i.e., vN2pc) curve). Note that the time-ranges differ between the vN2pc and vP3 components. Both virtual components replicate the pattern of time-delay shown in human ERPs (Figure 25). They further show a pattern of decreased amplitude for intrusion trials, which fits the human N2pcs (Figure 25A) especially well. Additionally, as argued in depth elsewhere (Chennu et al., 2009; Craston et al., 2008), vERPs should be seen as coarse approximations of human ERPs, as major physical properties of the human brain, skull and scalp are not considered. Therefore, one should only expect to obtain a qualitative match of vERPs to human ERPs. This qualitative fit supports Hypotheses 2b & 2c of this thesis, i.e., that the 2f-ST^2 model generates vERPs that replicate human ERPs obtained in distractor intrusion experiments.

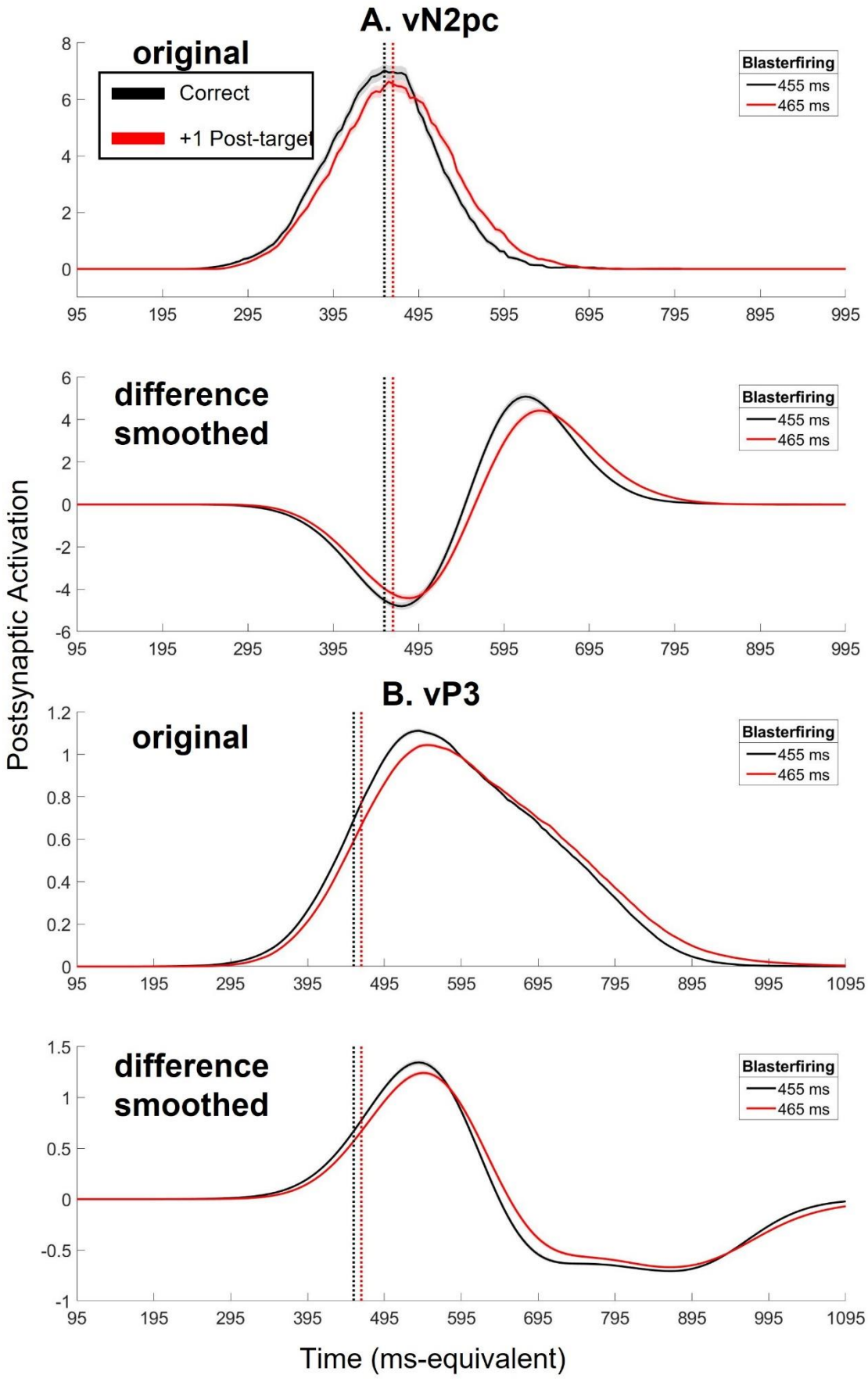


Figure 32. 2f-ST^{2}'s vN2pc (Panel A) and vP3 (Panel B) obtained after active task-filtering in the response TFL and a τK of 24 time-steps (120 ms-equivalents). Black and red traces correspond to vERPs of correct and intrusion trials, respectively. For both vERPs, we present the original (model-output) as well as the difference smoothed version. Vertical lines indicate average blaster firing latencies (with their numerical values being shown in the legend on the right).

Model Predictions

Importantly, since the 2f-ST^2 model encompasses the full stimulus-response range, we can investigate the consequences of manipulations at different stages in this full range. In particular, the simulations we have presented of Botella et al. (2001) and Zivony and Eimer's (2020a) empirical findings have centred on manipulating two levels of this range: 1) exogeneous stimulus-salience, by varying the key feature (annulus versus colour-marking) and 2) response competition, by varying the number of response feature items that could be reported as targets, through a category manipulation. The first of these is simulated as the presence or absence of an extra key feature delay to blaster firing, and the second as the presence or absence of a task filter in the response pathway, which constrains the extent of competition between response items due to lateral inhibition. To explore this, we present two configurations of the 2f-ST^2 model, for which equivalent human data is not available. Accordingly, these results make predictions to be tested by future research.

Inactive Response Task Filter and Key Delay Configuration

The first configuration lacking equivalent human data implements an inactive response task filter reflecting high response competition, like the analyses presented in the *The 2f-ST^2's virtual ERPs explaining RT patterns* section, but now additionally incorporating a large delay to key pathway processing of 24 time-steps (or 120 ms-equivalents) reflecting low stimulus-salience. The latter is the delay configuration we used to replicate Zivony & Eimer's (2020a) paradigm. In this setting, our model generates the response distribution shown in Figure 33 and the vERPs presented in Figure 34. As expected, such a large key delay leads to a substantial shift towards post-target intrusions, with the response distribution (Figure 33) now being roughly symmetrical around the +1-intrusion (i.e., the distractor stimulus directly following the target). Figure 34 also shows an expected pattern, as vN2pc as well as vP3 traces are clearly modulated by the response given, with vERP components being earliest for correct responses, intermediate for +1-intrusions and late for +2-intrusions. To reiterate, this pattern is to be expected because we delayed the firing of the blaster with our key-delay, which leads to the post-target shift in responses. The vERP pattern (Figure 34) is also in line with our expectations, as latency differences in vN2pcs (i.e., blaster-firings) should propagate to the vP3s, which they clearly do in this case. Note we have not plotted the -1-intrusion's vERPs, as they were very noisy due to the limited number of trials of this response condition.

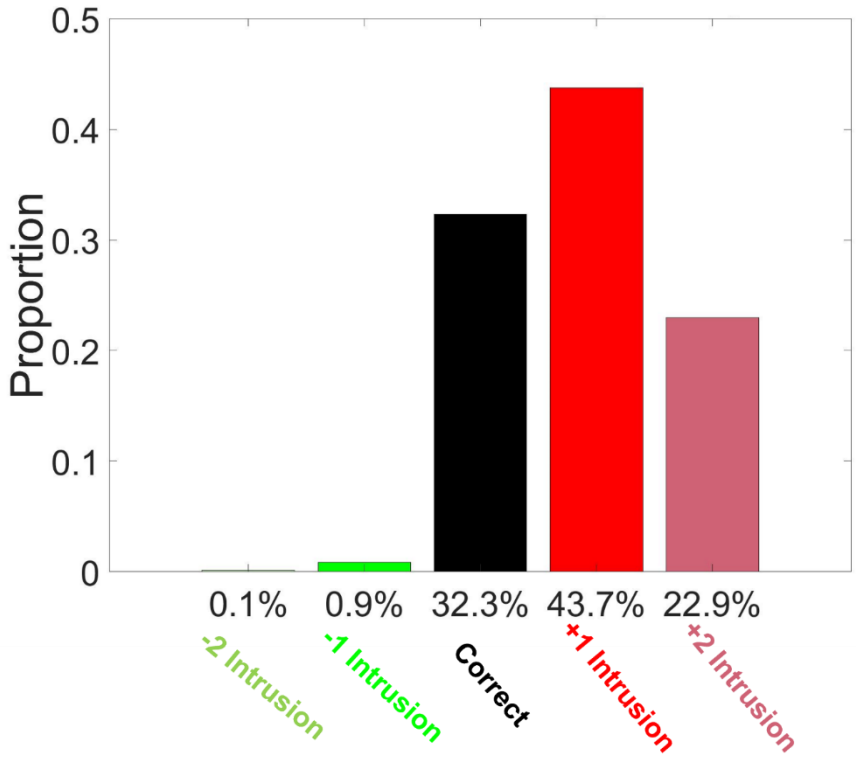


Figure 33. Response distribution of the 2f-ST^2 with inactive task-filtering in the response TFL and a τK of 24 time-steps (120 ms-equivalents).

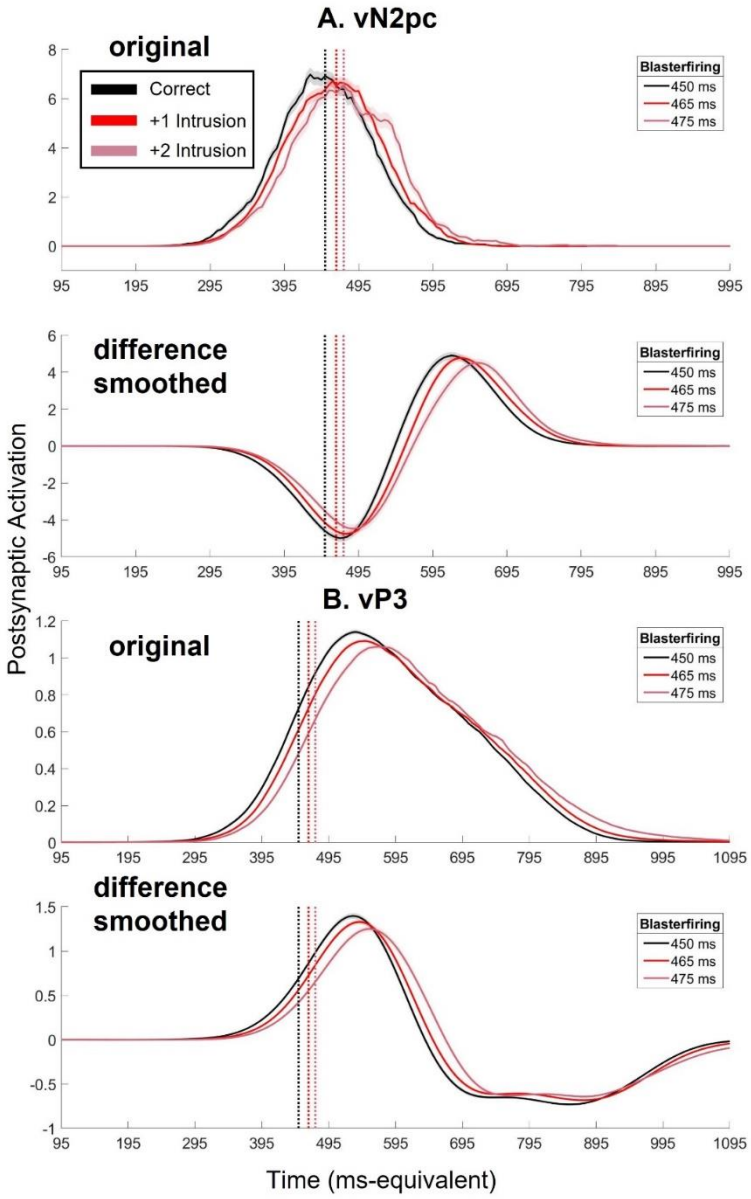


Figure 34. vERPs of the 2f-ST^2 with inactive task-filtering in the response TFL and a τK of 24 time-steps (120 ms-equivalents). Black, red, and light-red traces correspond to vERPs of correct, +1 intrusion & +2 intrusion conditions, respectively. Vertical lines indicate average blaster firing latencies (with their numerical values being shown in the legend on the right). For both vERPs, we present the original (model-output) as well as the difference smoothed version.

Active Task Filter and No Key Delay Configuration

We present the results of running the 2f-ST^2 model with an active task-filter in the response TFL/ type layer and without any added fixed delays in Figure 35 & Figure 34. This configuration lacks equivalent human data. The experimental paradigm that would resemble this model-configuration would involve an RSVP stream in which only two stimuli can for the participant be the correct response (active response TFL/ type layer, paradigm of Zivony & Eimer (2020a)) *and* the key-feature needs to be *very* salient, making key pathway processing swift (to match the absence of a key delay in this configuration). A central single RSVP stream consisting of grey distractor letters and two coloured stimuli following each

other, the target always being a digit and the +1-intruder being either a digit or a letter, could approximate this model configuration. However, one needs to consider that a single stream does not allow computation of the N2pc.

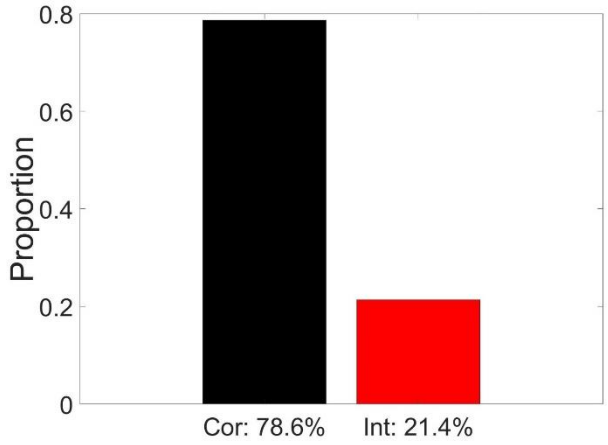


Figure 35. 2f-ST^{2}'s Response Distribution with active task-filtering in the response TFL and no additional processing delays to either pathway.

We present the accuracy distribution of this no delay configuration in Figure 35, in which 78.6% of trials were followed with a correct and 21.4% of trials with an intrusion response. The shift towards correct responses compared to Figure 31 is to be expected because the absence of a key delay (analogue to faster key feature processing speed in humans) implies an earlier blaster firing, which means that the correct response type is bound more frequently, and the succeeding intrusion type is bound less frequently.

The model's vERPs for correct and intrusion trials are shown in Figure 36. In contrast to the expected pattern presented in the previous section, it is striking that the latency differences presented in Figure 36 are much more pronounced than in Figure 32. This pattern of results is related to the *loss of responsiveness phenomenon* introduced earlier and is linked to intrusion trials having quite different characteristics from correct trials in the absence of a key delay. We explore these issues in the discussion.

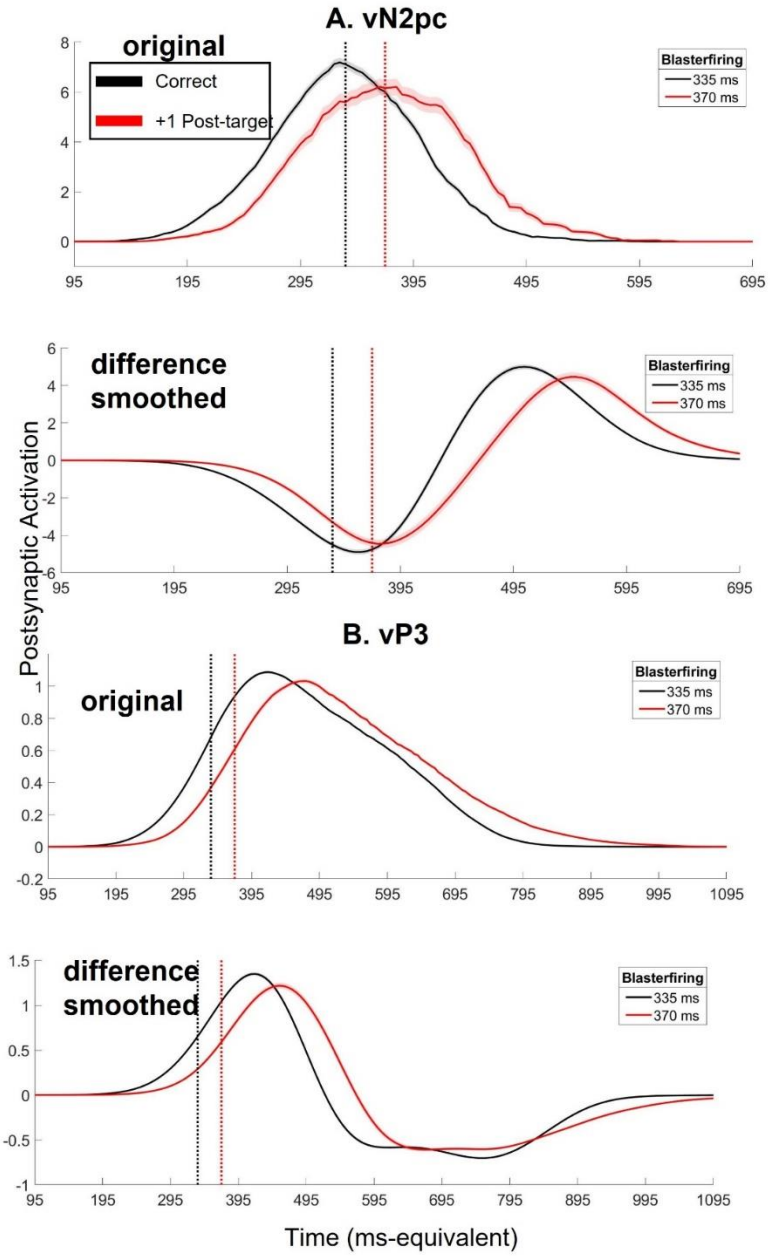


Figure 36. vERPs of the 2f-ST^2 with active task-filtering in the response TFL and no processing delays to either pathway. Black and red traces correspond to vERPs of correct and +1 intrusion, respectively. Vertical lines indicate average blaster firing latencies (with their numerical values being shown in the legend on the right). For both vERPs, we present the original (model-output) as well as the difference smoothed version.

Discussion

We present the 2-feature Simultaneous Type/ Serial Token (2f-ST^2) neural model, which is the first *neural* model that specifically accounts for the phenomenon of distractor intrusions regularly observed in RSVP experiments. Presenting the model's behavioural as well as virtual neuroimaging result patterns, we qualitatively replicate a variety of results obtained with human data. Additionally, using the original dataset, we replicated Zivony and Eimer's (2020a) N2pc latency differences and provided similar differences for Zivony and Eimer's (2020a) P3 components using Dynamic Time Warping (DTW) (Zoumpoulaki et al.,

2015). Finally, we provided a bootstrap analysis to support one main claim of our model-architecture, namely that the human N2pc and P3 EEG components are temporally correlated. The 2f-ST^2 model accounts for intrusion errors via a single mechanism of temporal variability underlying its TAE mechanism, the blaster, as well as implementing fixed delays to key or response pathway processing to replicate different empirical paradigms present in the literature. We hence argue that this model provides a complete yet parsimonious explanation of the cognitive mechanisms underlying intrusion errors in RSVP experiments or feature binding errors in the time domain in general. Much of the theory implied by the model, which was presented in initial form in Chennu et al. (2011), is consistent with the theoretical ideas presented in Zivony and Eimer's (2020b, 2020a) work.

The results of this chapter provide evidence in favour of Hypotheses 1 & 2 of this thesis, i.e., that the binding of multi-dimensional stimulus features into one percept depends on the timing of TAE, and, that the 2f-ST^2 model accounts for a broad range of findings obtained with distractor intrusion experiments. The latter was supported in the context of the 2f-ST^2 model replicating behavioural (supporting Hypothesis 2a) as well as electrophysiological (supporting Hypotheses 2b & 2c) data.

A series of behavioural findings supporting the 2f-ST^2 model

We conducted a series of analyses, which are worth some further discussion before moving on to an examination of the 2f-ST^2's results. First, the Fortunate Conjunction Experiment already presented with the initial 2f-ST^2 (Chennu et al., 2011) stands against a key prediction of Botella et al.'s (2001) model. In this data, a post-target shift in the behavioural response distribution was shown when symbol instead of letter stimuli were used as the key features in an RSVP stream of digit distractors. Hence, key feature processing speed was slower for the symbol condition. This post-target shift was accompanied with an *increase* in correct reports. In contrast, the Botella et al.'s (2001) model would predict a *decrease* in correct reports. The 2f-ST^2 model was indeed able to replicate these results solely via manipulation of τK, the delay to key pathway processing.

We further presented a re-analysis of the latency differences in ERPs between trials after a correct or intrusion response was provided. Specifically, we replicated the delay in N2pc onset shown by Zivony and Eimer (2020b, 2020a) in intrusion trials using dynamic time warping (DTW) (Zoumpoulaki et al., 2015). We additionally provide evidence that the same delay is present in the P3 component. These results extend those shown in the original studies (Zivony & Eimer, 2020b, 2020a), as the significant latency differences found with DTW suggest that not only the onset but the *whole component* was delayed after intrusion trials. The

fact that both ERP components, the N2pc as well as the P3, showed this delay in latency hints at the validity of a claim implied by the 2f-ST^2's architecture: that the N2pc and P3 components are temporally correlated. This was also mentioned by Zivony and Eimer (2020b), who stated that the results of their fourth experiment suggest a knock-on-effect of attention on the speed of WM encoding. This is reflected in our model because the time point at which the blaster fires dictates the time point at which types will be bound to tokens and hence encoded into WM.

To formally examine this strong claim of our model architecture, we additionally conducted a bootstrap analysis of N2pc and P3 DTW latencies. This analysis supported our model architecture, showing that if the same bootstrap sample is used to generate a surrogate N2pc, whichever temporal shift is observed due to the random sampling of the bootstrap will carry over to the bootstrapped P3. The N2pc has been extensively studied (Eimer, 1996; Woodman & Luck, 1999), being used as a temporal marker of attentional engagement (Callahan-Flintoft et al., 2018; Kiss et al., 2008; Wyble et al., 2020; Zivony & Lamy, 2018). Likewise, the P3 has been argued to be a marker of working memory consolidation (Kok, 2001; Polich, 2007; Verleger et al., 2005), also being linked to working memory load (Akyürek et al., 2010) and conscious perception (Pincham et al., 2016). However, a direct correlational link, as was established in our bootstrap analysis and our comparison of Correct and Intrusion trials, between the time courses of these two ERP components had yet to be established by the field. Our results are therefore the first to provide evidence that the cognitive processes underlying these two components are coupled, raising the possibility that, as suggested by the ST^2 model, the N2pc (ST^2's blaster) reflects an attentional gate that initiates an encoding/ conscious experience episode, which corresponds to the P3.

The 2f-ST2's behavioural and neuroimaging patterns

Behavioural

The 2f-ST^2 model qualitatively replicates the response distributions of Experiments 1A and 2 presented in Botella et al. (2001) with an open (i.e., no task-filter implemented) response TFL/ type layer, as well as Experiment 1 shown in Zivony and Eimer (2020a). In the latter case, this was with the implementation of a task-filter, which only allowed the target's and the +1-intruder's response type to progress to the response pathway's binding pool. We furthermore replicated the response distribution shown in the temporal feature binding experiment of Chennu et al. (2011). Additionally, using the time-point of response types completing binding to tokens, we were able to replicate a counter-intuitive pattern of reaction

times (RT) shown by Botella (1992). Crucially, the 2f-ST^2 model is able to replicate all these patterns solely via manipulations of τK and τR, the fixed delays added to key or response pathway processing, respectively. Replicating this variety of behavioural results provides strong support for our model's architecture and, in theoretical terms, the concept that variability in deployment of the TAE underlies a range of responses (such as correct or intrusion) (Zivony & Eimer, 2020a). It furthermore suggests that empirical manipulations that affect processing speeds of key or response features, such as used in Botella et al. (2001), change the symmetry of response distributions because they offset the *overall* processing of either feature, as is modelled by our two pathways. The latter concept was implemented rather similarly in Botella et al.'s (2001) model, albeit not as comprehensively as in the 2f-ST^2 model and only in its sophisticated guessing route. We have provided evidence that this mechanism is all that is required to generate the full range of conjunction patterns, whether fortunate or unfortunate.

Virtual Neuroimaging

2f-ST^2's virtual N2pc and P3 components provide a series of interesting patterns. First, the model configuration that replicated the RT findings of Botella (1992), exhibited virtual ERP components that were *earlier* for correct than *pre-target* intrusion trials (see Figure 24). This finding supports a straightforward explanation of the counter-intuitive RT finding of correct reports being *faster* than pre-target responses, even though the pre-target intruder item itself was presented *earlier in the RSVP sequence.* According to our model, the RT pattern of Botella (1992) occurred because there is an important difference in the TAE having its impact (i.e. the blaster-deployment in 2f-ST^2 and the critical time, t_c, in Botella et al.'s (2001) model), depending on whether a correct or intrusion response follows. For correct trials, the blaster on average fires temporally close to the activation peak of a strongly active target response TFL/ type neuron. This implies that the target neuron quickly wins the competition with co-active response TFL/ type neurons, as not much assistance from the blaster is required for that neuron to cross its threshold. For intrusion trials, however, the blaster's enhancement is deployed temporally further from the peak activation of the response type that gets bound. These intrusion types therefore progress more slowly to the point where they cross their threshold than target neurons in correct trials. To stress, when referring to types' threshold in this context we describe what Zivony and Eimer (2020b) call the "encoding threshold", which in the 2f-ST^2 model is located between the response TFL/ type layer and its binding pool; i.e. the TFL/ type unit being strong enough that it initiates a

binding to a token. The extra time that this requires manifests as slower RT for intrusion reports, even if the corresponding item was presented prior to the target item.

Turning on the response filter at the TFL/ type layer to simulate their empirical approach, we furthermore replicated the N2pc presented in Zivony and Eimer's (2020a) Experiment 1 (Figure 25), showing a pattern of decreased amplitude and delayed onset as well as overall time-course of the component in our vN2pc (Figure 32A) for intrusion trials. Additionally, we presented the author's (Zivony & Eimer, 2020a) P3 (Figure 25), which demonstrated a delayed component after intrusion trials. Importantly, both intruder-induced delays affected the *whole* ERP components (i.e., not *only* their onsets) and were found to be statistically significant after running the DTW analysis. We presented the 2f-ST^2's virtual P3 component for this model-configuration in Figure 32B, which again qualitatively replicates the pattern of decreased amplitude and a delayed overall component. Adding to the temporal link between the human N2pc and P3 suggested by our bootstrap analysis (Figure 30), these virtual neuroimaging findings provide further support for a not just temporal, but functional link between these two components in human cognition. Such a link was hinted at in Zivony and Eimer (2020b), who observed a knock-on-effect of attention on speed of WM encoding. This is precisely in line with the 2f-ST^2's architecture and that of STST in general ((Bowman et al., 2008; Bowman & Wyble, 2007), in which the time-point of blaster-firing (attention) largely determines the time-point of TFL/ type-nodes crossing the corresponding binding pool threshold (WM encoding).

In a final exploratory set of analyses, we generated the vERPs of our model without task-filtering in the response TFL/ type layer and a τK of 24 time-steps (120 ms equivalent) as well as with an activated response TFL/ type layer and no delays to either pathway. The former analysis yielded an expected pattern of results (Figure 34), with vERPs of +1-intruder trials now sitting in-between vERP traces of correct and +2-intruder trials, which were presented temporally before and after the +1-trace, respectively. This pattern is what might be expected, since the sequence of vERP-traces matches that of virtual RSVP items presented in the stream (as opposed to, for example, the pattern shown in Figure 24, in which correct trials were associated with earlier vERP traces than -1-intruder trials, even though correct items were presented *after* -1-intruder items). However, and strikingly, the latter analysis, in which the model was run with an active response TFL/ type layer and no delays to processing whatsoever, yielded an unexpected pattern of vERPs. Figure 36 displays these interesting vERP traces, which show substantially larger latency differences between correct and intruder

trials than Figure 32, in which a large key delay was implemented in order to replicate Zivony and Eimer's (2020a) Experiment 1 response distributions. We undertook a detailed analysis to investigate the origin of this pattern and to understand its underlying dynamics, which we present next.

Conclusion

To conclude, in this chapter we (alongside empirical analyses) modified and adopted the 2f-ST^2 model to simulate experiments, such as those conducted by Zivony and Eimer (2020a), in which only the target and the +1-intruder item shared the same response dimension and hence could have been the target. Despite accomplishing the latter successfully, we encountered some unexpected behaviour of the model during the process. Pursuing the origins of this behaviour, we uncovered particularly interesting dynamics intrinsic to the 2f-ST^2 model, which will be presented in detail next.

Chapter 5 – The 2f-ST^2 Model's Loss of Responsiveness Phenomenon

Introduction

In Chapter 4 we presented the 2f-ST^2 model and demonstrated how it replicates a variety of behavioural as well as neuroimaging findings yielded by various experimental paradigms conducted by Botella and colleagues (Botella et al., 2001), as well as Zivony and Eimer (2020a). The findings presented in the previous chapter thus highlight the flexibility and simplicity of the 2f-ST^2 model, as the model is able to replicate various empirical and neuroimaging findings without major modifications to its underlying architecture or internal processes.

However, it is worth emphasizing that the 2f-ST^2 model was initially developed by Chennu and colleagues (Chennu, 2009; Chennu et al., 2011) to specifically replicate the experimental paradigm adopted by Botella et al. (2001), who ran experiments in which all stimuli could have theoretically been the target stimulus as they all shared the same response dimension (e.g., all being digits). Adopting this initial version of the 2f-ST^2 model and extending it to additionally simulate and account for the experiments of Zivony and Eimer (2020a), who ran experiments in which *only one* distractor stimulus shared the response dimension of the target and hence represented the only possibility of committing an intrusion error, was accompanied with a few significant challenges. To reiterate, in the model we

simulated the latter experimental paradigm by only allowing the target- and the +1-neuron to be active in the response TFL/ type layer as that layer implements task-relevance (e.g., category-matching of distractors to the target) in the response dimension. In contrast, in the initial 2f-ST^2 model all neurons in this layer could excite their corresponding response binding pool unit.

Figure 37 presents the results of running the 2f-ST^2 model after configuring it to simulate Zivony and Eimer's (2020a) Experiment 1 via the implementation of task-filtering in the response TFL/ type layer as well as a τK value of 24 time-steps (the equivalent of approximately 120 milliseconds (120 ms-equivalents)). Note that the results of Figure 37 were previously presented in Figure 31 & Figure 32. A τK of 24, meaning a significant delay in key pathway processing, was adopted to resemble the less salient key feature of a shape surrounding target stimulus used in Zivony and Eimer's (2020a) Experiment 1. Figure 37A shows the response distributions of the model, which were in line with those found empirically (see the previous chapter for more detail). Importantly, the most intriguing and challenging finding after running the 2f-ST^2 model in this configuration were the model's virtual ERPs, particularly the vP3 (Figure 37B).

We argued in the previous chapter that one can only expect a qualitative fit between virtual and human ERPs. Looking at the vP3 patterns of Figure 37B (see Figure 32A for the corresponding vN2pc), one might argue that the vERPs are qualitatively in line with the human ERPs, as both show earlier and higher-amplitude components for correct than intrusion trials. Nonetheless, these amplitude and latency-differences seem minimal in the vP3, especially next to the equivalent effects yielded by the model with active task filtering in the response TFL/ type layer but without any delays to either pathway's processing (presented previously in Figure 36 and discussed in detail later on). This intriguing pattern motivated us to look more thoroughly into the 2f-ST^2's internal dynamics and processes to truly understand the impact of task-filtering in the response TFL/ type layer. This was important to understand since we added response pathway task-filtering in a model which was originally not designed for this purpose. Doing so, we uncovered a series of unexpected findings that exposed interesting ways in which the 2f-ST^2 model behaves. Besides introducing these findings, we will explain how an intriguing phenomenon, which simultaneously accounts for the diminished latency differences in vERPs mentioned above *and* allows us to replicate Botella's (1992) counterintuitive reaction-time (RT) findings, emerged from this series of analysis: *The loss of responsiveness phenomenon of the 2f-ST^2 model*.

The current chapter will introduce the loss of responsiveness phenomenon sticking to the approximate chronology with which we conducted analyses. Therefore, a different structure than in the other research chapters will be adopted, as we will present analyses' findings before discussing their implications one at a time and thereby continuously augment our understanding of how responsiveness is lost in the 2f-ST^2 model. Specifically, the chapter will commence with a simple explanation on why the vP3 latency-differences presented in Figure 37B came about, before considering the 2f-ST^2 model's behaviour without processing delays to either pathway and demonstrating that τK is the parameter that modulates the extent of responsiveness being lost. Subsequently, we will demonstrate that the loss of responsiveness allows us to replicate the counterintuitive RT patterns of Botella (1992) and, finally, provide predictions and recommendations for future research that might further elucidate the loss of responsiveness.

This chapter will provide evidence in favour of Hypothesis 1 of this thesis, i.e., that the binding of multi-dimensional stimulus features into a single percept depends on the timing of TAE. In addition, we will not only provide evidence in favour of this hypothesis, but also provide results that emphasise the importance of whether key and response feature processing occurs synchronously. This chapter further relates to Hypothesis 2c of this thesis, which states that the 2f-ST^2 model's virtual P3 dynamics replicate electrophysiological findings obtained with the human P3 component. Even though we will not present further evidence in favour or against this hypothesis, the analyses of this chapter will provide insight into *how* the 2f-ST^2 model's vP3s are generated, particularly analysing and explaining cases in which the model's loss of responsiveness impacts the extent of qualitative fit between empirical and equivalent virtual P3s.

The 2f-ST^2 Model's Loss of Responsiveness

A simple aspect: The pattern of vP3s shown in Figure 37B involves a number of factors that interact with each other quite intricately. The simplest aspect associated with this pattern is that neurons in the response pathways' TFL/ type layer require the blaster's enhancement to their membrane potentials (VM) in order to cross the effective threshold and excite their respective binding pool units. Figure 38 & Figure 39 show this requirement for enhancement, presenting the response TFL/ type layer's VMs of the target- and +1-neuron in correct and intrusion trials. The blaster's enhancement is uniquely important in correct trials to enable the target neuron to cross the threshold to the response binding pool (henceforth called T-C trials and plotted as a black solid line in Figure 38 & Figure 39). This is because the target neuron

in correct trials has a VM close, but just below, the threshold (at .35, indicated by the green horizontal line) for a rather long time-period prior to the blaster firing, see Figure 38. As mentioned in Chapter 4, Figure 38 - Figure 40 display results after a single run of the model, implementing the same rng-configuration as the initial 2f-ST^2 model (Chennu et al., 2011), as adopting five runs while saving individual neurons' activation (membrane potential or input channel) levels was not feasible due to the substantially increased computational cost. Figure 38 clearly shows that the target neuron for T-C outcomes (black solid line) starts to ascend *earlier* than the +1-neuron in intrusion trials (called +1-I, red dashed line). For example, at around 150 milliseconds, the black solid line is considerably above the red dashed line. If this latency difference of their VM-onsets would translate to their postsynaptic potentials, it would lead to vP3s that more clearly resemble the latency-differences between correct and intrusion trials shown in Zivony and Eimer's (2020a) human ERPs (Figure 25), i.e., correct P3s substantially earlier than intrusion P3s. However, because the target neuron for T-C outcomes struggles to cross the threshold (at .35) without help from the blaster, it crosses the threshold at about the same time as the +1 neuron in +1-I (the +1-intruder being bound in intrusion trials) outcomes (see solid black and dashed red lines crossing green line at around 395 milliseconds), leading to the vP3s of Figure 37. Note that blaster-firing latencies in vERP or neurons' postsynaptic activation plots, such as those of Figure 37 - Figure 39, depict *averages* across all trials of a given response condition, and for some trials, the firing will be a good deal earlier or indeed later. We will present the distributions around these blaster-firing latency-averages later (Figure 41).

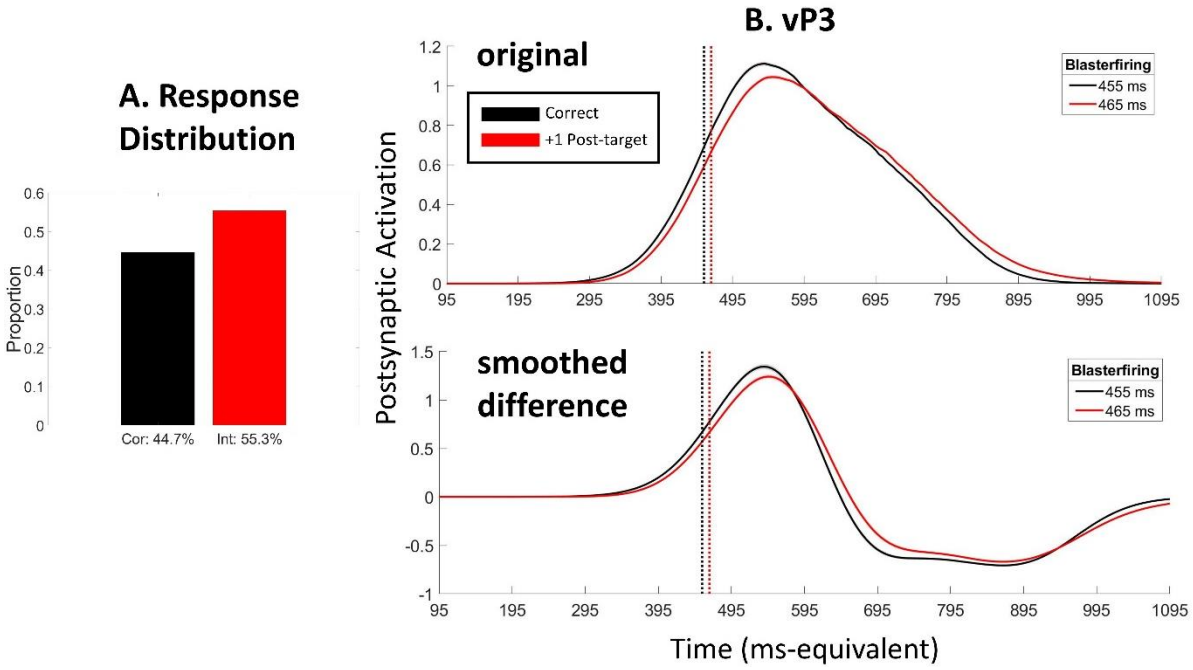


Figure 37. Results of running the 2f-ST^2 model with task-filtering in the response pathway and a τK of 24 time-steps, replicating Zivony and Eimer's (2020a) Experiment 1. We present the 2f-ST^2 model's response distribution and vP3s in Panels A & B, respectively. Note that these results were previously presented in Figure 31 & Figure 32.

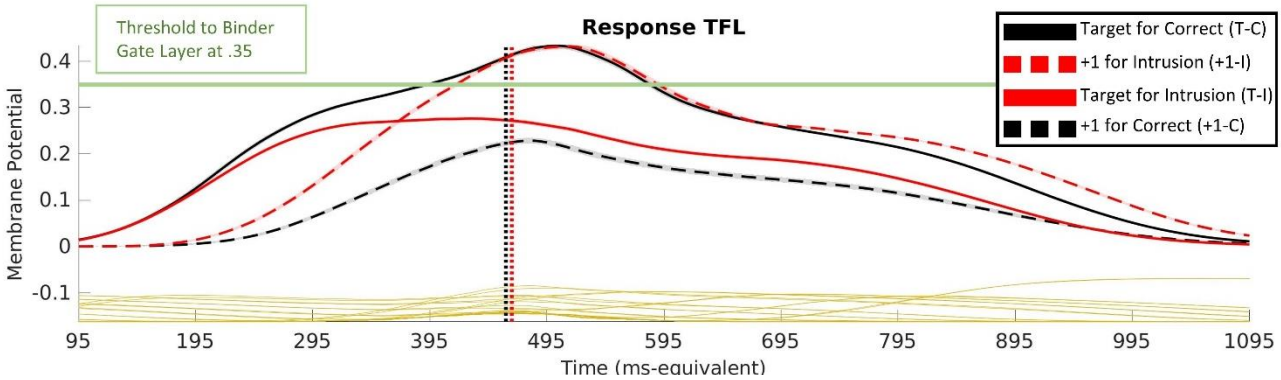


Figure 38. Response TFL/ type layer's neuronal membrane potentials (VMs). Black = correct- & Red = intrusion-trials. Solid = target unit & dashed = +1 unit. Therefore, black solid, red dashed, red solid, and black dashed lines represent VMs of target unit in correct trials (T-C), +1 unit in intrusion trials (+1-I), target unit in intrusion trials (T-I), and +1 unit in correct trials (+1-C). This is also indicated by the legend in the top right. The green horizontal line indicates the VM threshold to respective binding pool units at .35. Vertical dotted lines indicate average blaster firing latencies of correct (black) and intrusion (red) trials. Gold traces indicate category-nonmatching distractor types, which do not meet task demands.

The T-C's slow VM-ascent: A noticeable characteristic of the T-C's VM (black solid) is its slow ascent towards threshold just before the blaster hits (on average at ~ 455 ms), depicted in Figure 38 & Figure 39. This is, as we argue, the most crucial aspect of all dynamics in this configuration of the 2f-ST^2 model, i.e., with an active response TFL/ type layer and a large τK. Again, if this ascent was steeper, the target neuron for the T-C outcome would cross the threshold earlier and the vP3 pattern would resemble that of humans much more closely. In fact, the T-C trajectory is one out of only two cases across all settings of the model in which a unit is bound to a token (because of the timing of blaster-firing), while the rate of VM-

increase (i.e., its gradient) is *decreasing*. We will present the second such case shortly when introducing how the 2f-ST^2 model's loss of responsiveness accounts for the counterintuitive RT findings of Botella (1992).

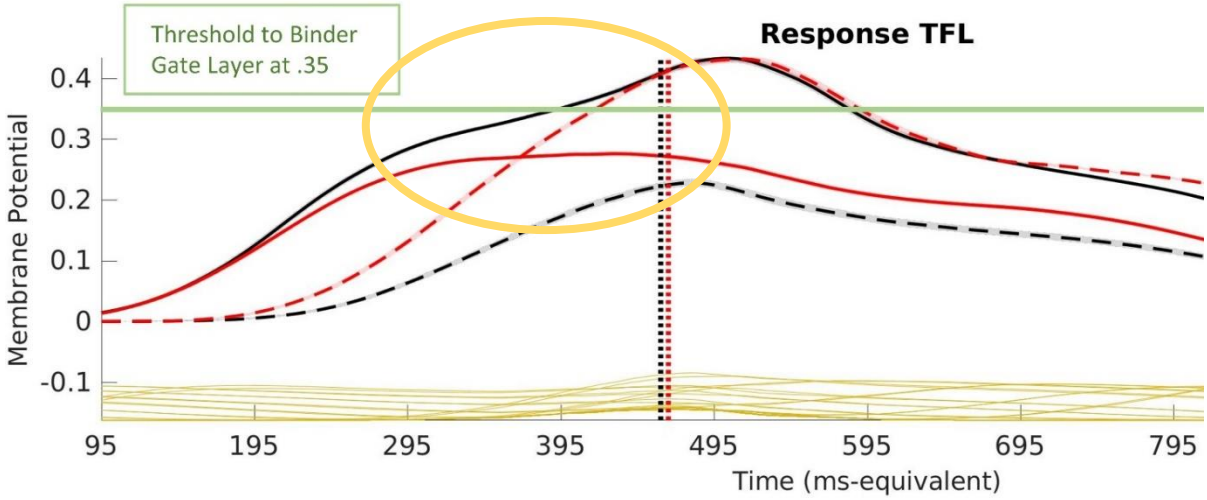


Figure 39. Response TFL/ type layer's VMs (plot presented in Figure 38) zoomed in to more clearly show the time-interval of 295-495 ms-equivalents, in which the T-C unit fails to excite its binding pool unit sufficiently without the blaster's enhancement, which on average comes at about 455 ms.

Decreased excitation from the item layer is the most crucial reason for this behaviour of the T-C's VM-trace. The item layer is the *main driving force* in the model's architecture that excites the TFL/ type layer. Figure 40 illustrates the response item layer's postsynaptic activation, showing how the postsynaptic activation values for the T-C outcome (which in turn excites the respective TFL/ type neuron) has been decreasing during the time-interval of interest. This is simply due to the sequential nature in which items are presented to the model. Also note how much the item layer's postsynaptic potential has declined when the blaster fires (vertical black dotted line at ~455 ms). Note that blaster-firing latencies at ~455 ms indicate averages, which vary across trials (we present the distributions around these averages later in Figure 41). In other words, delaying the key pathway, delays blaster firing, giving time for target response item layer traces to decay into a less responsive state. This is why, even though response TFL/ type neurons almost always require the blaster's enhancement to cross the effective threshold to their binding pool units, requiring the blaster's enhancement is particularly relevant in cases in which the bottom up excitation from the item layer has been decreasing (meaning a loss of responsiveness). Further, the T-C's decreasing response item layer activation at the time of blaster firing, shown in Figure 40, justifies an important statement made repeatedly in this chapter: types that have lost their responsiveness require more assistance from the blaster (since they are actually decaying at the response item layer)

to cross the effective threshold to their binding pool units.

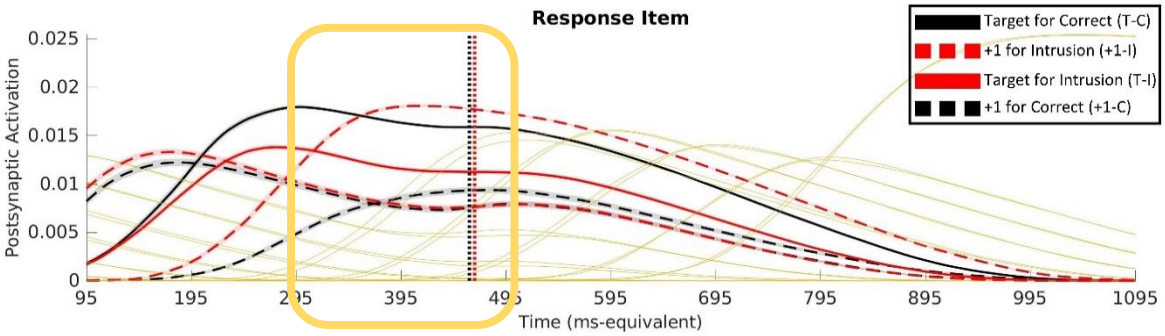


Figure 40. Postsynaptic activation values of response item layer units showing a decline in excitation from this layer to the subsequent response TFL/ type layer during the time-interval of interest of 295-495 ms (indicated by yellow box). The plotting conventions follow those of Figure 38 & Figure 39, except that dashed lines now indicate activation levels of -1 in addition to +1 units. This plot indicates that the T-C's item layer activation has been decreasing, leading to decreasing response TFL/ type layer VM values before the blaster fires at ~455 ms on average. Note that the blaster fires with temporal variability and that the dotted vertical lines only indicate the mean blaster firing latency of a condition, not the distributions around means. We will provide blaster firing latency distributions later in Figure 41. Gold traces indicate category-nonmatching distractor types, which do not meet task demands. Note that gold traces' amplitudes are comparable to red and black traces' amplitudes because the model's item layers do not implement task filtering.

A summary: The loss of responsiveness of the target and +1-intrusion neurons: As key-delay increases, trials that end up with a correct or an intrusion response become progressively similar in terms of their characteristics, e.g., stimuli's input-strengths. Without any key-delay, only trials in which, due to chance, the blaster was fired late, the target-neuron's input was weak and/or the +1-neuron's input was strong led to the model binding the +1-neuron's response feature to the token and hence committing an intrusion error. However, with a large τK of, e.g., 24 time-steps, trials in which the model commits an intrusion error are much like trials in which the model encodes the correct response feature to the token in terms of stimuli's input-strengths and blaster-firing latencies. The latter can be seen in Figure 41 in which blaster-firing latency distributions are plotted. Importantly, vertical dotted lines indicate across-trial average latencies of correct (black) and intrusion (red) trials, which sit closer together in the key-delay condition. We call this phenomenon a *loss of responsiveness* of the response TFL/ type target- & +1-neurons: These neurons were in close to no competition with each other without a key-delay and it was usually unambiguous which response feature will be encoded to a given token. However, once τK is large (24 time-steps), an important competition (due to lateral inhibition) emerges between them, and the winning neuron usually beats its competitor by a small margin only. Hence, the range of trials that are now categorized as correct or intrusion changes and become more similar to one another. This is why their vP3s come out being so similar: because the underlying trials' characteristics (in terms of blaster-firing latencies and input stimulus strengths) are.

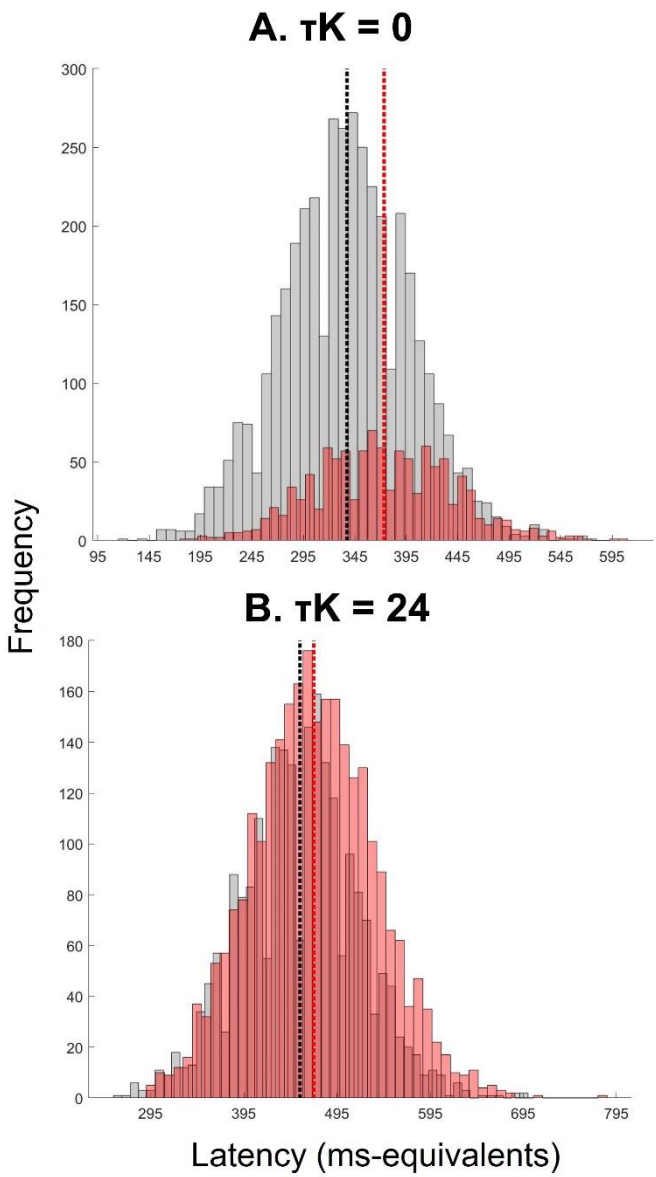


Figure 41. Distributions of blaster-firing latencies of the 2f-ST^2 model with active task-filtering in the response pathway and τKs of 0 (A) & 24 time-steps (B) for correct (black) and intrusion (red) trials. Note that y-axis limits differ between panels. Mean blaster firing times are shown with vertical dotted lines.

The model without a key-delay (τK = 0): The mechanisms and interactions illustrated earlier imply a major change to our model: the target as well as the +1-intrusion type lose their responsiveness. This also means that trials are re-distributed between response conditions as one increases τK to a value of, e.g., 24 time-steps. The re-distribution of trials into correct and intrusion conditions means that the characteristics of the trials contained in a particular condition change, too. Trial-characteristics refer to trials' blaster-firing latencies and stimulus input strengths. Note that a full run of the model always involves the same set of virtual RSVP streams and that the described change in conditions' trial-characteristics is due to the re-distribution of trials into conditions, not due to trials themselves changing in any way. Figure 42 shows the response distributions as well as the vP3s if *no delays* are

implemented in the model (previously shown in Figure 35 & Figure 36). In this setting, intrusions occur because the blaster fires with some variability and stimuli vary in their input-strengths across trials. Hence, intrusions are the result of a clearly later blaster firing as well as of a weak target- and/or a strong +1-unit response feature. Intrusions in this setting are rather unambiguous, meaning they only occur in rare circumstances in which the target unit's response-TFL/ type layer activation has decayed so much that the +1 is bound. In humans, this unambiguity could relate to an illusory percept that has been seen well, i.e., the subject reporting high confidence and the 'wrong' stimulus-identity. In the next paragraph, we will show how τK, the delay to the key-pathway, is the parameter that modulates the target- & +1-unit's responsiveness, decreasing it progressively as τK is increased.

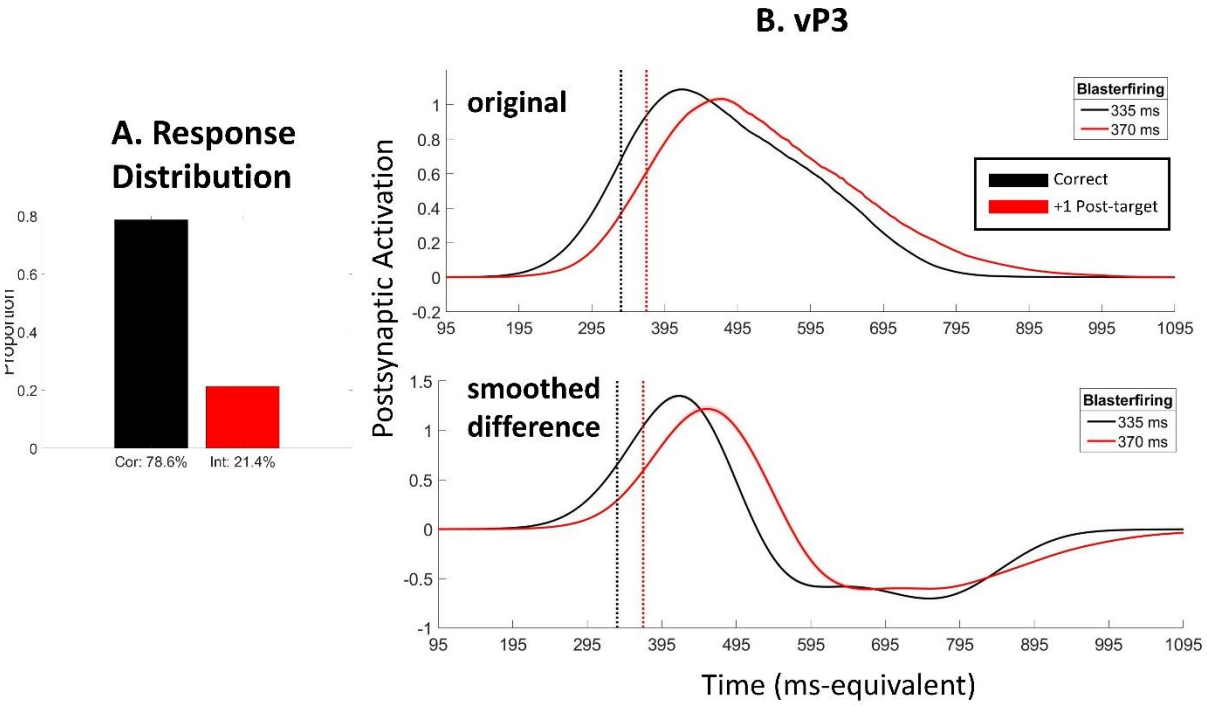


Figure 42. Results of running the 2f-ST^2 model with task-filtering in the response pathway and no delays to either pathway's processing. We present the 2f-ST^2 model's response distribution and vP3s in Panels A & B, respectively. Note that these results were previously presented in *Figure 31.* Response distributions of Zivony & Eimer's (2020) experiment 1 (A: Original Study, B: Replication) (left) and 2f-ST2 (right) with task-filtering in the response pathway and a τK of 24 time-steps (120 ms-equivalents). Black and red bars correspond to correct and intrusion trials, respectively. Letter distractor in Panel A correspond to trials in which a distractor letter was presented immediately after the target digit and, thus, no intrusion errors were possible. Digit distractor refers to trials in which a digit distractor was presented immediately after the target digit, hence allowing for intrusion errors to occur.Figure 35 & Figure 36.

τK is the Parameter that modulates loss of responsiveness

In the following analyses, we investigated τK values of 0, 6, 12, 18 and 24 time-steps to show that the value of τK modulates the loss of responsiveness. This parameter space corresponds to ms-equivalents of 0, 30, 60, 90 & 120 ms. The results presented in this

paragraph (see Figure 43 - Figure 45) were obtained after a single run of the model, implementing the same rng-configuration as the initial 2f-ST^2 (Chennu et al., 2011).

Same Neuron, Different Delays: First, we are presenting the response TFL/ type layer's postsynaptic activation traces of the same neuron across our key-delay (τK) range. For example, the top left panel of Figure 43 shows the T-C's response TFL/ type layer's activation-trace. The top row of Figure 43 show the 'winning'-neurons, meaning the neuron that was later bound to the token, either being the T-C or the +1-I. For these neurons, increasing τK leads to rather small changes in activation trace form: the trace of T-C gets slightly broader and is of decreased amplitude as τK increases and an opposite pattern is hinted at with respect to +1-I. We furthermore present the two 'losing' items, the +1-C (+1-node in correct trials) and the T-I (target note in intrusion trials) in the bottom two panels of Figure 43. The +1-C's (bottom left in Figure 43) activation increases steadily with an increase in τK. The T-I pattern (bottom right) is not as clear-cut due to its no-delay trace, but overall suggests the opposite pattern: a decrease in activation as τK increases. The most important panel here is the bottom left, +1-C. It suggests a fundamental reason for response TFL/ type units' loss of responsiveness is that increasing τK introduces a *competition* between units, which is absent without such a delay. Specifically, if the blaster fires early (as it does without a τK), most of the time the target unit is hit by it, which in turn inhibits the +1-unit strongly, leading to low +1-C amplitude (purple trace in the bottom left panel of Figure 43). However, once a large τK is implemented, the target-unit's bottom-up driving force (from the item layer) is decaying when the blaster hits the response TFL/ type layer, as shown in Figure 40, meaning that the target and the +1-unit regularly have similar response TFL/ type layer-amplitudes at that point in time. Therefore, the units are *inhibiting each other to a similar extent*. Now, which response feature is bound to the token is not as clear-cut as it is with a τK of 0. As τK increases, the competition between the target- & the +1-unit is getting fiercer, and the winner now is usually winning only by a small margin, leading to higher amplitude of the teal trace in the +1-C panel. We present the same plots of Figure 43 as a larger version in Figure 44.

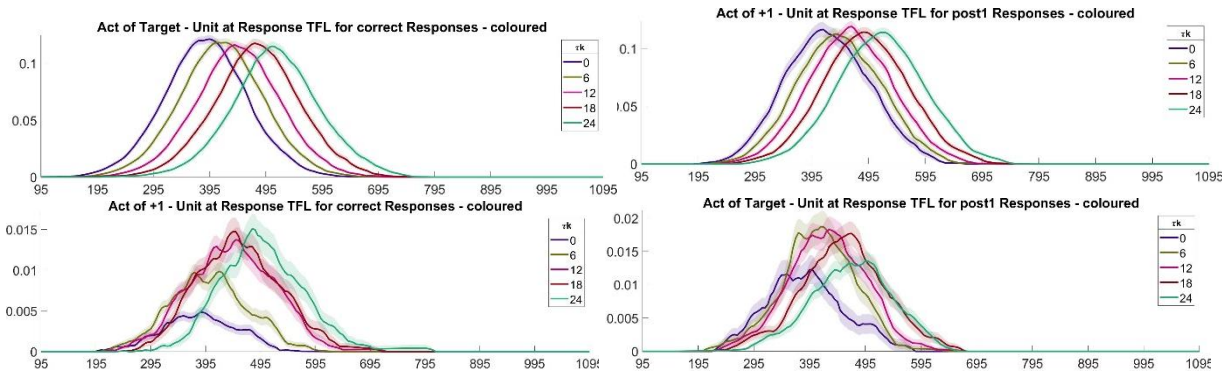


Figure 43. Layout 1 - Response TFL/ type layer's postsynaptic potentials for the same neuron across τK values of 0, 6, 12, 18 & 24 time-steps. We present the target neuron in correct trials (T-C, top left), the +1 unit in intrusion trials (+1-I, top right), the +1 unit in correct trials (+1-C, bottom left), and the target unit in intrusion trials (T-I, bottom right).

These plots illustrate the impact of increasing τK on each neuron's response TFL/ type postsynaptic potential values, which in respect of form of the activation trace are rather minimal for the top two panels, showing neurons that win the competition of this layer. The bottom two panels show the 'losing' neurons. Especially the bottom left panel, showing the +1 unit in correct trials, illustrates how increasing τK introduces a fiercer competition between the target and the +1 neuron in this layer. This can be seen since a τK of 0 means low amplitude of the losing +1 unit, because the winning target unit is inhibiting the +1 unit strongly. However, this inhibition is getting weaker as τK increases, leading to increased amplitudes of the +1 unit in correct trials as τK increases and, therefore, a fiercer competition.

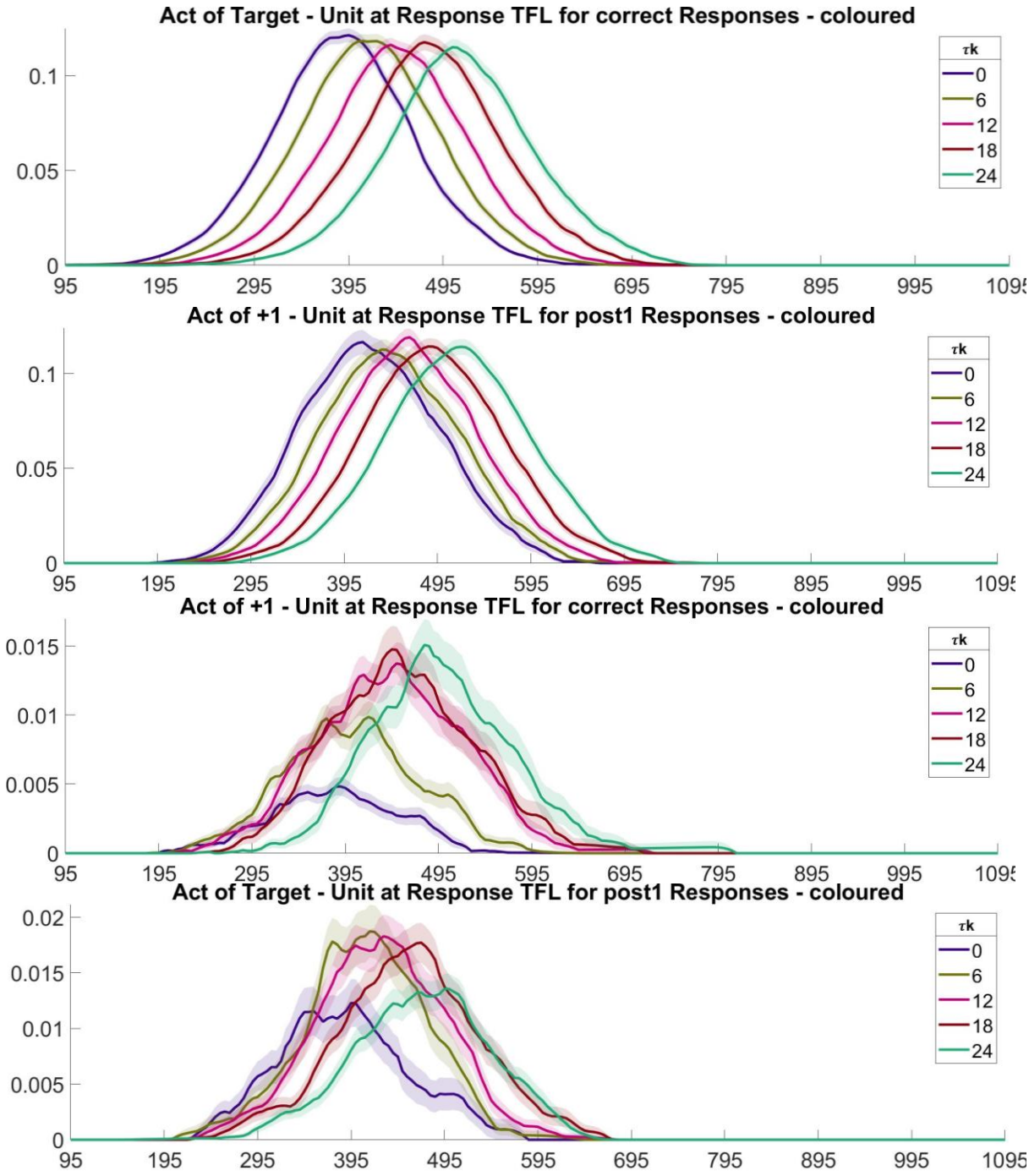


Figure 44. Layout 2 - Response TFL/ type layer's postsynaptic potentials for the same neuron across τK values of 0, 6, 12, 18 & 24 time-steps. We present the target neuron in correct trials (T-C, top left), the +1 unit in intrusion trials (+1-I, top right), the +1 unit in correct trials (+1-C, bottom left), and the target unit in intrusion trials (T-I, bottom right). These plots are identical to those presented in Figure 43, just in a larger format to improve visibility.

Differences between target- & +1-unit: We conducted another set of analyses in which we investigated the hypothesis that the target- & +1-units are becoming more similar as τK increases. To this end, we subtracted the response TFL/ type layer's postsynaptic activation as well as membrane potential time-series of the +1-unit from that of the target-unit (i.e., target - +1), separately. We then computed the area between resulting curves (AUC) from horizontal

lines at zero. If such an AUC value is zero, it means that the time-series of interest of the target- and the +1-unit are identical. A larger AUC value would therefore mean substantial differences between the two time-series being subtracted. Our loss of responsiveness argument would manifest in such an analysis as a decreased AUC value as τK increases. We present the results of this analysis in Figure 45, which shows the membrane potentials in the left column and the postsynaptic activation time-series in the right column, with neurons' subtraction time-series being superimposed and plotted in green. In line with previous plots, time-series of T-C are plotted as black solid, the +1-C as black dashed, the T-I as red solid, and the +1-I as red dashed lines. Vertical dotted lines indicate average blaster-firing latency of correct (black) and intrusion (red) trials. The patterns in Figure 45 support our hypothesis of a progressive loss of responsiveness, as both, membrane potentials as well as postsynaptic activations display a progressive decrease in AUC as τK increases.

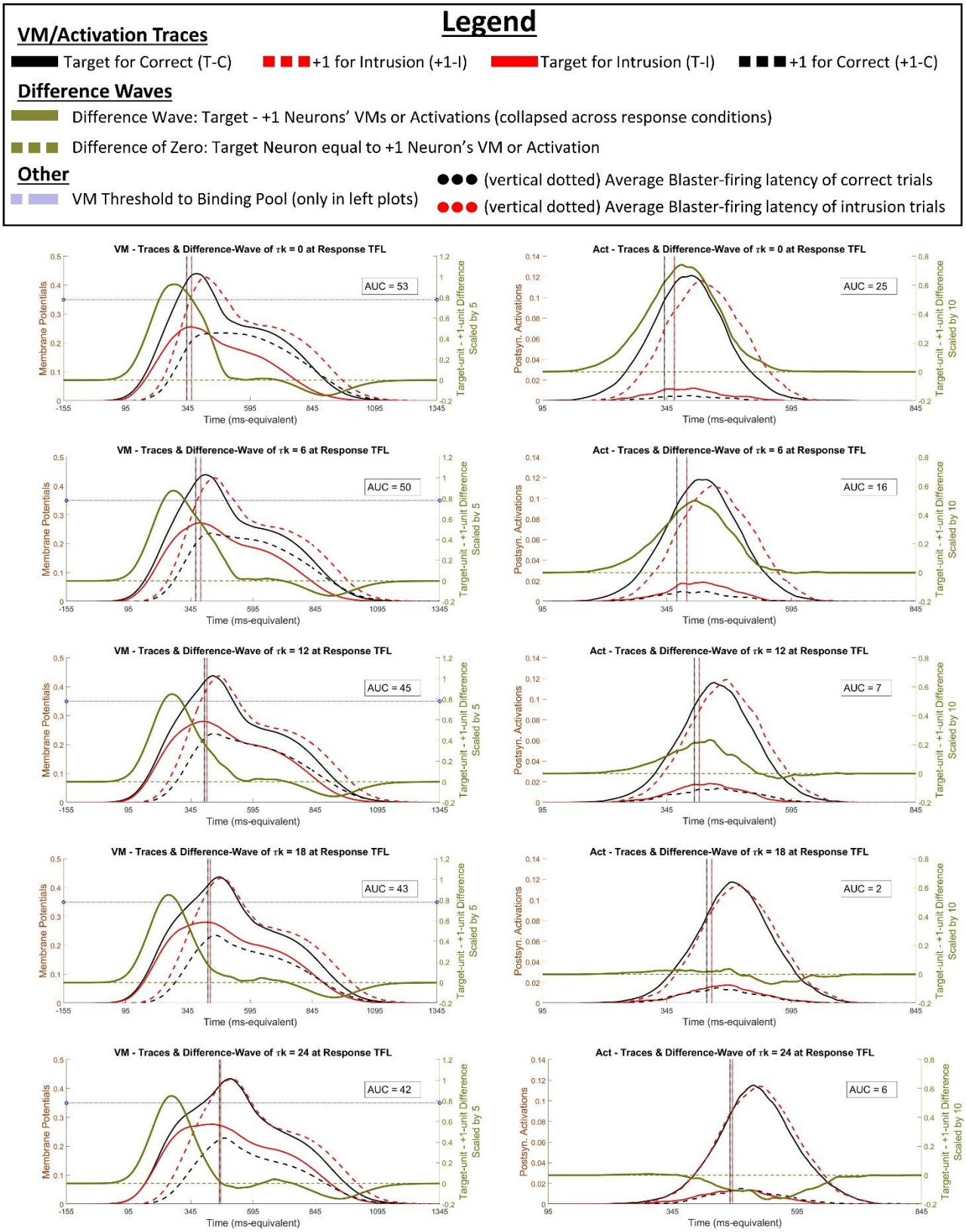


Figure 45. The target and +1 units' response TFL/ type layer's VM (left) as well as postsynaptic activation (right) time-series become more similar as τK increases. τK increases in steps of 6 from 0 to 24 time-steps (i.e., 0, 6, 12, 18 & 24) from top to bottom.

In line with previous plots, we plot time series of the target unit in correct trials (T-C) as black solid, the +1 unit in correct trials (+1-C) as black dashed, the target unit in intrusion trials (T-I) as red solid, and the +1 unit in intrusion trials (+1-I) as red dashed lines. Further, we present VM (left) and postsynaptic activation (right) differences between target and +1 units over time (i.e., target - +1) as green lines. Blue horizontal lines in VM plots (left) indicate the threshold to the binding pool at .35. AUC values between green curves and time-series differences of zero (meaning identical time-series of target and +1 units) are provided in plots' legends. Both analyses support our claim of a progressive loss of responsiveness since both analyses demonstrate a progressive decrease in AUCs as τK increases, implying that the target and +1 units' time-series become more similar as τK increases.

Explaining counterintuitive RT patterns

Figure 46 displays the membrane potentials of the response TFL/ type layer for the 2f-ST^2 model with task-filtering inactive and no delays to either pathway. Light green, green, black, red, and light red lines in Figure 46 correspond to the VMs of -2, -1, target, +1, and +2 neurons, respectively. The membrane potentials shown in Figure 46 were obtained after a single run of the model, implementing the same rng-configuration as the initial 2f-ST^2 (Chennu et al., 2011). Panels A, B, D and E show -2-, -1-, +1- and +2-intrusions, respectively. Panel C shows the correct condition. Figure 46 can therefore be viewed as the analogue of Figure 38 and Figure 39, in which we displayed the response TFL/ type layer's membrane potentials to illustrate the loss of responsiveness phenomenon in our other model configuration (with active task-filtering in the response TFL/ type layer). Figure 46B shows the membrane potential of the -1-neuron in -1-trials, which is strikingly similar to that of the target-neuron in correct trials (previously called T-C), shown in Figure 38 and Figure 39. As described in depth with respect to the T-C neuron, the -1-neuron in -1-trials demonstrates a similar decrease in its membrane potential's gradient just before the blaster enhances neurons' activation-levels (indicated with a yellow circle in Figure 46B). This membrane potential gradient-decrease is even more visible looking at the -2-neuron in -2-trials (Figure 46A). We argue, therefore, that this scenario is, also, a case of pre-target neurons losing their responsiveness: pre-target intrusions in this context occur after the blaster enhances the respective pre-target neuron at a point in time at which its driving item-layer excitation has already started to decline. As a result, the pre-target neuron requires comparably more time to receive enough enhancement from the blaster to excite its corresponding binding pool unit sufficiently for its type to be bound to the currently active token. This additional time ultimately results in the counterintuitive RT patterns found by Botella (1992) in humans and previously shown in Figure 23 in the 2f-ST^2 model. The VM threshold of response TFL/ type neurons to the binding pool at .35 is indicated by the horizontal blue line in Figure 46.

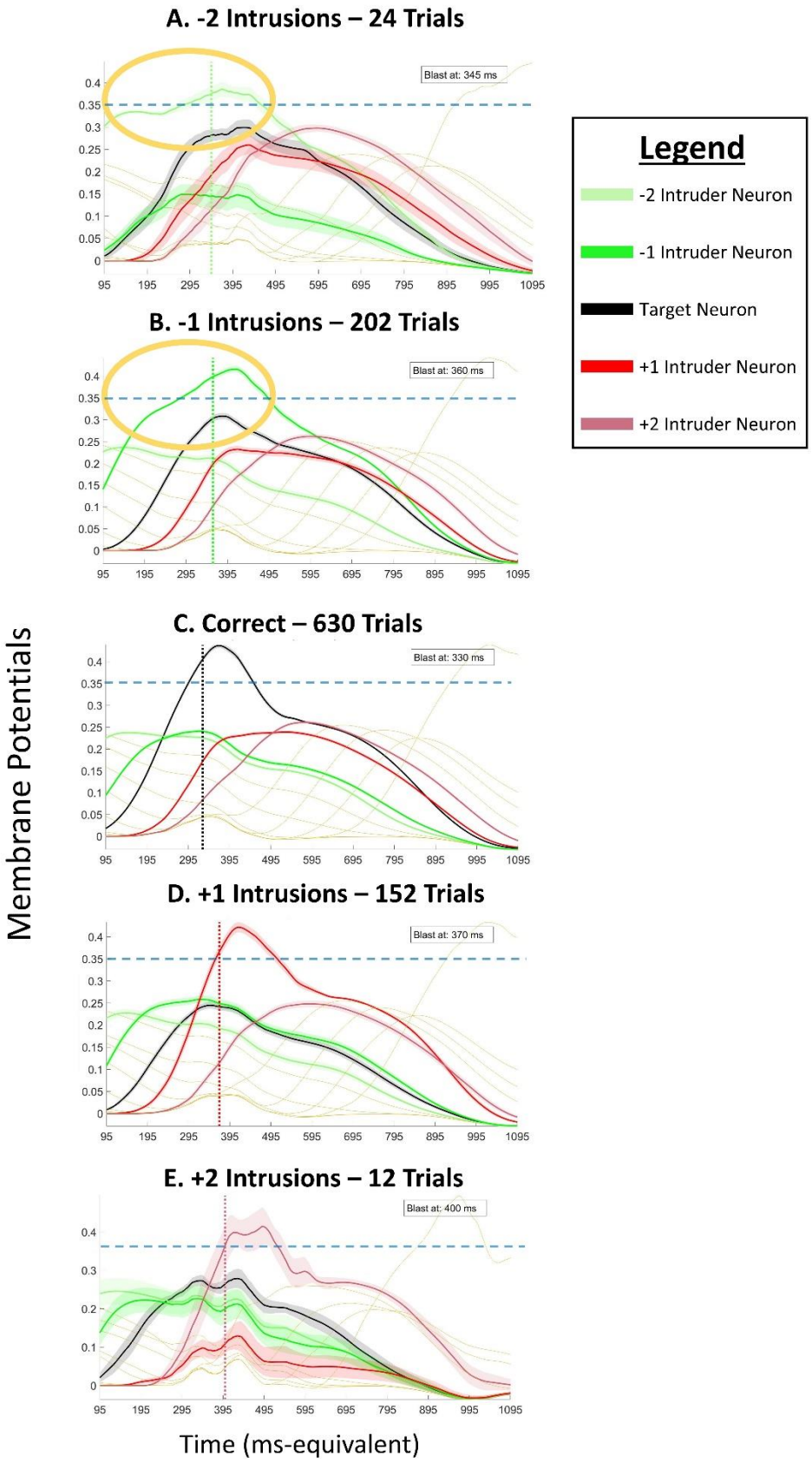


Figure 46. Loss of responsiveness explaining counterintuitive RT patterns. The membrane potentials of the response TFL with the response filter being turned off and without any delays to either pathway are plotted. . Light green, green, black, red, and light red lines correspond to the VMs of -2, -1, target, +1, and +2 neurons, respectively. The number of trials of each condition is provided, vertical lines indicate average blaster-firing latencies (respective numerical values being provided in plots' top-right textboxes) and horizontal lines indicate the threshold to the binding pool at 0.35.

We provide a schematic illustration of how the loss of responsiveness of pre-target intruder-neurons is not only helpful in this configuration of the 2f-ST^2 model, but presumably necessary for our model to replicate the counterintuitive RT patterns. Figure 47 displays two hypotheses of how intrusion errors might occur in each of its columns. Each row of Figure 47 depicts how different responses are provided according to the two hypotheses. In these plots, curved and coloured lines represent item-representations along the response dimension (e.g., response features and response pathway types in the 2f-ST^2 model). Light green, green, black, red, and light red colours correspond to -2 intruder-, -1 intruder-, target-, +1 intruder, and +2 intruder-items, respectively. The encoding threshold, as proposed by Zivony and Eimer (2020b) and as modelled in the 2f-ST^2 model as the connection between response TFL/ type layer and the binding pool, is represented as a blue dashed line. Time-intervals of transient attentional enhancement (TAE, the blaster in 2f-ST^2 model) are depicted as yellow areas. We stress that item-representations specifically refer to response features (or response types in 2f-ST^2 model terms) and that the time-interval of TAE (or the blaster-firing) is the only aspect of the key feature (or key pathway) present in these plots. A key concept in this illustration is that items need to cross the encoding threshold to be encoded into WM.

Hypothesis 1 (left column of Figure 47) displays probably the simplest and most likely explanation of how intrusion errors occur. According to Hypothesis 1, intruders' (response) item-representations cross the encoding threshold governed by their display position (i.e., temporal position in the RSVP stream). Therefore, which item is encoded into WM is only a matter of *when* the encoding process itself takes place. Importantly, this presumably still involves TAE-mechanisms. This concept has similarities to the sophisticated guessing route of Botella et al.'s (2001) model, with the critical time, t_c, determining which response feature would be bound. However, such dynamics would not lead to the RT patterns observed by Botella (1992). Instead, RT patterns reflecting the order of item-representations would be the result: RT (pre-target intrusions) < RT (correct) < RT (post-target intrusions). The counter-intuitive RT patterns observed in humans (Botella, 1992) in addition to the response TFL/ type layer's VMs shown in Figure 46, however, strongly support the second hypothesis illustrated in Figure 47 (right column).

According to Hypothesis 2, the range of TAE-timings is centred around the correct item. Note that we based the TAE-intervals shown in Figure 47's right column on the average blaster-firing latencies presented in Figure 46. An important aspect of Figure 46's blaster-firing latencies is that those of correct trials on average occur earlier than those of pre-target

intrusions. As explained in the previous chapter, this is because in correct trials, the target key type is generally stronger and thus activates the blaster circuit earlier. Hypothesis 2 proposes that pre-target intrusions are associated with late TAE deployment, when pre-target intruder item's representations have already started to decrease in magnitude. This can be seen in the top two rows of Figure 47's right column, in which both green curves have already started to decline just before the time-interval of TAE (yellow box). This is in contrast to how correct reports are provided according to Hypothesis 2 (third row of Figure 47's right column). TAE impacts the target item (black curve) in correct trials at a point in time at which its item representation has been increasing in magnitude. Therefore, less assistance is required from TAE to cross the encoding threshold in this case, which leads to the target item in correct trials being encoded faster than pre-target intruder items in pre-target intrusion trials. This is the case despite the target item itself being presented after pre-target intruder items in the RSVP stream. This latency-difference in encoding is illustrated by the time-interval at which the black and green curves cross the 122ncodeng threshold in the first three rows of Figure 47's right column. The last two rows of Figure 47's right column illustrate how post-target intrusions are made according to Hypothesis 2. +1-intrusions are made because TAE impacts processes at a time interval at which the target-item's representation has been decreasing, but the +1-intruder's representation has been increasing, leading to the +1-intruder to win the competition and be encoded into WM. +2-intrusions are made similarly, with +2-intruder now winning the competition against the +1-intruder, with the latter decreasing in representation magnitude at the time interval of TAE.

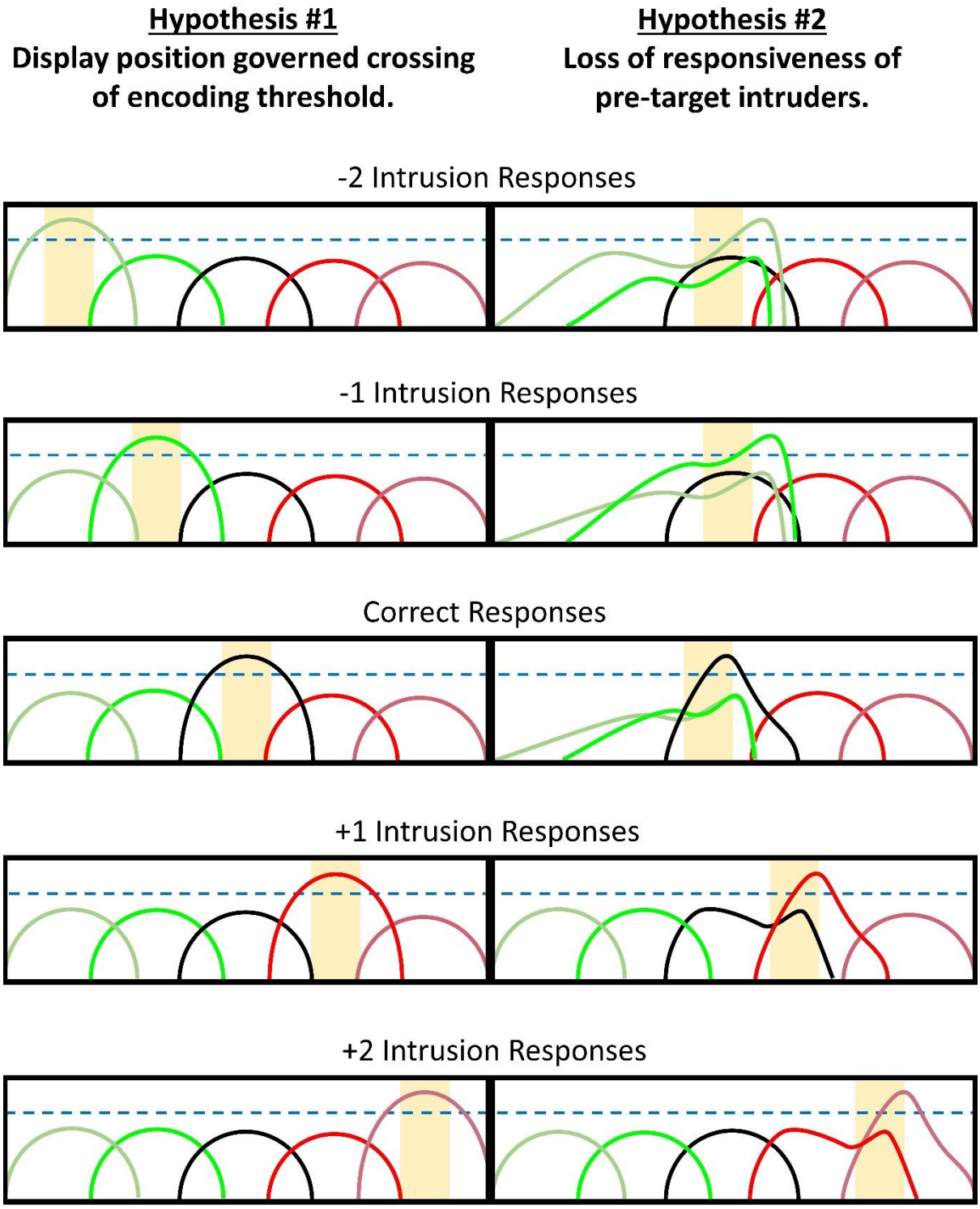


Figure 47. Schematic outline of two possible scenarios (hypotheses) of intrusion errors occurring in human cognition. Light green, green, black, red, and light red curves indicate item-representations of -2, -1, 0 (target), +1, and +2 items, respectively. The blue dashed lines indicate the encoding threshold, which item-representations have to cross to be encoded into WM and subsequently provided as the response. Yellow rectangles indicate transient attentional enhancement (TAE). Note that item representations specifically refer to items' response dimension and that time-intervals of TAE are the only key-feature related element in these plots. Rows illustrate how -2 to +2 intrusions as well as correct responses are generated according to two hypotheses, each positing different item representation dynamics. Hypothesis 1 (left column) states that item-representations (coloured curves) rise and fall sequentially and that the time-interval of encoding is governed by items' temporal display positions in the RSVP stream. The second hypothesis (right column) states that the TAE's timing is centred around the target-item's representation. According to Hypothesis 2, item-representations of pre-target intruders in pre-target intrusion trials (top two rows) cross the encoding threshold later than target items in correct trials (third row), which in turn leads to the counterintuitive RT patterns shown by Botella (1992). Time intervals of TAE deployment depicted for the second hypothesis are based on the blaster-firing latencies found with the 2f-ST^2 model (Figure 46).

Importantly, and as shown in the 2f-ST^2 model's response TFL/ type layer VMs in Figure 46, responses of pre-target intrusion errors are delayed because the TAE's assistance impacts item representations late. As just discussed, late in this context specifically means a point in time at which pre-target intruders' item-representations were already decaying in magnitude, hence requiring more time to cross the encoding threshold. We call this the loss of responsiveness. Crucially, item-representations require the TAE's assistance *in both hypotheses.* To understand the interplay between TAE and (response) item-representations posited by Hypothesis 1, it is useful to consider whether key and response feature processing occurs synchronously or not. A straightforward way of understanding, for example, -2 intrusions according to Hypothesis 1 is that the target's key feature was detected *much* earlier than its response feature. Concretely, detection of the target's key feature would occur so early that the -2 intruder's response feature (highly active light-green item-representation in Figure 46's top left panel) is most active, leading to a -2-intrusion response. This logic would extrapolate to other types of intrusions, too. For example, a +2-intrusion response would be given because the target's key feature was detected with such a substantial time-delay that the +2 intruder's response feature was most active (compare Figure 46's bottom left panel). Such divergent latencies of key and response feature processing would most likely emerge out of a system in which the two features are processed asynchronously. Such an asynchrony between pathways (of feature processing) could, for example, exist if there is a substantial source of temporal variability with which key, but not response features are processed. In theory, asynchronous key and response feature processing could be the case in a computational model such as the 2f-ST^2 model as well as human cognition.

However, the 2f-ST^2 model in its current form works synchronously and, thus, according to Figure 47's Hypothesis 2. This is the case because τD, the gaussian distributed random temporal noise added to item presentation lengths, is the only source of temporal variability present in the model. Therefore, τD is the only source of temporal variability affecting blaster firing latencies. Note that a slight source of asynchrony between the 2f-ST^2 model's key and response pathway exists due to variable input stimulus strengths. However, we argue that this asynchrony is negligible for the model's overall behaviour in this context. We further stress that the 2f-ST^2 model *could* theoretically be configured asynchronously, too, configuring, for example, τR to be fixed, but τK to vary randomly across trials. τK varying randomly would lead to blaster firing latencies varying accordingly and, importantly, independently of the response pathway. This would resemble the display position governed behaviour of TAE deployment shown in Figure 47's Hypothesis 1. However, because items in

this case would cross the encoding threshold according to their display positions, the 2f-ST^2 model configured in this way would not replicate Botella's (1992) RT findings any longer. Instead, RTs would reflect items' display positions, too, and thus, pre-target intrusion errors would be made faster than correct reports. We provide further thoughts about the temporal variability underlying blaster firing latencies being implemented via τD or via a separate mechanism at the end of this chapter.

These findings therefore do not only provide support in favour of Hypothesis 1 of this thesis, i.e., that the binding of multi-dimensional stimulus features into a single percept depends on the timing of TAE, they also emphasise the importance of whether key and response feature processing occurs synchronously. A discussion of the fact that the same loss of responsiveness phenomenon in this case allows the 2f-ST^2 model to replicate human behaviour accurately, but in the case of active task filtering in the response TFL/ type layer and a large τK value leads to rather diminished latency differences between correct and +1-intrusion vP3s will be provided next.

Discussion - Predictions and directions for future research

The 2f-ST^2 model implies a set of intriguing predictions for human cognition, especially due to the loss of responsiveness dynamics introduced above. To summarise, when configured with active task-filtering in the response TFL/ type layer, the large key delay allows enough time for a competition between the target and +1-intruder neuron to emerge. This competition has the effect that winners emerge victorious usually by a small margin only, making the trials that lead to a correct or an intrusion response very similar in terms of input stimulus strengths and blaster-firing latencies. Crucially, *this same phenomenon* is responsible for replicating the RT patterns of Botella (1992), shown in Figure 23. In this context, the fact that the pre-target intruder requires more assistance from the blaster to cross the effective threshold to the response binding pool (or, in more theoretical terms, the encoding threshold (Zivony & Eimer, 2020b)), leads to pre-target intrusions occurring *later* (in terms of RT) than correct trials, even though pre-target items were presented *earlier* than target-items in the virtual RSVP stream. Hence, the same loss of responsiveness phenomenon allows the accurate replication of a counter-intuitive behavioural (RT) pattern in one configuration of the 2f-ST^2 model, providing a compelling argument for why these patterns were observed in humans, but leads to diminished latency-differences between correct and +1-intruder vERPs in another configuration of the model. Notably, the latter, rather important, result of the loss of responsiveness impacting vERPs was completely driven by the necessity

of replicating the behavioural pattern observed in Zivony and Eimer's (2020a) Experiment 1, requiring us to increase τK substantially, which in turn led to this crucial change to the target-intruder competition introduced above. The loss of responsiveness phenomenon nonetheless entails a couple of interesting predictions for human cognition, which will be discussed next.

For experiments in which only two stimuli in the RSVP stream carry relevant task-information (resembling the 2f-ST^2 model with active task-filtering in the response TFL/ type layer), the following two predictions can be made. Firstly, empirical manipulations that increase key pathway processing-time (make processing slower) should lead to *decreased* latency differences in the human N2pc as well as P3 between correct and intrusion trials. Zivony and Eimer's (2020a) Experiment 2, when compared to their first experiment, however, provided evidence in the opposing direction. The authors sped up key pathway processing in their second experiment, using a more salient target-defining key feature, item-colour instead of an item-surrounding geometrical shape. As expected, when speeding up key-feature processing, the response distribution shifted towards an increase in correct reports compared to Experiment 1. However, the N2pc onset latency-difference between target and intrusion trials *was decreased* in Experiment 2. We evaluate the fact that the 2f-ST^2 model would have predicted increased latency-differences in this experiment as indicating a limitation of our computational model, particularly with respect to how the temporal variability underlying TAE is simulated. This limitation as well as an introduction to a more accurate way of modelling TAE dynamics with the 2f-ST^2 model will be discussed later in detail.

Secondly, intrusion trials should be accompanied with reports of high confidence if key pathway processing is fast due to the experimental paradigm implemented. Vice versa, if key pathway processing is slow, people should report low confidence. This is because as the competition between the target- & the +1-response feature gets fiercer, winning neurons come out victorious by an increasingly smaller margin, which in humans should manifest as a loss of certainty/confidence about the correct response. To our awareness, the only study recording confidence ratings in an experiment in which only one target and one intruder item shared their response dimensions was the second experiment of Zivony and Eimer (2020b). Therefore, a comparison study in which key pathway processing was either faster or slower is still lacking. In the context of the last two predictions, the fierceness of the target-intruder competition, which is the main mechanism underlying the loss of responsiveness in the 2f-ST^2 model with active task-filtering in the response TFL/ type layer, depends largely on τK, the delay to key pathway processing. Importantly, the target-intruder competition gets fiercer as

the two respective TFL/ type neurons become increasingly similar in their activation *before the blaster hits* on average.

With respect to the 2f-ST^2 model with no task-filtering in the response TFL/ type layer (replicating the experimental paradigm of Botella and colleagues (e.g. Botella et al. (2001)), a straightforward direction for future research would be to measure the N2pc as well as P3 ERP components. The potential discovery of pre-target intrusion trials being accompanied with *later* ERP components would provide strong evidence in favour of our model architecture and particularly our loss of responsiveness phenomenon generating the RT pattern shown by Botella (1992). The only relevant ERP study in this context was run by Botella and colleagues (Botella et al., 2008) and their results were overall well in line with the 2f-ST^2 model. The authors recorded the P3 component at Pz and showed increased amplitude after correct trials compared to intrusion trials and later P3s for post-target than pre-target trials. Even though these findings clearly fit the 2f-ST^2 model, a study specifically comparing N2pc or P3 latency of pre- and post-target intrusion *as well as* correct trials is still lacking, as Botella et al. (2008) combined the two types of intrusions into a single condition when comparing their P3s to those of correct trials.

The final proposals for future research involve the 2f-ST^2 model itself. First, in its current form, the 2f-ST^2 model does not model the case proposed by Zivony and Eimer (2020b) of both (target and intruder) items being encoded into WM on a subset of trials. As mentioned earlier, this concept was realized in the original ST^2 (Bowman & Wyble, 2007) via the process of joint tokenization, which could also be implemented in the 2f-ST^2 model. Furthermore, in Zivony and Eimer (2020a), non-target digit stimuli were presented in pre-target frames to ensure that attentional resources were deployed due to the key feature and not the categorical shift from letters to digits. In the 2f-ST^2 model, this would mean that in each trial, some types in the response TFL/ type layer meet response pathway task demands prior to the target-type. Types corresponding to virtual non-target digit stimuli would not be affected by the blaster, because their corresponding key-types do not meet task demands, but they would affect other response TFL/ type neurons due to the lateral inhibition implemented in this layer.

Finally, the concept of temporal variability of the TAE underlying intrusion errors is central to theoretical accounts, such as that of Zivony and Eimer (2020a) or Botella and colleagues (Botella et al., 2001), albeit only being relevant for the sophisticated guessing process of the latter. The time-point of the blaster impacting later Stage 1 layers of the 2f-ST^2 model is likewise the main mechanism leading to intrusion errors. We therefore acknowledge

the significance of modelling the TAE's temporal variability as accurately as possible. The 2f-ST^2 model currently manipulates the blaster's temporal variability via τD, the random delay added to each item's presentation length. Importantly, even though this delay is different for different items, it is held constant *across the two pathways* for a given item and trial. This approach keeps the model's two pathways synchronised (although note that a slight asynchrony exists due to different item input strengths between pathways). Applied to human cognition, the implementation of the blaster's temporal variability via τD resembles the case of a very early (in terms of the brain's processing sequence) source of temporal variability that affects key and response feature processing identically. As a result, if our computational model successfully and completely simulates human data, it implies that the rather simple implementation of blaster variability via τD suffices for modelling intrusion errors in human cognition. Moreover, it would concretely suggest that human TAE varies in time between trials due to an early source of variability that affects key and response feature processing similarly. However, if there should be findings that our model fails to replicate accurately, it would be indicative of some aspect of our model being imprecise.

The diminished latency differences between correct and intrusion vP3s when simulating Zivony and Eimer's (2020a) Experiment 1 presented in this chapter suggested that some aspect of the 2f-ST^2 model likely was imprecise. We believe this result to indicate a limitation especially since the 2f-ST^2 model was initially not designed to simulate experiments, such as those of Zivony and Eimer (2020a), in which only two stimuli meet response feature task demands, e.g., being the only two digits in a stream of letters. This is because the 2f-ST^2 model's architecture and parameter space was set up to replicate Botella et al.'s (2001) experiments. It would therefore be conceivable that only adding a response pathway task-filter does not suffice for the accurate and complete replication of Zivony and Eimer's (2020a) empirical paradigm. Note that the 2f-ST^2 model with active response pathway task-filtering generates vERPs that show decreasing latency differences between correct and intrusion responses as τK increases *in general* and that Zivony and Eimer (2020a) found N2pc latency differences between conditions to increase as key features become less salient (resembling larger τKs). Our computational model thus not only struggles to replicate the ERPs of Zivony and Eimer's (2020a) Experiment 1, it also features an interaction between vERP latency differences and τK that is in the opposite direction to that found empirically.

These findings suggest that the implementation of the blaster's temporal variability via τD is the limiting factor in this context. This is suggested by the following two points. First, since key and response feature processing are separate from one another one would expect

that they will be affected by distinct sources of temporal variability. The limitations of our implementation of the blaster's temporal variability via τD is further indicated by the notion that TAE variability should decrease as key features become more salient (i.e., key pathway processing speeds up), which was indicated by Zivony and Eimer's (2020a) N2pcs. In response, we added a mechanism that samples the blaster's latency not from (for example) a gaussian distribution with fixed width, as is implemented in τD at the moment, but instead lets the variance of the distribution from which blaster-latencies would be sampled interact with τK. Importantly, this added source of temporal variability only affects the 2f-ST^2 model's key pathway. This novel way of modelling TAE's temporal variability with the 2f-ST^2 model will be presented in the next chapter.

Conclusion

In this chapter, we chronologically presented our investigation of the loss of responsiveness phenomenon of the 2f-ST^2 model. In doing so, we provided further evidence in favour of Hypothesis 1 of this thesis, i.e., that the binding of multi-dimensional stimulus features into one percept depends on the timing of TAE. In this context, our results suggested an important role of the extent of synchrony with which key and response feature processing occurs. Further, and while not providing additional results in favour or against Hypothesis 2c, stating that the 2f-ST^2 model's vP3 dynamics replicate electrophysiological findings obtained with the human P3 component, we provided analyses that illustrated the impact of the loss of responsiveness phenomenon on the vP3s generated by the 2f-ST^2 model.

In essence, the loss of responsiveness phenomenon occurs whenever the blaster enhances activation levels of later layers at a point in time at which the response TFL/ type neuron that ultimately will be bound to the token has already been decaying in activation strength. Interestingly, this behaviour on the one hand leads to diminished vERP latency differences between correct and +1-intrusion trials if only the two corresponding response types are able to progress to their respective binding pool units due to active task filtering in the model's response pathway to replicate Zivony and Eimer's (2020a) experiments. On the other hand, it is due to the 2f-ST^2 model's loss of responsiveness that the model is able to replicate and account for the counterintuitive RT findings presented by Botella (1992). After the chronological presentation of the loss of responsiveness phenomenon, we finally discussed the phenomenon and additionally provided a set of predictions as well as recommendations for future research.

As stated earlier, the findings and conclusions presented in this chapter motivated us to modify the 2f-ST^2 model with respect to how the temporal variability underlying its TAE mechanism (the blaster) is modelled. The motivation for modifying the model was based on the following three aspects. First, we believe this modification to increase the biophysiological plausibility of our model, since we argue that it is unlikely that the brain's key and response feature processing is affected by the same single source of temporal variability, which τD reflects in the current 2f-ST^2 model. Second, we understood the aforementioned limitations of our computational model when replicating Zivony and Eimer's (2020a) ERP findings to indicate that only basing the blaster's temporal variability on τD does not accurately represent the cognitive processes underlying TAE. Third, and as indicated by Zivony and Eimer's (2020a) N2pc findings, key feature salience should have a direct impact on the extent of temporal variability underlying TAE, with TAE varying less as key features become more salient. In the framework of the 2f-ST^2 model, this would be equivalent to blaster-firing latencies varying less in time as τK decreases.

In the next chapter, we will first probe the hypothesised interaction between key feature salience and the temporal variability underlying TAE empirically before adding a new source of temporal variability with which the blaster fires. We argue that the 2f-ST^2 model presented in the next chapter thus represents its most accurate and complete version.

Chapter 6 – Temporal variability in the key feature pathway: empirical and modelling explorations

Introduction

In Chapter 5 we introduced the loss of responsiveness, a phenomenon observed in the 2f-ST^2 computational model, which reflects different later layer dynamics according to whether the response TFL/ type layer involved active (to replicate Botella et al.'s (2001) paradigm) or inactive (to replicate Zivony and Eimer's (2020a) paradigm) task filtering. Specifically, whereas loss of responsiveness allowed us to qualitatively replicate the counterintuitive reaction time data presented by Botella (1992), it simultaneously led to rather small vERP latency-differences between correct and intrusion trials, when replicating the paradigm of Zivony and Eimer (2020a). As stressed before, we adopted the core architecture of the model previously presented by Chennu et al. (2011), which was explicitly designed to model the experimental paradigm of Botella et al. (2001). The architecture of the initial 2f-ST^2 model contained one particularly influential design choice that limited its efficacy in

replicating Zivony and Eimer's (2020a) experimental paradigm: τD. In the initial 2f-ST^2 model, τD was the only source of latency variability in the model. Even though the 2f-ST^2 model presented in the previous two chapters enabled the implementation of fixed latency differences between key and response pathway processing via τK and τR, respectively, inter trial latency variability, modelled via τD, sat on top of the fixed latency differences between pathways. As a consequence, τD effectively forced key and response pathway processing to be synchronised for a given trial. This synchronous processing of the two pathways can be justified on Occam's razor grounds, i.e., since it means a simpler model. However, cases in which the model's simplicity (particularly the synchrony between the two pathways) is suggested to limit the model's ability to replicate some empirical finding would justify a modification to this aspect of the 2f-ST^2 model. We argue that the diminished vERP latency differences between correct and intrusion trials when replicating Zivony and Eimer's (2020a) Experiment 1, discussed in the last chapter, suggest that the synchrony between key and response pathway processing limits the 2f-ST^2 model in this context. Therefore, we will add a novel source of temporal variability to the key pathway of the 2f-ST^2 model in this chapter. To add to the diminished latency differences between vERPs suggesting the necessity for another source of temporal variability, we will explore an experimental phenomenon using Zivony and Eimer's (2020a) Experiment 1 and 2 N2pc data in this chapter. We will argue that these analyses further suggest that key and response pathway in the 2f-ST^2 model need to be desynchronised.

The experimental phenomenon examined in this chapter to probe whether key and response feature processing occur with distinct sources of temporal variability in the brain, which would suggest the necessity for the 2f-ST^2 model's key and response pathways to be desynchronised, is based on key feature salience interacting with the deployment of transient attentional enhancement (TAE). This analysis concerns the temporal variability with which TAE is deployed when a task-relevant key feature is detected in the RSVP stream. Specifically, TAE should not only be deployed earlier if a more salient key feature is adopted in the RSVP stream (such as experiments using colour rather than shape as the key feature). TAE should further be deployed with less temporal variability between trials if the key feature adopted is very salient. Conceptually, this is because of the floor effect with which very salient key features are detected. In terms of reaction times (RT), there exists a fastest possible RT for the most salient key features, but not a slowest possible RT for the least salient key features. If one would speed up RTs using very salient key features, being easily detected by our attentional system, the distribution of RTs would be compressed towards the fastest RT.

Extrapolating this to the N2pc component, the most salient key features should lead to highly similar N2pc latencies across trials and more dissimilar (i.e., more variable) N2pc latencies for less salient key features. This conceptual notion is supported by the N2pc findings presented by Zivony and Eimer (2020a), which showed not only earlier, but much narrower N2pc components in Experiment 2 (adopting the salient key feature colour) than 1 (adopting the less salient key feature shape). We present comparison plots of the two experiments' N2pc components later on (Figure 73 & Figure 78). To formally assess the notion that TAE is deployed with lower temporal variability if key features are more salient, we empirically analysed the datasets of Zivony and Eimer's (2020a) Experiments 1 and 2.

We further used these analyses to inform us about where in the 2f-ST^2 model's processing pipeline the new source of temporal variability should be added. This is a key issue, since the stage at which the source of temporal variability resides, e.g., being very early (i.e., almost exclusively prior to the separation into two pathways) or later (i.e., within a given pathway), carries distinct implications for the cognitive dynamics leading to intrusion errors. To this end, all analyses examining the interplay between key feature salience and the variability of TAE deployment were conducted separately for two time-windows of the N2pc, capturing the component's onset only (150 – 200 ms) or the full component (150 – 400 ms). Our reasoning is that if increased key feature salience should, for example, only lead to more variable TAE deployment in the N2pc's onset, this would suggest the source of temporal variability of key feature processing to be located early on in the processing pipeline. We will provide a more thorough justification of our choice of time-windows later.

Moreover, and to illustrate how differently salient key features were implemented, we depict the experimental paradigms of Zivony and Eimer's (2020a) Experiments 1 (top panel, identical to Figure 17) and 2 (bottom panel) in Figure 48. Experiment 1's dataset has been analysed in Chapters 4 and 5; the key (target-specifying) feature was an annulus or square surrounding the digit in the target frame. In Experiment 2, participants had to report the first coloured digit. The following post-target distractor stimulus was either a digit or a letter and either presented in grey or coloured (Figure 48's bottom panels A-D depict these scenarios). Zivony and Eimer (2020a) demonstrated a response shift towards correct responses as well as an earlier and sharper N2pc component in Experiment 2 than 1, supporting the notion that the more salient key feature (colour) not only induced earlier TAE deployment, but also that TAE was deployed with less temporal variability.

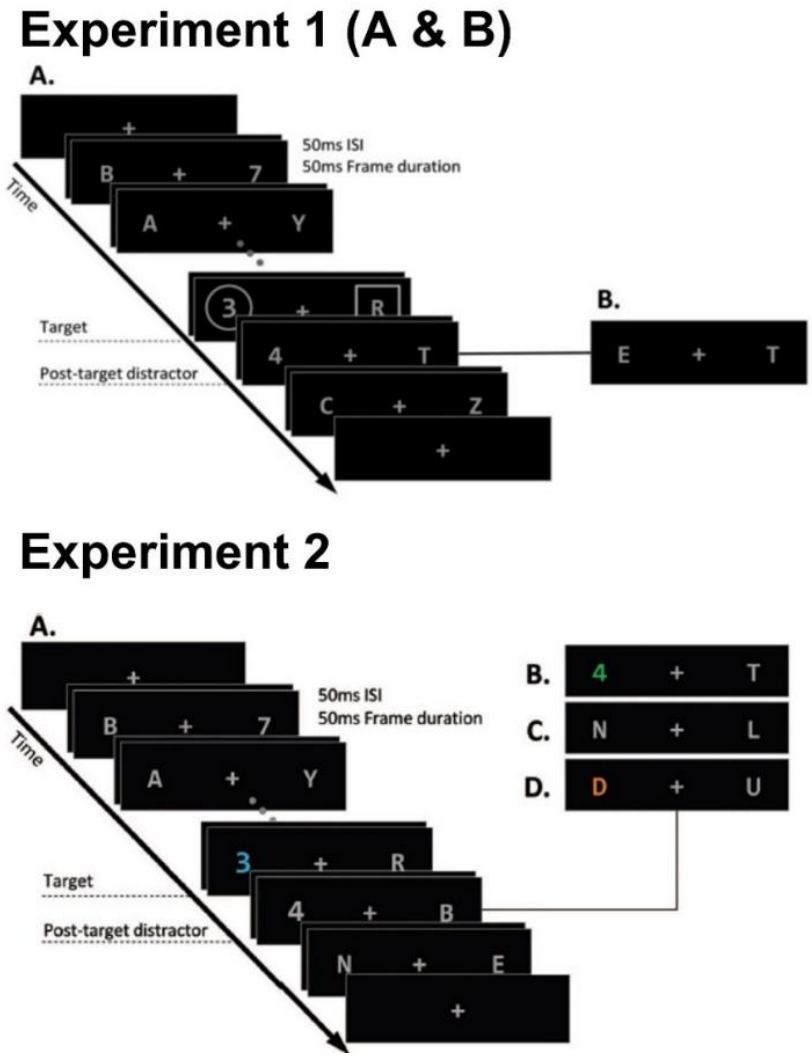


Figure 48. Experimental paradigms of Zivony and Eimer's (2020a) Experiments 1 and 2. Experiment 1, discussed in detail in Chapter 4, adopted an annulus or square surrounding the digit in target frames, whereas in Experiment 2, target stimuli were coloured. Hence, Experiment 2 implemented a more salient key feature, which meant earlier and narrower N2pc components compared to those observed in Experiment 1.

This chapter is dedicated to investigating Hypothesis 3 of this thesis, i.e., that increased key feature salience induces less temporally variable TAE deployment in human cognition, which can further be computationally modelled with the 2f-ST^2 model. It will begin with an empirical analysis to assess whether an increase in key feature salience leads to less temporally variable TAE deployment. To this end, we will present several analyses that contrast the temporal jitter (i.e., latency variability) of N2pcs of Zivony and Eimer's (2020a) Experiments 1 and 2. Thus, the operationalized hypothesis is that N2pc temporal jitter (TJ) should be larger in Experiment 1 than 2. We were further interested in potential TJ differences between correct and intrusion trials. This resulted in a 2x2 interaction design with the factors experiment (1 (low key feature salience) and 2 (high key feature salience)) and response condition (correct and intrusion). A jackknife procedure was adopted to test the interaction

effect, whereas all main and simple effects were analysed with a permutation procedure. In general, permutation procedures were desirable for these analyses, since we used Dynamic Time Warping (DTW) to measure N2pc TJ, which threatened the normality assumptions of parametric approaches, such as ANOVAs. However, as will be explained in detail later on, a permutation approach was not feasible for the interaction effect. Moreover, as previously discussed, we performed the TJ analyses separately for the N2pc's onset time-window as well as for the full component.

We will subsequently probe the second part of Hypothesis 3 and add a mechanism to the 2f-ST^2 model to simulate the interaction between key feature salience and the temporal variability of TAE deployment. This mechanism adds gamma-distributed noise to the latencies with which the blaster enhances later Stage 1 as well as Stage 2 processes in the model. Importantly, the width (i.e., variance) of the gamma distributions is based on the value of τK, the fixed delay to key pathway processing, in a given run of the model. This was implemented in this fashion as τK and the blaster in the 2f-ST^2 model correspond to key feature salience and TAE in experimental paradigms, respectively. Hence, and on the assumption that key feature salience modulates TAE's temporal variability in human cognition, it is sensible that τK modulates the extent of temporal variability of the blaster firing in the 2f-ST^2 model.

In the final part of this chapter, we will present the 2f-ST^2 model's main results after the addition of the modification introduced above. In this context, we will present an additional delay configuration of the 2f-ST^2 model to replicate Zivony and Eimer's (2020a) Experiment 2 and thereby provide further evidence in favour of Hypothesis 2 of this thesis, i.e., that the 2f-ST^2 model accounts for a broad range of findings obtained with distractor intrusion experiments.

Methods

N2pc Temporal Jitter Analyses

To demonstrate the effect proposed in Hypothesis 3 that transient attentional enhancement (TAE) is deployed with less temporal variability as key features become more salient, we conducted a 2x2 interaction analysis with the factors experiment (1 vs 2) and response (correct vs intrusion). Hypothesis 3 particularly refers to the interaction's between-experiment main and simple effects (the latter assessing differences between experiments in correct and intrusion trials respectively). Nonetheless, we analysed the full 2x2 interaction because we were also interested in exploring potential between-condition effects, as such

effects would carry important implications about the neural processes that generate correct or intrusion percepts. Similar to our analyses in Chapter 4, we collapsed the data of Zivony and Eimer's (2020a) Experiments 1A and 1B when testing it against the data of Experiment 2. We measured N2pc TJ using dynamic time warping (DTW) and analysed the interaction effect adopting a jackknife procedure before fitting a repeated measures ANOVA model to N2pc TJ values. Main and simple effects were tested via permutation procedures. We performed all TJ analyses separately for the N2pc onset (defined as 150-200 ms) and the full component (defined as 150-400 ms, compare Figure 49).

We determined the two time-intervals of interest using the fully flattened average (FFA) plotted in Figure 49. For the FFA, we initially concatenated all single trial N2pcs. Specifically, this meant that the N2pcs across different participants, response conditions (correct & intrusion) and experiments were combined into a large (11462 (trials) x 450 (timepoints)) two-dimensional array. Taking the average wave across all (11462) trials then yields the FFA. The FFA was presented in detail by Bowman, Brooks, Hajilou, Zoumpoulaki and Litvak (2020), who advocated its value as an unbiased tool for determining analysis parameters, such as regions-of-interest (ROIs). Importantly, the FFA is blind to differences between participants, experimental conditions, or experiments, since these factors are collapsed out before plotting the FFA. This practice prevents an experimenter to adjust their decisions post-hoc according to the contrast of interest. Figure 49 also depicts the two time-intervals of interest that were determined using the FFA: 150-200 ms for the N2pc's onset and 150-400 ms to capture the whole N2pc component.

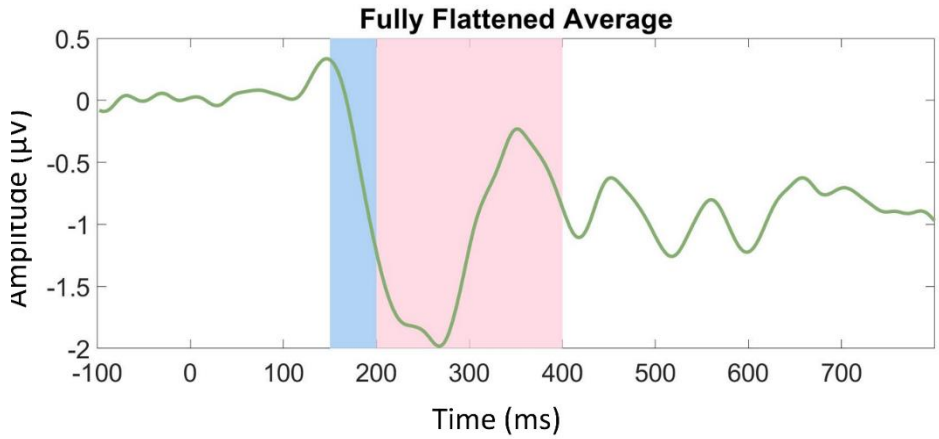


Figure 49. Fully flattened average (FFA) N2pc. The FFA (green line) was computed by first combining subjects of both experiments, then combining all trials across different experimental response conditions, subjects and experiments. We used the FFA to manually choose the two time-intervals of interest, being 150-200 ms and 150-400 ms to capture the N2pc's onset and the whole component, respectively. The N2pc time-window chosen to reflect the component's onset (150 – 200 ms) is depicted as a blue area and the time-window reflecting the full N2pc (150 – 400 ms) is depicted as a blue+pink area.

Reasons for testing the interaction via jackknifing: We had to adopt a different method to test the interaction effect, since the preferred bootstrap permutation procedure adopted for main and simple effects was not feasible for the interaction. The bootstrap permutation procedure is preferred since it is free of parametric assumptions and N2pc TJ, being operationalised as the variance of DTW distance distributions, frequently failed to meet gaussian assumptions (see Chapter 4 for distributions of DTW distances). We initially attempted to develop a permutation procedure for the interaction, assessing the difference of differences in N2pc TJ. However, these proved to be invalid. We assessed the validity of our permutation procedures for the interaction by running simulations in which we controlled the datasets' ground truth effects.

For these simulations, we generated simulated datasets using the following equation: $Y = X + \varepsilon(Experiment) + \varepsilon(Response\ Condition) + \varepsilon(Subject) + \varepsilon(Interaction)$. All variables were sampled from a standard gaussian distribution. X simulated a random intercept which was fixed throughout a given dataset. $\varepsilon(Experiment)$ & $\varepsilon(Response\ Condition)$ were fixed for a given experiment or a given response condition, respectively. Thus, this simulated both main effects to be non-null (e.g., there being a non-zero difference between experiments in the dependent variable Y). $\varepsilon(Subject)$ was fixed for a given subject (we simulated Experiments 1 and 2 to consist of 23 and 12 subjects, respectively, equalling the sample sizes of our EEG datasets). $\varepsilon(Interaction)$ was fully random, i.e., being sampled from a standard gaussian distribution for each entry of Y anew. Thus, in a statistical sense, there was no interaction effect in these ground truth data sets, i.e., the interaction was (statistically) null. Therefore, a valid statistical test on the interaction effect should yield a uniform distribution of p-values across the 50000 data sets generated. However, our permutation procedure generated p-values under the null that were highly inflated (i.e., strongly skewed towards low p-values).

This finding was due to permutation tests for interactions not being available in general (Anderson & Robinson, 2001). This is the case since permuting an outcome variable within levels of (e.g., two) factors does not remove the interaction effect, which would be required for generating a distribution of permuted interaction effects under the null against which the true observed interaction effect is tested. A valid approach for testing interaction effects was proposed by Buzkova (2016), which permutes the residuals after fitting a statistical model (e.g., an ANOVA) lacking an interaction term to the data (containing the intercept as well as, e.g., both main effects). The same (e.g., ANOVA) model *after adding the*

interaction term is then fitted to the residuals, with yields a distribution of the test statistic (e.g., F) under the null. One finally fits the full model (e.g., ANOVA with main as well as interaction terms) to the "observed" dataset and tests the obtained true observed F-value against the permutation distribution of F-values under the null. We attempted to adopt this procedure for our interaction analysis of N2pc TJ, also, fitting linear mixed effects models to obtain the residuals required for permuting the null. Even though this procedure was found to not inflate the false-positive rate using simulations under the null, it was not applicable to our N2pc TJ data. In essence, the challenge with our N2pc TJ analyses is due to subjects' single-trial EEG data being too noisy to be used. Therefore, we were required to perform our TJ analyses on the subject-level average N2pcs. The bootstrap permutation procedure we developed to this end is a valid approach for main and simple effects because we can ensure exchangeability for these analyses. Essentially, these bootstrap tests just adopt the permutation equivalents of independent and dependent sample t-tests for testing the across- and within-subjects effects of experiment (1 vs. 2) and response condition (correct vs. intrusion), respectively. Assessing the interaction effect required a value of the outcome variable (N2pc TJ) for each subject in each response condition (i.e., each bin of the 2x2 interaction) to which a statistical model could be fit. We obtained these values in our interaction analysis, presented next, via jackknifing and DTW. Importantly, jackknifing N2pcs meant that any model that is fit to our N2pc TJ values will contain correlated residuals. This issue is accounted for by Ulrich and Miller's (2001) correction, which will be applied to the F-values obtained with the model. In theory, since we will compute a N2pc TJ value for each subject in each response condition, we could have run a permutation procedure instead. However, we would need to correct for correlated residuals due to the jackknife in such an analysis as well. The development of an appropriate correction in a permutation context would have constituted a rather complicated task. We therefore decided against a permutation procedure for the interaction effect, especially since Ulrich and Miller's (2001) correction has been proven to be accurate.

2x2 Interaction Jackknife DTW Analysis

In the usual way, the jackknife procedure is used here to enable latencies to be reliably quantified. This is typically done by calculating latencies on jackknifed averages, which are constructed from the ERPs of all participants apart from one. The difference here is that, since we are quantifying temporal jitter, these jackknifed collections of ERPs are subject to a DTW procedure that enables us to calculate the latency variability for each jackknifed collection.

Our 2x2 interaction jackknife DTW approach is illustrated in Figure 50. We first computed single-subject N2pc ERPs by averaging across trials for each subject in each response condition separately. For each experiment and each response condition (i.e., correct & intrusion trials), we first removed a given participant's N2pc (Figure 50.1), to give a jackknifed collection. Note that we implemented an unbalanced procedure, as Experiments 1 and 2 involved the N2pcs of 23 and 12 participants, respectively. This discrepancy in sample sizes was considered in our statistical analyses, explained in detail later. The N2pcs of a given jackknife sample were then averaged to generate a jackknife (n-1) Grand Average (GA) N2pc (Figure 50.3). Subsequently, and adopting the exact methodology that was introduced in detail in Chapter 4, the DTW area-difference (i.e., the standardized area between the DTW's warping path and its main diagonal, henceforth called DTW distance) between each subject-level N2pc in a given jackknife sample and that jackknife sample's n-1 GA N2pc was computed (Figure 50.4). Repeating the last step for all subject-level N2pcs of a given jackknife sample yields a distribution of DTW distance values (Figure 50.5). The variance of this distribution was operationalised as the TJ of the *excluded subject's* N2pc (Figure 50.6). The variance of DTW distance values measures TJ because DTW distances indicate latency differences between the jackknife n-1 GA N2pc & subject-level N2pcs. Therefore, the variance of a jackknife sample's distribution of DTW distances indicates the variability of subject-level N2pc latency differences with respect to the jackknife n-1 GA N2pc (i.e., the temporal jitter). For example, if subject-level N2pcs of a jackknife sample unfold with temporal characteristics similar to that of the jackknife n-1 GA N2pc, the variance of their DTW distance values will be small (i.e., less TJ).

We consider the computation of the TJ of the N2pc of Experiment 1's Subject 3 in correct trials to provide an example of this procedure. It was computed by first excluding the corresponding (i.e., Subject 3's N2pc in correct trials) N2pc from the N2pc pool (i.e., all subject-level N2pcs of Experiment 1 for correct responses). The remaining subjects' (i.e., Subjects 1-2 & 4-23) N2pcs in correct trials were separately dynamic time warped against the n-1 GA N2pc, the latter being the Grand Average of the current jackknife sample's N2pcs (i.e., Experiment 1 & correct trials, except Subject 3's). This procedure (Figure 50.1-6) was performed for each subject's N2pc in each response condition separately (Figure 50.7). This provided a value of N2pc TJ for all subjects (each one being excluded once) in each bin of the 2x2 interaction. We ran a repeated measures ANOVA on these TJ values, implementing one across-subject (experiment) and one within-subject (response condition) factor (Figure 50.8). To stress, we conducted this DTW analysis to generate a measure of N2pc TJ on *jackknife*

N2pcs, because they exhibit a much better signal-to-noise ratio than, for example, single-trial N2pcs, which is particularly important for quantifying latencies (and, thus, temporal jitter).

Moreover, we explored an additional approach to computing N2pc TJ for the interaction effect. This analysis, presented in Appendix D, adopted a bootstrap procedure for the computation of N2pc TJ. Specifically, for each subject in each response condition, the respective jackknife collection was taken (i.e., Figure 50.1-2) and the jackknife n-1 GA N2pc was generated (Figure 50.3). A bootstrap sample of the jackknife collection's N2pcs was taken and the Grand Average N2pc of these bootstrap N2pcs was computed (i.e., the bootstrap n-1 GA N2pc). The DTW distance between the jackknife and the bootstrap n-1 GA N2pcs was computed next and stored in a distribution. This procedure was repeated a large number of times, yielding a distribution of DTW distances. The variance of this distribution was again operationalised as N2pc TJ, a repeated measures ANOVA was performed and the F-values were corrected according to Ulrich and Miller (2001). We ran this analysis as well since there is an argument to be made for DTW to be more appropriate if conducted on two Grand Average ERPs instead of one Grand Average and one subject-level ERP. However, because the results did not change substantially with this alternative N2pc TJ measure we decided to abstain from bootstrapping in this context for the sake of simplicity.

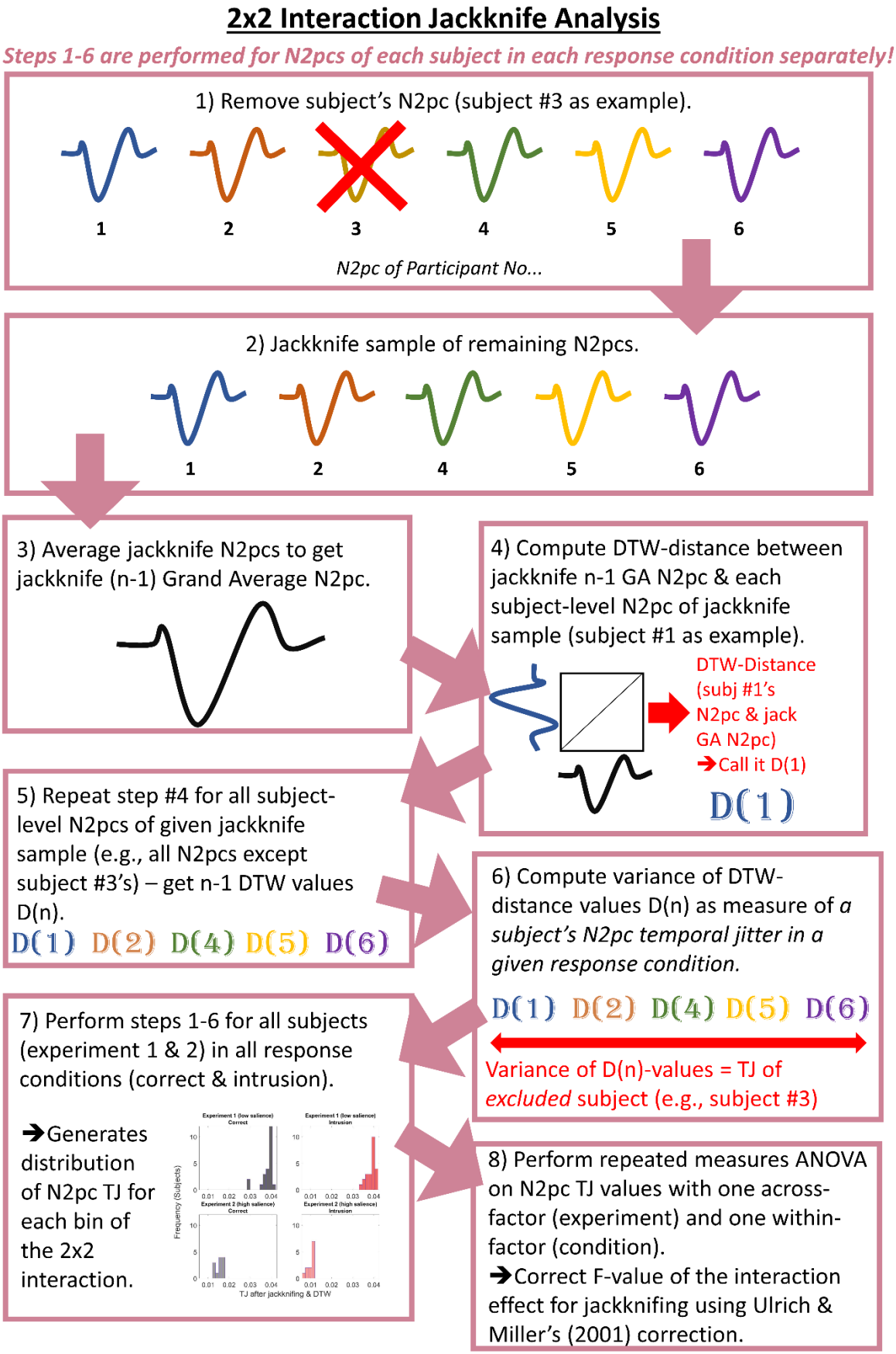


Figure 50. N2pc TJ 2x2 Interaction Jackknife procedure. Steps 1-6 are performed for each participant in each response condition separately. First, a respective participant's N2pc is removed from the sample, to give a jackknife collection (1 & 2). The jackknife n-1 GA N2pc is computed next (3). The DTW distance of each subject-level N2pc in the jackknife sample to the jackknife n-1 GA N2pc is computed for all subject-level N2pcs in the jackknife sample separately (4). Performing Step 4 for all subject-level N2pcs in a given jackknife sample gives n-1 DTW distance values D(n) (5). The variance of Step 5's n-1 D(n) values is computed and used as a measure of a subject's N2pc TJ in a given response condition (6). Performing Steps 1-6 for all subjects in both response conditions separately generates a distribution of N2pc TJ for each bin of the 2x2 interaction (7). Finally, a repeated measures ANOVA is performed on the N2pc TJ values obtained in Step 7 and the interaction effect is tested (8). Note that we correct the interaction's F-value for jackknifing using Ulrich and Miller's (2001) correction.

Importantly, because the jackknife-procedure artificially inflates the degrees of freedom, we corrected the resulting F-value according to Ulrich and Miller (2001), who recommended an adjustment of the within-group error variance for unequal designs (i.e., different sample-sizes between groups), as was the case in our analysis. We therefore multiplied the within-group mean square error (MSE) by both group's sample size minus one (n-1), i.e.: $MSE\ (corrected) = MSE(uncorrected) * 22 * 11$. The corrected MSE value was then used to compute F: $F = \frac{MSG}{MSE(corrected)}$. MSG in this context refers to the interaction's effect. The corrected F-value was then used conventionally to determine its corresponding p-value using MATLAB's fcdf function. Even though the latency measures differed for our TJ analysis, this correction was conceptually identical to that used by Zivony and Eimer (2020a) to determine statistical significance of N2pc onset-latency differences between correct and intrusion trials.

Finally, we ran a simulation analysis to assess the validity of Ulrich and Miller's (2001) correction when applied to variance values of jackknifed datasets. These simulations were similar to those presented earlier that showed the invalidity of our bootstrap permutation approach for testing the interaction effect. For these datasets, we again implemented the two main effects, subject-level variability and no interaction effect. The null hypothesis was therefore true for all simulated datasets' interaction effect. After generating a given dataset, we jackknifed as explained above for each bin of the 2x2 interaction separately. For a given bin (e.g., correct trials in Experiment 1), the jackknifing resulted in n vectors of length n-1, with n equalling 23 and 12 for our simulated Experiments 1 and 2, respectively. The variance of a given jackknife sample was, identical to the procedure displayed in Figure 50, operationalised as an excluded subject's measure of interest (being N2pc TJ in Figure 50 and the variance of numbers generated in these simulations). We again performed repeated measures ANOVA with the between-subjects factor of experiment (1 & 2) and the within-subjects factor of response condition (correct & intrusion) to the variance values, before correcting the interaction effect's F values for jackknifing according to Ulrich and Miller (2001). Corresponding p-values were again obtained in a conventional manner using MATLAB's fcdf function. This procedure was repeated 50000 times, yielding a distribution of corrected p-values under the null. Figure 51 displays this distribution of corrected p-values. A valid statistical test should yield a uniform distribution of p-values under the null, since in the absence of an effect, a given value of p in a given run is completely due to chance and, therefore, as likely as any other p-value. Figure 51 demonstrates an approximately uniform

distribution, suggesting the validity of Ulrich and Miller's (2001) correction for our approach. We acknowledge that the distribution of p-values presented in Figure 51 demonstrates a slightly decreased likelihood of obtaining small p-values, indicating that Ulrich and Miller's (2001) correction is slightly conservative (i.e., decreasing F-values too much) when adopted in this context. We ran the same simulations using the mean of jackknife samples instead of the variance, which yielded a uniform distribution of p-values. Therefore, Ulrich and Miller's (2001) slight over-correction of F-values is likely due to the fact that the dependent variable of interest was variances.

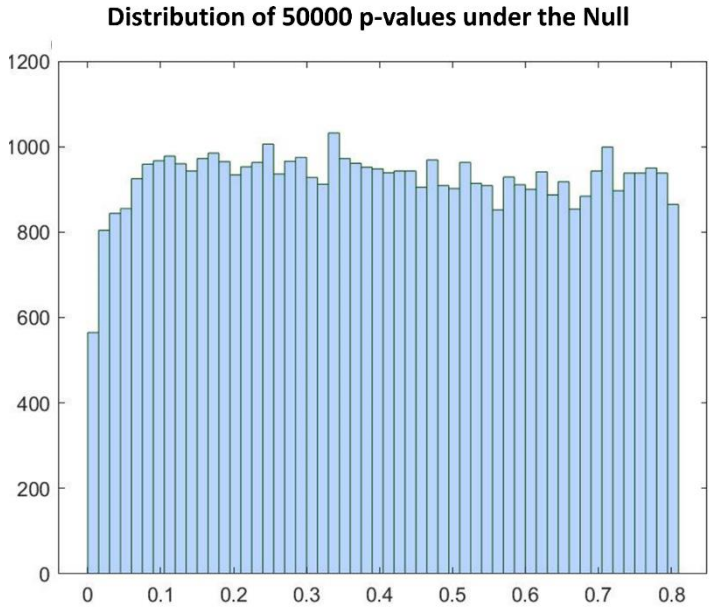


Figure 51. Distribution of p-values after correcting the F-values of repeated measures ANOVAs' interaction effects according to Ulrich and Miller (2001) under the null. 50000 datasets were generated that contained two experiment's data with 23 and 12 simulated subjects each, two main effects, subject-level variability and no interaction effect. Repeated measures ANOVAs were performed on jackknifed variance values, resembling the approach of our 2x2 interaction jackknife analysis. P-values were found to be approximately uniform, indicating Ulrich and Miller's (2001) correction to be valid when applied to our analysis.

DTW bootstrap permutation approach for testing N2pc TJ Main and Simple Effects

We probed both main effects (experiments & response condition) and the four simple effects (experiments for correct & intrusion trials separately and response condition for each experiment separately) with a Dynamic Time Warping (DTW) (Zoumpoulaki et al., 2015) bootstrap permutation analysis. We acknowledge that the standard procedure of this analysis would have been to obtain a measure of N2pc TJ for each subject, but as stressed before, this was not feasible since the trial-level N2pcs were too noisy.

Experiment 1 vs. Experiment 2 (across-subject) main and simple effects: The procedure of this analysis is illustrated in Figure 52 for the case of between-experiment analyses, which were of main interest for this chapter. The analysis commenced with a bootstrap analysis of

the subject-level N2pcs, which yielded a true observed difference in TJ between experiments (Figure 52, top panel). Subsequently, a permutation analysis was conducted to obtain a distribution of TJ differences under the null against which the true observed difference in TJs was statistically tested (Figure 52, bottom panel).

The DTW bootstrap analysis was conducted separately for Experiments 1 and 2. The procedure was identical for both experiments. First, the subject-level N2pcs were averaged to obtain GA N2pcs (henceforth referred to as 'the experiment's GA N2pc', Figure 52.1-2). We then bootstrap sampled the subject-level N2pcs (i.e., with replacement), keeping the same number of subjects of each respective experiment (Figure 52.3). Next, the bootstrap N2pcs were averaged to obtain a bootstrap GA N2pc (Figure 52.4). DTW between the experiment's GA N2pc and the bootstrap GA N2pc was computed next (Figure 52.5). The obtained DTW distance was stored in a distribution and this procedure was repeated 1000 times, yielding one distribution of DTW distances (DTW-distribution) for each experiment (Figure 52.6-7). Each DTW distance indicates how much the bootstrap GA N2pc shifted temporally compared to the experiment's GA N2pc. Furthermore, a *distribution* of DTW distances indicates the *general characteristics* of the underlying subject-level N2pcs that were bootstrap sampled (with reference to the experiment's GA N2pc). This is similar to the reasoning provided above in the context of the 2x2 jackknife interaction procedure. The DTW distance distribution's width indicates the *typical range of temporal shifts* between the bootstrap and the experiment's GA N2pcs in the underlying dataset (in this case being subject-level N2pcs). If a DTW-distribution is very narrow, bootstrap GA N2pcs typically unfold at very similar time-intervals to the experiment's GA N2pc. However, if the DTW-distribution is very wide, bootstrap GA N2pcs vary substantially in time based on the underlying bootstrap sample, at times unfolding much earlier or later than the experiment's GA N2pc. We therefore defined the variance of a DTW-distribution as a measure of N2pc TJ (Figure 52.8). Finally, Experiment 2's DTW-distribution variance was subtracted from Experiment 1's (i.e., $Var\ (Exp\ \#1) - Var\ (Exp\ \#2)$, Figure 52.9). This difference between variances (i.e., TJs) was our true observed measure of interest, which was tested for statistical significance using a permutation procedure next.

For permutation tests in general, one simulates the measure of interest according to the unit of statistical inference under the null, i.e., in the absence of, for example, the experimental manipulations that induced differences along the dimension of interest in the data. Once this null-condition is ensured, usually using random permutations (e.g., via

randomising class-labels) of the dataset, one then needs to repeatedly run the same procedure that yielded the original measure of interest to obtain a distribution of that measure under the null against which the true observed value can be tested. A p-value is then defined in the typical way as the proportion of values in the permutation distribution equal to or more extreme than the observed value. To ensure the aforementioned null-condition, we combined both experiments' datasets and randomly labelled participants as either belonging to Experiment 1 or 2. Importantly, we made sure that the original number of participants (23 and 12 for Experiments 1 and 2, respectively) was maintained in our (surrogate) permutation datasets (Figure 52.10). We then ran the bootstrap analysis as explained in the previous paragraph on the permutation datasets, again using 1000 bootstrap samples for each (permutation) experiment (Figure 52.11). This bootstrap analysis yielded one difference value between (permutation) DTW-distributions' variances (i.e., TJ differences between experiments under the null, Figure 52.12). This procedure was repeated 10000 times and the resulting permutation TJ differences were stored in a distribution (Figure 52.13). Finally, our true observed difference of TJs between Experiments 1 and 2 was tested against this permutation distribution, providing a p-value reflecting the proportion of TJ differences equal to or greater than our true observed value, implying a one-tailed test due to our clear a-priori hypothesis of Experiment 1's N2pc TJ being larger than Experiment 2's (Figure 52.14).

This procedure was performed after collapsing subjects' single-trial EEG data across the two response conditions to test the main effect of experiment and for correct and intrusion trials separately to test the respective simple effects.

Correct vs. intrusion (within-subject) main and simple effects: Finally, the within-subject main and simple effects of response condition (correct vs. intrusion) were tested via a similar bootstrap permutation procedure to the one detailed above. In this context, the main effect was tested after combining all subjects' EEG data from both experiments, effectively collapsing out the main effect of experiment from the analysis. Simple effects were tested by performing the analysis separately for Experiments 1 and 2. The within-subjects effects of response condition were tested as follows.

In the analyses testing the within-subjects effects of response conditions, the bootstrap distributions shown in Figure 52 (e.g., steps 7 & 13) represented correct and intrusion trials instead of Experiments 1 and 2. For example, the left column of Figure 52 would now not correspond to Experiment 1, but to correct trials. Thus, we extracted all participants' correct trials and computed their respective subject-level N2pcs. We then bootstrap sampled these

subject-level N2pcs and took the average wave to obtain the bootstrap GA N2pc (of correct trials), before computing the DTW distance between this bootstrap GA N2pc and the GA N2pc of correct trials. Thus, this yielded a bootstrap distribution of latencies of correct trials. This procedure was repeated for intrusion trials. We again computed the variance of the two distributions and then subtracted them from each other to obtain our true observed difference in N2pc TJs. The two analysis pipelines therefore are identical so far, the only difference being what is being fed into them (correct & intrusion N2pcs instead of N2pcs across experiments). Importantly, the full analysis pipeline procedurally only differs in one step for the test of the within-subjects effect of response condition. The only exception in procedures is the step illustrated in Figure 52.10, the random shuffling of participants' N2pcs into belonging to either of the two experiments.

Randomly shuffling participants' N2pcs into either experiment is the permutation equivalent of a two-sample independent T-test in this context. However, this was not appropriate for tests of the within-subject effects of response condition, since these are equivalent to the case of a two-sample dependent T-test. This is because all subjects have two observations in the dataset, in this case being subject-level N2pcs of correct and intrusion trials. Therefore, we performed the permutation equivalent of a two-sample paired T-test and adopted a condition swapping procedure to simulate the null, similar to the one presented in Chapter 4 in the context of our DTW analysis of correct and intrusion N2pc latencies. See Manly (2018) for a discussion of permutation equivalents of dependent and independent two-sample T-tests. Specifically, for the present DTW N2pc TJ analysis testing differences between response conditions, we randomly (and with equal likelihoods) either kept the real labels of a participant's subject-level N2pcs as representing correct and intrusion trials or we flipped the labels, labelling that participant's correct N2pc as representing intrusion trials and vice versa. This ensured that exactly one subject-level N2pc was in both permutation samples shown in Figure 52.10. Finally, we again adopted a one-tailed significance test for the analysis of N2pc TJ differences between correct and intrusion trials. In this context, we had the a-priori hypothesis that intrusion trials should imply a larger N2pc TJ than correct trials because, as we argue, correct trials tend to involve a clearer percept of the reported stimulus than intrusion trials. This is related to the loss of responsiveness phenomenon presented in Chapter 5: correct stimuli arguably need less assistance from TAE mechanisms, and hence should involve more similar N2pc latencies across trials (i.e., low TJ). On the other hand, intrusions are percepts that are more ambiguous and hence require more assistance from TAE mechanisms (or, in 2f-ST^2 terms, weaker types that depend on the blaster's enhancement

more to excite their corresponding binding pool unit). This means larger temporal variability with which the TAE impacts processing across trials, therefore leading to higher TJ in the N2pcs of those trials.

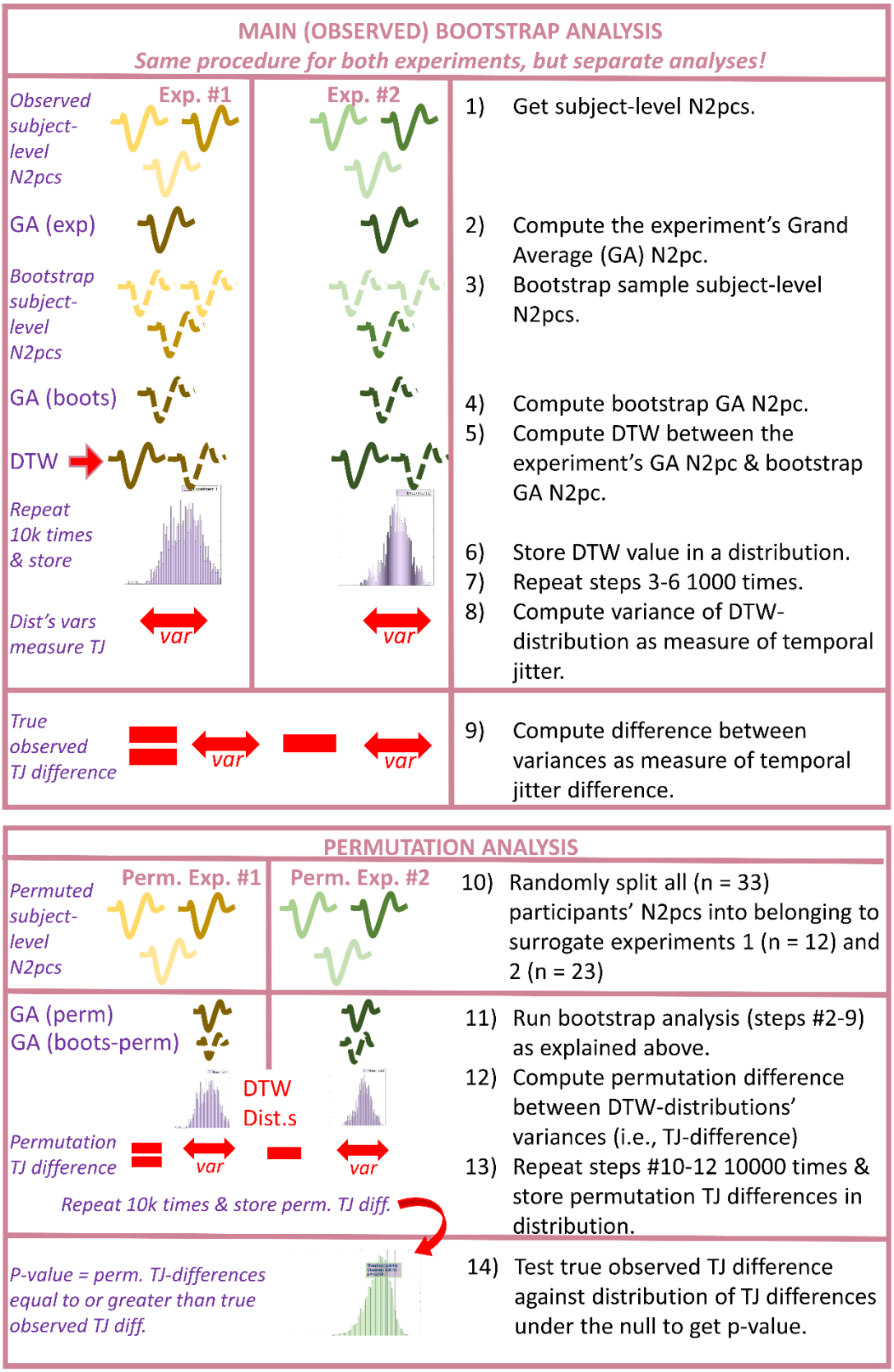


Figure 52. DTW N2pc TJ bootstrap permutation procedure. Top and bottom panels illustrate bootstrap and permutation procedures, respectively. The bootstrap procedure is computed separately and identically for both experiments, commencing with the computation of the GA N2pcs (1-2). Bootstrap samples are computed next and a bootstrap GA N2pc is computed by averaging bootstrap subject-level N2pcs (3-4). The DTW is computed between the experiment's GA N2pc and bootstrap GA N2pcs (5-6), and these steps (3-6) are repeated 1000 times (7), yielding a distribution of DTW distances (7) of which the variance serves as a measure of the N2pc's temporal jitter (8). The difference between experiments' variance values is computed and used as the observed jitter-difference between experiments (9), which is tested for statistical significance using a permutation procedure next (bottom panel, 10-14). For the permutation procedure, the pool of all participants' subject-level N2pcs is randomly assigned to Experiments 1 and 2 (10), before performing the bootstrap analysis as explained before (10-12). This yields a distribution of temporal jitter differences between experiments under the null (13) against which the true observed jitter-difference (9) is tested (14). This procedure was adjusted to test the within-subject effects of response condition with the implementation of a condition-swapping procedure.

Gamma Noise model of TAE deployment

To probe the second part of Hypothesis 3, i.e., that less temporally variable TAE deployment after increased key feature salience can be computationally modelled with the 2f-ST^2 model, we added gamma distributed noise to the 2f-ST^2's key pathway processing. *In each trial,* there existed the possibility of τK (the delay to key pathway processing) to be smaller or larger than its fixed value. This mechanism added a new kind of variability to the temporal dynamics of the model's key pathway processing, which affected the timing of the task-relevant key TFL/ type-neuron crossing the threshold to the blaster circuit. This variability (henceforth called Gamma Noise) propagated to the latencies with which the blaster fired and affected later Stage 1 layers in the key *as well as* the response pathway (the blaster affecting those layers is how TAE-deployment is modelled in the 2f-ST^2 model). The implementation of Gamma Noise further meant control over the extent of the noise added to blaster-firing latencies, which enabled us to accurately simulate the proposed interaction between key-feature salience and the TAE's temporal variability (or jitter).

Configuration of Gamma Noise Distribution & Addition to τK

The Gamma Noise distribution was created *separately for separate runs* of the model. For each trial in a given run, one value of Gamma Noise was sampled from this distribution. We present example distributions in Figure 53, which were generated as follows. For a given distribution (i.e., 'original' distributions, plotted in the left column of Figure 53), a shape and a scale parameter had to be set. The shape parameter controls the distribution's shape, either leading to a distribution that is highly right-skewed (low shape value) or to a normal distribution (high shape value). For our purposes, a low shape value was desired to ensure that most values in the distribution were close to zero and only few large values were present. We argue that a low shape-value is especially crucial in this context because processes in human selective attention are bound to a temporal limit concerning *how quickly* they can occur, but not so much w.r.t. *how slowly* they can occur. As stressed previously in the context of RT's floor effect, in experimental trials in which the TAE is deployed *very early*, the TAE will never be deployed *before previous cognitive processes have been completed* (as illustrated by the N2pc, which indexes selective attention, and typically carries an onset latency of 150-200 ms (Eimer, 1996; Zivony & Eimer, 2020a)). In contrast, in experimental trials in which the TAE is deployed *very late*, it is not bound temporally to any subsequent cognitive processes to the same extent. Hence, late TAE deployment occurs with a wider range of latencies compared to early TAE deployment. For these reasons, a low shape value was desired. Concerning the numerical value of the shape parameter, a valid argument can be made in

favour of the parameter varying based on τK. This argument is based on the speed-accuracy interactions of RTs in experimental paradigms (Draheim et al., 2019). In essence, RTs are distributed more similarly to a gaussian distribution as one instructs participants to respond as accurately as possible. RT distributions become more skewed to the right (i.e., more probability mass for fast RTs) if participants are instructed to respond as fast as possible. We could have used this speed-accuracy interaction of RTs to justify the shape parameter to interact with τK, as low τKs lead to fast RTs (characterised by a low shape value of gamma distributions) and large τKs lead to slow RTs (characterised by higher shape values). However, for the sake of keeping the 2f-ST^2 model and our addition to it as simple as possible, we decided to only vary the scale parameter and keep the shape parameter fixed at 1, which generates gamma distributions that are highly skewed to the right.

Gamma distributions also require a scale parameter. The scale of a distribution controls its width, i.e., the extent to which larger values occur in it. This parameter enabled us to control the extent of variability added to the model's τK and, thereby, the temporal jitter added to the latencies with which the blaster fired. Thus, the scale parameter was based on the value of τK, as the proposition to be simulated was that larger τK values (simulating less salient key-features in experiments) should imply more variability with which the blaster fires (i.e., the TAE is deployed in human cognition). We therefore set the scale to 1 for τK values smaller than 0 and to τK+2 for all other values. For the model-configurations of (main) interest presented in Chapter 4, τKs of 0 and 24, this implied scale values of 2 and 26, respectively. Figure 53 illustrates the impact of different scale values to the distributions, showing the distributions adopted for τKs of 0 (Figure 53, A & B), 9 (Figure 53, C & D) as well as 24 (Figure 53, E & F). The configuration of τK = 9 & τR = 0 was adopted to simulate Zivony and Eimer's (2020a) Experiment 2 and will be introduced in detail later on.

Generating the gamma distributions with the shape and scale parameters explained above led to the distributions plotted in Figure 53, A, C & E for τK values of 0, 9 and 24, respectively (called 'original' distributions, Figure 53, left column). After the generation of a given distribution, we shifted the distribution to ensure that the respective value of τK represented the distribution's median ('final' distributions in Figure 53's right column). To this end, we first computed the original distribution's median and subtracted it from τK. The resulting difference value was then added to the whole distribution, shifting it either to the left if τK < median (compare Figure 53, A & B) or to the right if τK > median (Figure 53, C - D & E-F). We sampled a value of τK for all trials anew from the 'final' distribution shown in

Figure 53's right column. The 'final' distributions consisted of values that were smaller or larger than τK half the time, i.e., it was equally like to sample a τK value that was smaller or larger than the initially set fixed τK for a given trial. For example, if setting a model run to implement a fixed τK value of 24 time-steps, the value of τK adopted for an individual trial would be sampled from the distribution shown in Figure 53F. Therefore, a given trial would be as likely to be run with a τK faster than the original 24 time-steps as it would be to be run with a slower τK. Critically, even though values smaller or larger than τK both made up 50% of our 'final' distributions, the difference of values > τK to τK was generally larger than that of values < τK, simulating the characteristic of the hard limit to fast TAE-deployment described earlier.

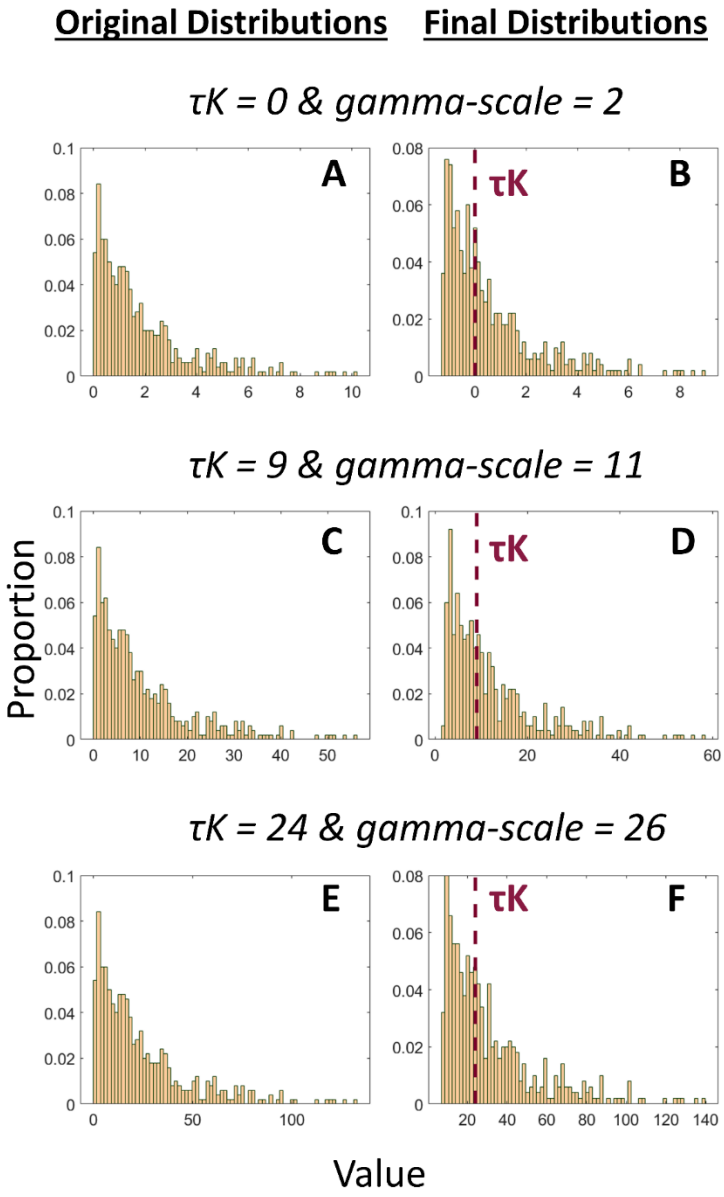


Figure 53. Gamma Noise Distributions. Notice the change in x-axis with row. Left and right columns illustrate original (i.e., output of MATLAB's gamrnd function) and final distributions, respectively. All distributions were generated with a shape parameter of 1, leading to highly right-skewed distributions. The scale parameter, in contrast, differed based on the value of τK, being 1 if τK < 0 and to τK+2 if τK ≥ 0. Final distributions (right column) are the original distributions shifted so that a distribution's median was equal to τK. Top, middle and bottom rows illustrate the distributions for τKs of 0, 9 & 24. For a given run of the model, a final distribution was generated and used for all trials. Importantly, a given trial's value of τK was sampled from the final distributions shown in the figure's right column. We made sure that key-pathway processing is sped up 50% of times and slowed down 50% of times (simulating temporal jitter around the deployment of the TAE in human cognition). We based the scale parameter (and hence the distributions' widths) on τK to simulate the proposed interaction between key-feature salience and the temporal variability underlying human TAE deployment. Note that due to distributions being skewed to the right, τKs were increased more strongly than they were decreased, despite Gamma Noise values being larger or smaller with equal likelihoods of 50%.

Replicating Zivony and Eimer's (2020a) Experiment 2 with a τK of 9 and Gamma Noise

As illustrated in Figure 48, Zivony and Eimer (2020a) incorporated coloured target digits followed by either grey or coloured digits or letters in their second experiment. Compared to Experiment 1, which showed all stimuli in grey, a response shift to an increased proportion of correct trials was observed in Experiment 2 (Zivony & Eimer, 2020a). This shift

in responses was argued to be due to a swifter deployment of TAE thanks to the more salient (coloured) key-features presented in Experiment 2 (Zivony & Eimer, 2020a). We therefore implemented a reduced delay to key-pathway processing (τK) to the 2f-ST^2 model to replicate Zivony and Eimer's (2020a) second experiment. Whereas τK was set to 24 time-steps to replicate Experiment 1, we now adopted a τK value of 9 time-steps. The fixed delay to response-pathway processing, τR, was set to 0 and τD, the randomly sampled delay to both pathways, was set to 4 time-steps in both cases. Furthermore, Gamma Noise was implemented in the 2f-ST^2 model in all delay-configurations presented in the current chapter. We argue that the implementation of Gamma Noise resembles a more complete model compared to that presented in the previous two chapters, as we now account for an additional mechanism proposed in humans: TAE being deployed with modulated temporal variability due to key-feature salience. The results of the latter will therefore constitute further evidence for Hypothesis 2 of this thesis, i.e., that the 2f-ST^2 model accounts for a broad range of findings obtained with distractor intrusion experiments.

Results

N2pc Temporal Jitter (TJ) Analyses

We present the results of our temporal jitter (TJ) analysis on N2pc-latencies to investigate Hypothesis 3, i.e., that transient attentional enhancement (TAE) is deployed with less temporal variability as key features become more salient. This hypothesis particularly refers to the main effect of experiment in our 2x2 interaction design, which employed the factors experiment (1 vs. 2) and response condition (correct vs. intrusion). In this section, we will present the results of the 2x2 interaction jackknife procedure first, before presenting the results of both main effects as well as the four simple effects, which were obtained adopting a bootstrap permutation approach. All analyses of main and simple effects will first be presented for the N2pc onset time-window (150 – 200 ms) and subsequently for the full N2pc time-window (150 – 400 ms).

2x2 Interaction Jackknife

Figure 54 presents the results of our 2x2 interaction jackknife analysis for the N2pc onset (Figure 54A) as well as the whole component (Figure 54B). The top eight panels of Figure 54 display the subject-level distributions of N2pc TJ after jackknifing and computing the variance of a jackknife sample's DTW distances to that jackknife n-1 GA N2pc. Figure 54's top eight panels show the N2pc TJ values of Experiments 1 and 2 in top and bottom rows and the values of correct and intrusion trials in left and right columns. We added an inset to

Experiment 2's N2pc TJ distribution in intrusion trials for the N2pc onset (bottom right quadrant of Figure 54A, top panels), since keeping the x-axes ranges constant across the four distributions of a given time-window meant that only one bin was plotted in the histogram initially. Note that Experiment 1 (top row) and intrusion trials (right column) were hypothesised a-priori to be correlated with more N2pc TJ. Repeated measures ANOVAs were performed on these distributions, which resulted in non-significant p-values of $p = .335$ and $p = .6814$ for the N2pc onset and the full N2pc, respectively. Note that these p-values were obtained after performing Ulrich and Miller's (2001) correction for jackknifing. The bottom two panels illustrate the interaction plots for the two time-windows of interest, displaying the average N2pc TJ values of each bin of the 2x2 interaction. Note that the range of N2pc TJ values differs between time-windows in these plots, with the full N2pc (Figure 54B) demonstrating overall lower values of N2pc TJ compared to those of the N2pc's onset (Figure 54A).

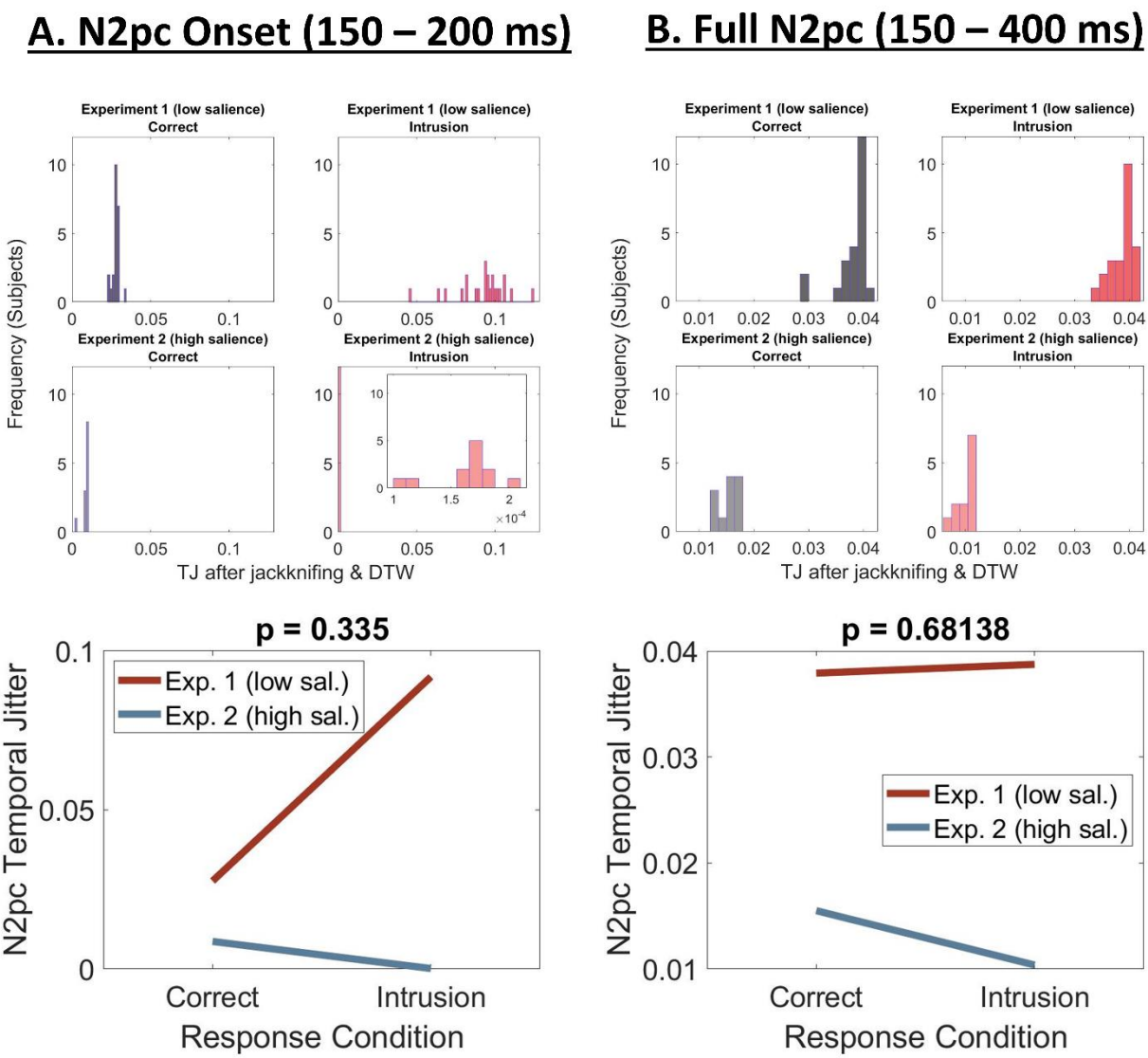


Figure 54. 2x2 Jackknife Interaction. Panels A and B show the results of performing the jackknife interaction analysis for the time-windows of N2pc onset (150 – 200 ms) and the full N2pc (150 – 400 ms), respectively. The top eight panels display the subject-level distributions of N2pc TJ after DTW and jackknifing on which the repeated measures ANOVAs were conducted. We provide an inset to the bottom right quadrant of Panel A's top panels (N2pc onset, Experiment 2, intrusion trials) since the initial histogram of this condition only had one bin due to the x-axis range being too large. The bottom two panels illustrate the interaction plots, depicting the average N2pc TJ values of each bin of the 2x2 interaction. Note that N2pc TJ value-ranges (as in, e.g., the y-axis of the interaction plots) is lower for the full N2pc (B) than for the N2pc's onset (A).

DTW bootstrap permutation approach testing main and simple effects

The results of our DTW N2pc TJ bootstrap permutation analyses are presented next. For these analyses, a bootstrap procedure was employed, which yielded one true observed value of TJ-differences for each effect of interest. Embedding this bootstrap procedure in a permutation loop allowed us to simulate a null-distribution of our N2pc TJ difference of interest and, hence, to probe whether the observed N2pc TJ difference was statistically significant (see Figure 52 for a detailed illustration of our DTW bootstrap-permutation approach). We will present all analyses for the N2pc onset (150 – 200 ms) first, before showing the results of the full N2pc (150 – 400 ms).

Main and simple effects of experiment: Since it is of main interest for Hypothesis 3 of this thesis, i.e., that increased key feature salience induces less temporally variable TAE deployment in human cognition, we will present the results of the between-experiment effects first. Figure 55 & Figure 56 illustrate the results of testing the main effect of experiment for the N2pc's onset and the full N2pc, respectively. The top panel in these figures present each respective experiment's (true observed) bootstrap DTW-distance distributions. Concretely, this means the distributions of DTW distances between bootstrap GA N2pcs and that experiment's GA N2pc (see Figure 52). Similar to our 2x2 interaction jackknife N2pc TJ analyses, we operationalized the variance of the DTW-distance distributions (illustrated by purple arrows in the figures) as our measure of N2pc TJ and hence computed the difference of the two variance values as our measure of statistical interest (i.e., the N2pc TJ difference between experiments). As explained in detail previously (and as illustrated in Figure 52), embedding our bootstrap procedure in a permutation-loop yielded a simulated null-distribution of our measure of statistical interest, which is presented in the bottom panels of Figure 55 and Figure 56. For example, the bottom panel of Figure 55 shows the distribution of 10000 permuted pairs of DTW-distance distributions' variance (i.e., N2pc TJ) differences for the N2pc's onset. Hence, one datapoint of Figure 55's bottom panel distribution corresponds to a pair of null-distributions similar to the true observed distributions plotted in the top panel of Figure 55 and computing the difference between their variances. The one-tailed statistical test yielded a significant result for the N2pc's onset (observed N2pc TJ difference = 0.00167, $p < .0001$) and a non-significant result assessing the full N2pc (observed N2pc TJ difference = 0.00115, $p = .1953$). We acknowledge the striking shape of the permutation distribution shown in Figure 55's bottom plot. This shape is generated due to the unequal sample sizes in the two experiments (23 in Experiment 1 and 12 in Experiment 2), which affects the permutation distributions for the N2pc's onset in particular. We ran the same analysis with equal numbers of subjects in both experiments (i.e., using a random subset of only 12 subjects for Experiment 1). This led to the permutation distribution becoming approximately symmetrical around zero, while the p-value remained highly significant. This same pattern was found to be true for the two simple effects contrasting N2pc TJ between experiments for the N2pc's onset, shown next.

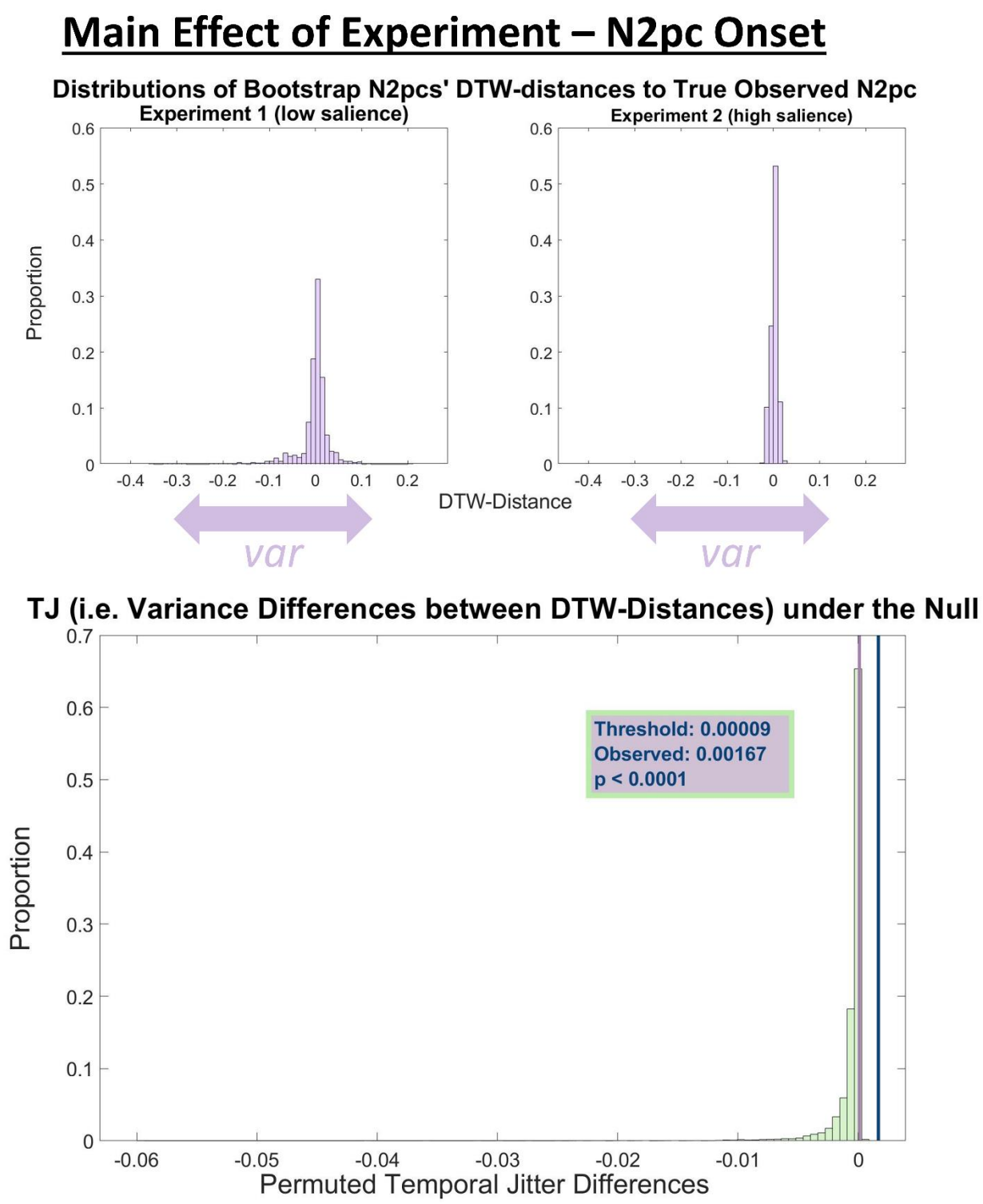


Figure 55. N2pc TJ DTW bootstrap-permutation test of the between-experiment main effect for the N2pc's onset (150 – 200 ms). The top panel illustrates the two true observed distributions of 1000 DTW distance values computed on a given bootstrap versus that experiment's GA N2pc. Left and right distributions in the top panel depict these distributions for Experiments 1 and 2, respectively. The difference in variances between these two distributions was our measure of statistical interest (i.e., N2pc TJ difference between experiments) that was statistically tested via a permutation procedure. The bottom panel shows the simulated null-distribution our permutation procedure yielded. Our bootstrap procedure was embedded into the permutation loop, meaning that in each of the 10000 permutation repetitions, 1000 bootstrap samples were performed. Hence, each data-point of the distribution in the bottom panel represents the difference between variance-values of two permuted (i.e., null) distributions of DTW distances, similar to those plotted in the top panel. The statistical test yielded a significant result, as is indicated by the blue (true observed variance-difference between the distributions plotted in the top panel) and purple (5% significance threshold) lines in the bottom panel. Note that we investigated the striking shape of the bottom panel's distribution and found it to be due to the unequal number of subjects in the two experiments. A rerun of the same procedure adopting the same number of subjects in the two experiments yielded a similarly significant p-value while generating a distribution that was approximately symmetrical around zero.

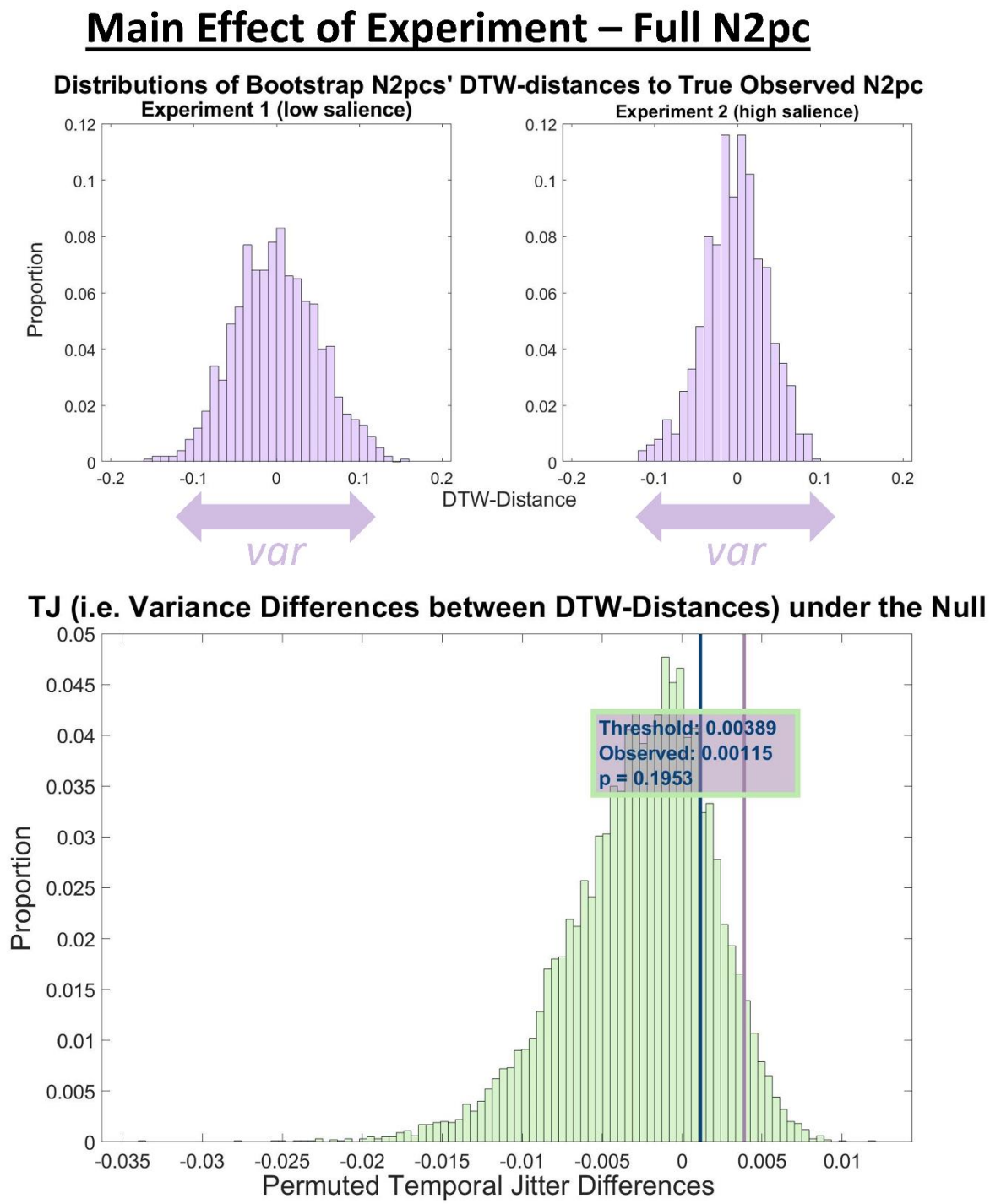


Figure 56. N2pc TJ bootstrap permutation results of the experiment main effect assessing the full N2pc (150 – 400 ms) component. Plotting conventions are identical to those of Figure 55. The main effect of experiment yielded a non-significant finding for the full N2pc.

We tested the between-experiment simple effect, assessing correct trials only, next. These results are presented in Figure 57 & Figure 58, for the N2pc's onset and the full N2pc, respectively, keeping the same plotting conventions as in previous plots (e.g., Figure 55). Similar to the main effect presented before, the simple effect between experiments for correct trials turned out to be statistically significant for the N2pc's onset (observed N2pc TJ

difference = 0.00203, $p < .0001$, compare Figure 57), whereas the equivalent test for the full N2pc component provided a non-significant result (observed N2pc TJ difference = -0.00016, $p = .3418$, compare Figure 58). The results of the between-experiment simple effect for intrusion trials are presented in Figure 59 & Figure 60. These analyses showed the same pattern of significance, with the N2pc's onset yielding a significant result (observed N2pc TJ difference = 0.034, $p < .0001$), whereas the full N2pc's analysis yielded a non-significant p-value (observed N2pc TJ difference = 0.0037, $p = .0943$, compare Figure 60).

Simple Effect of Experiment (Correct) – N2pc Onset

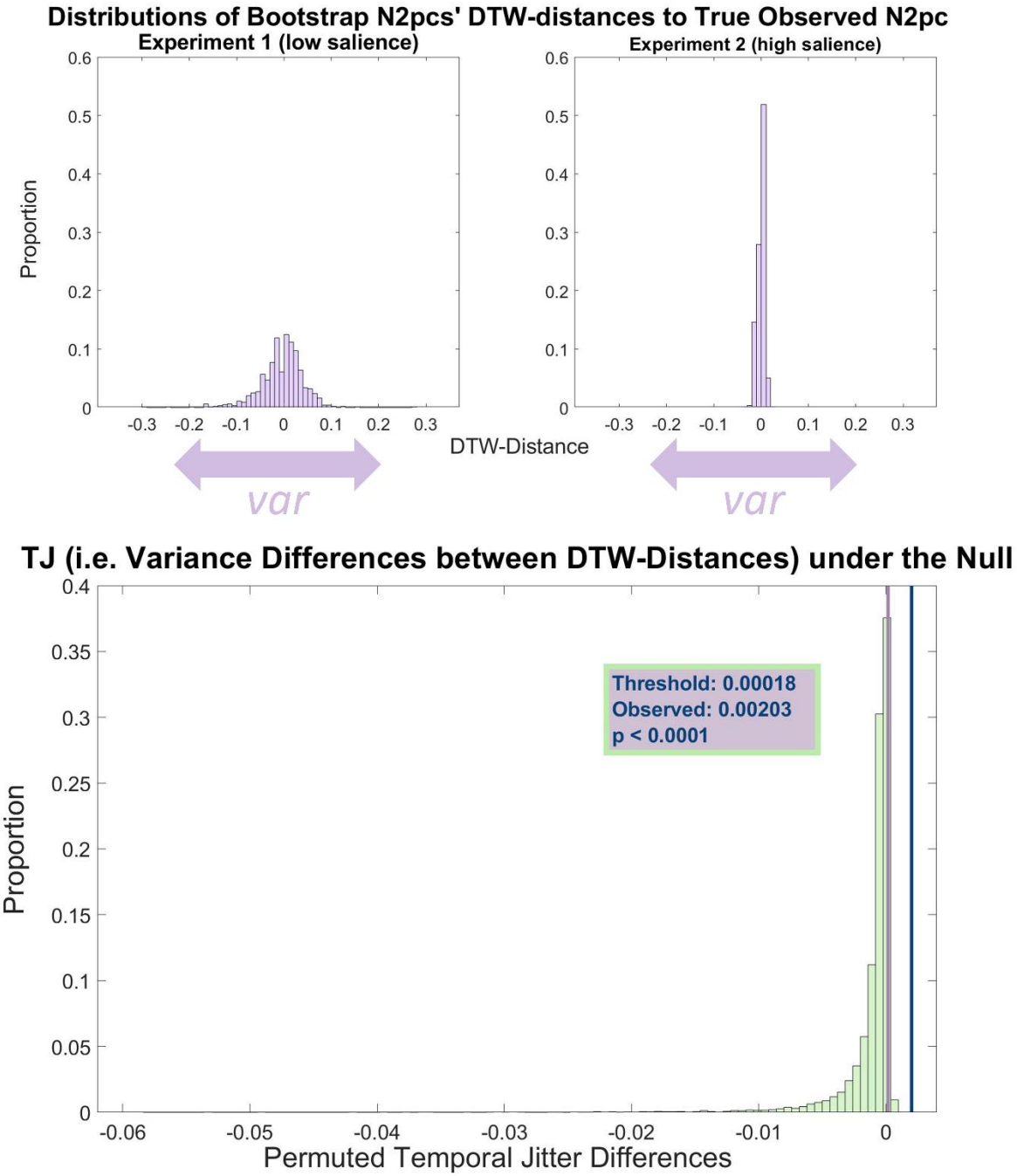


Figure 57. N2pc TJ bootstrap permutation results testing the simple effect of experiment-differences in correct trials and assessing the N2pc's onset (150 – 200 ms). Plotting conventions are identical to those of Figure 55. A statistically significant TJ difference was found between experiments in correct trials for the N2pc's onset.

Simple Effect of Experiment (Correct) – Full N2pc

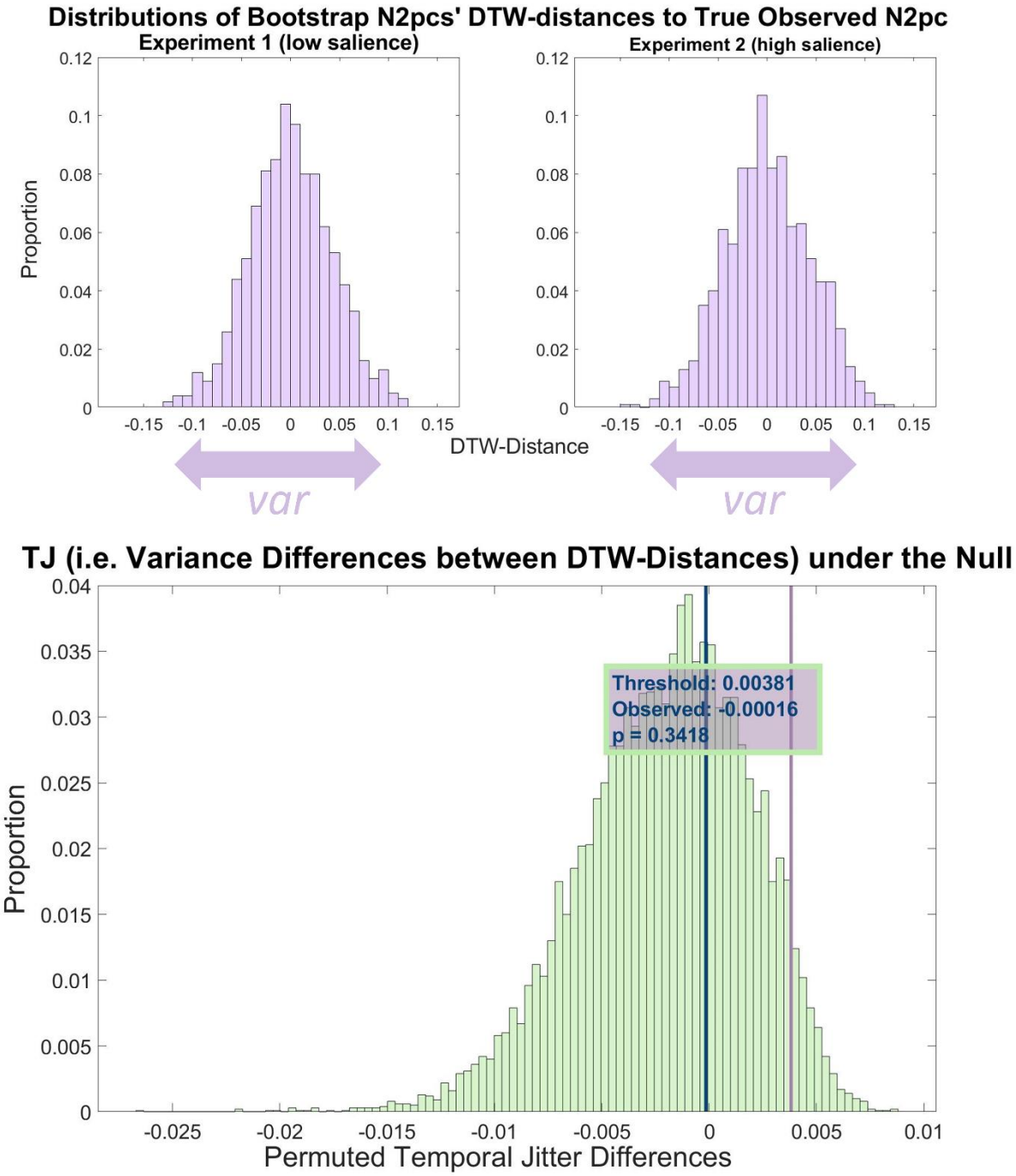


Figure 58. N2pc TJ bootstrap permutation results testing the simple effect of experiment-differences in correct trials and assessing the full N2pc (150 – 400 ms) component. Plotting conventions are identical to those of Figure 55. The TJ difference between experiments in correct trials for the full N2pc component was found to be non-significant.

Simple Effect of Experiment (Intrusion) – N2pc Onset

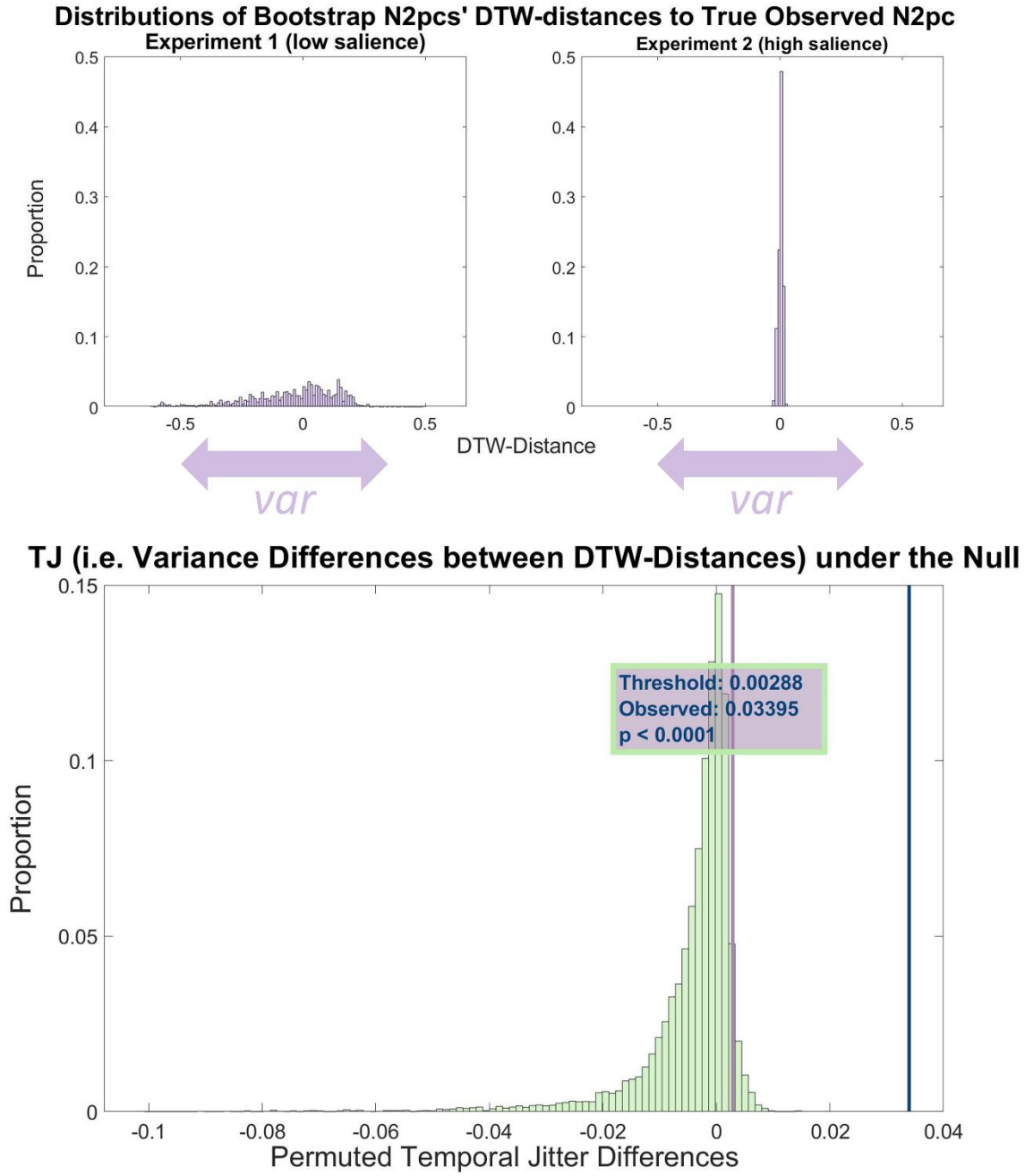


Figure 59. N2pc TJ bootstrap permutation results testing the simple effect of experiment-differences in intrusion trials and assessing the N2pc's onset (150 – 200 ms). Plotting conventions are identical to those of Figure 55. A statistically significant TJ difference was found between experiments in intrusion trials for the N2pc's onset.

Simple Effect of Experiment (Intrusion) – Full N2pc

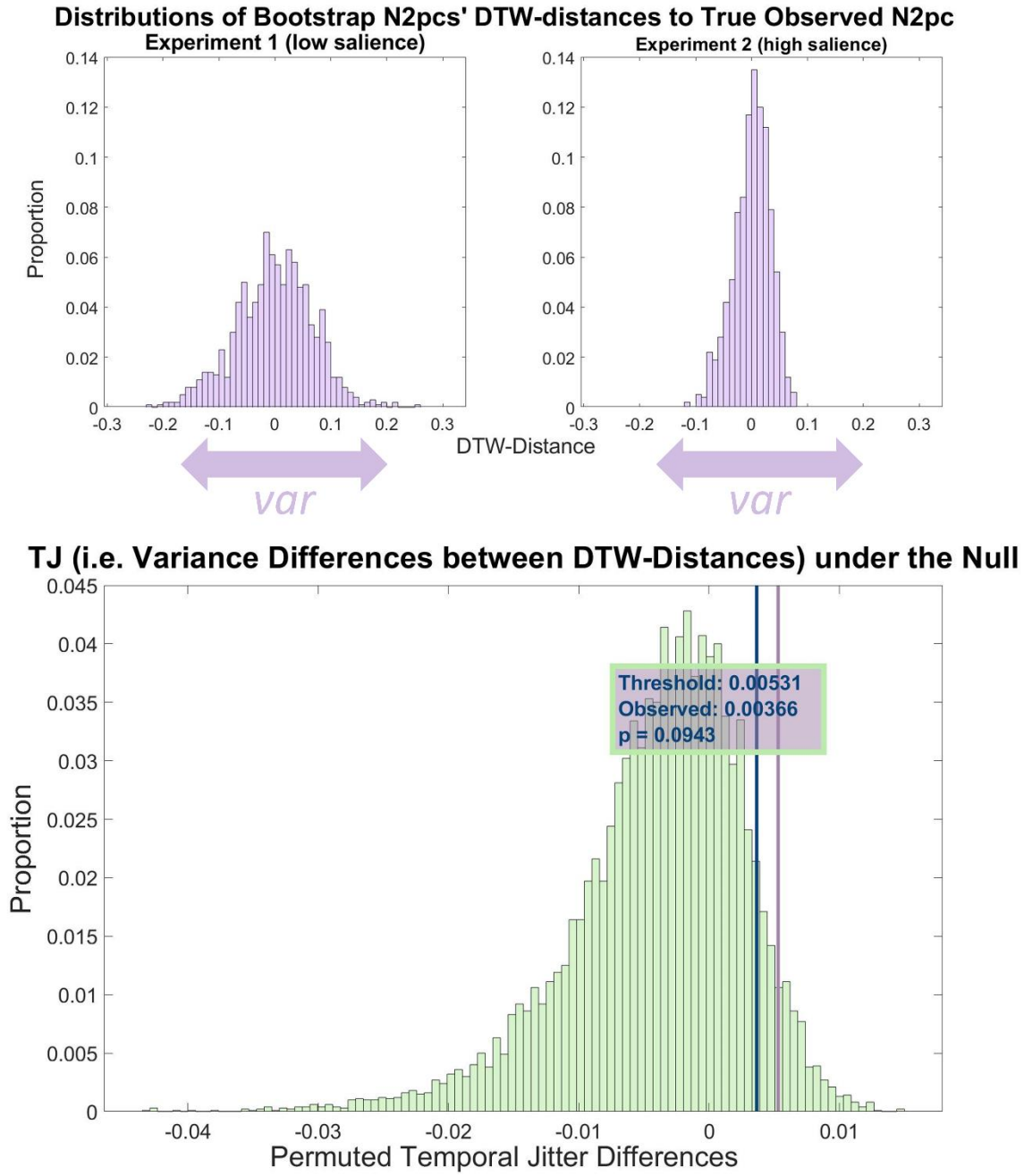


Figure 60. N2pc TJ bootstrap permutation results testing the simple effect of experiment-differences in intrusion trials and assessing the full N2pc (150 – 400 ms) component. Plotting conventions are identical to those of Figure 55. The TJ difference between experiments in intrusion trials for the full N2pc component was found to be non-significant.

Main and simple effects of response condition: We finally assessed the main and simple effects of response condition (correct vs. intrusion) with our bootstrap permutation procedure, adopting the permutation equivalent of a two-sample *dependent* T-test. The results of the main effects are presented in Figure 61 & Figure 62 for the N2pc's onset and the full N2pc, respectively. These plots are again similar to those presented earlier in the context of our between-experiment analyses with the only difference being that the distributions presented in top panels now correspond to correct and intrusion trials instead of corresponding to Experiments 1 and 2. For the main effect of response condition and the N2pc's onset, a statistically significant finding was obtained (observed N2pc TJ difference = 0.00025, p = .0269, see Figure 61). Testing this main effect for the full N2pc did not yield a statistically significant result (observed N2pc TJ difference = 0.00103, p = .3061, see Figure 62). The simple effect of differences between response conditions for Experiment 1 was tested next (Figure 63 & Figure 64). This analysis yielded a non-significant result for the N2pc's onset (observed N2pc TJ difference = 0.03474, p = .0993, see Figure 63). In contrast, a statistically significant finding was obtained for the full N2pc (observed N2pc TJ difference = 0.00314, p = .0212, see Figure 64). Note that the pattern of significance is reversed compared to that demonstrated earlier for the between-experiments effects, with the full N2pc generating a significant finding and the N2pc onset time-window generating a non-significant result. The results of testing the simple effect of response condition for Experiment 2 is presented in Figure 65 & Figure 66. This effect was found to be non-significant for the N2pc's onset (observed N2pc TJ difference = 0, p = .4333, see Figure 65) as well as the full N2pc (observed N2pc TJ difference = -0.00083, p = .8934, see Figure 66).

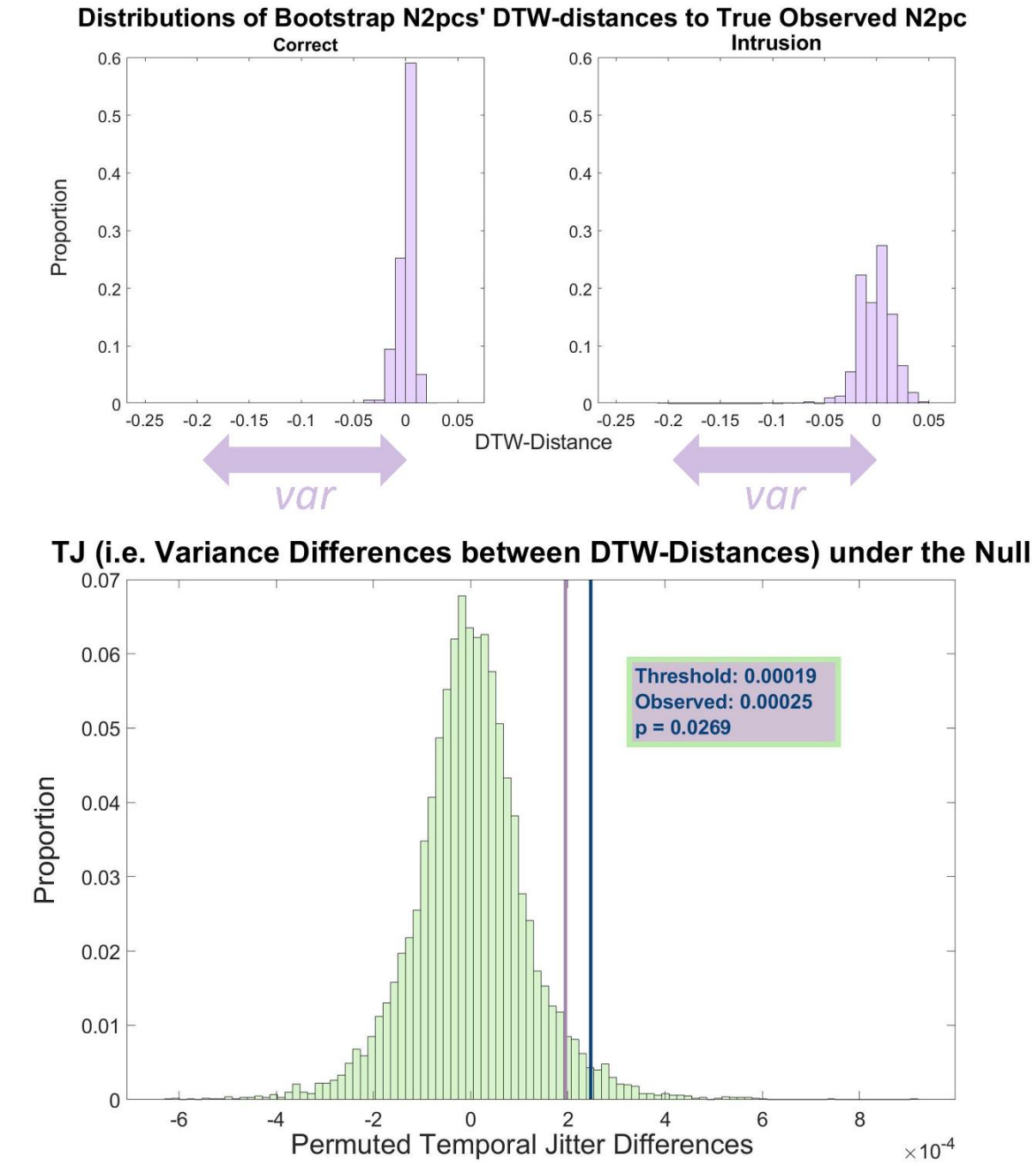


Figure 61. N2pc TJ bootstrap permutation results of the response condition main effect assessing the N2pc's onset (150 – 200 ms). Plotting conventions are identical to those of Figure 55. The main effect of response condition yielded a statistically significant result for the N2pc's onset.

Main Effect of Response Condition – Full N2pc

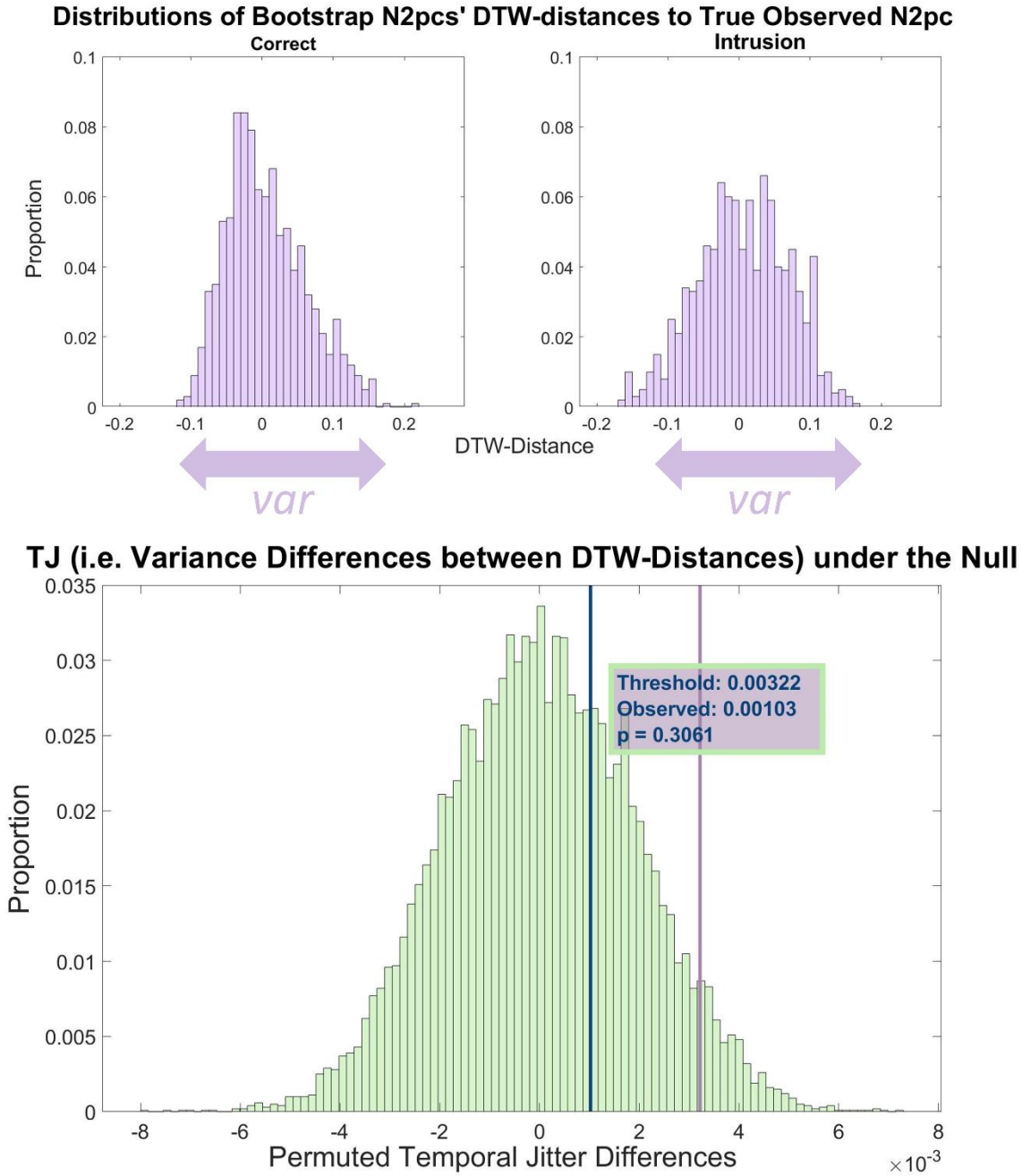


Figure 62. N2pc TJ bootstrap permutation results of the response condition main effect assessing the full N2pc (150 – 400 ms) component. Plotting conventions are identical to those of Figure 55. The main effect of response condition yielded a statistically non-significant result for the full N2pc.

Simple Effect of Condition (Exp. 1) – N2pc Onset

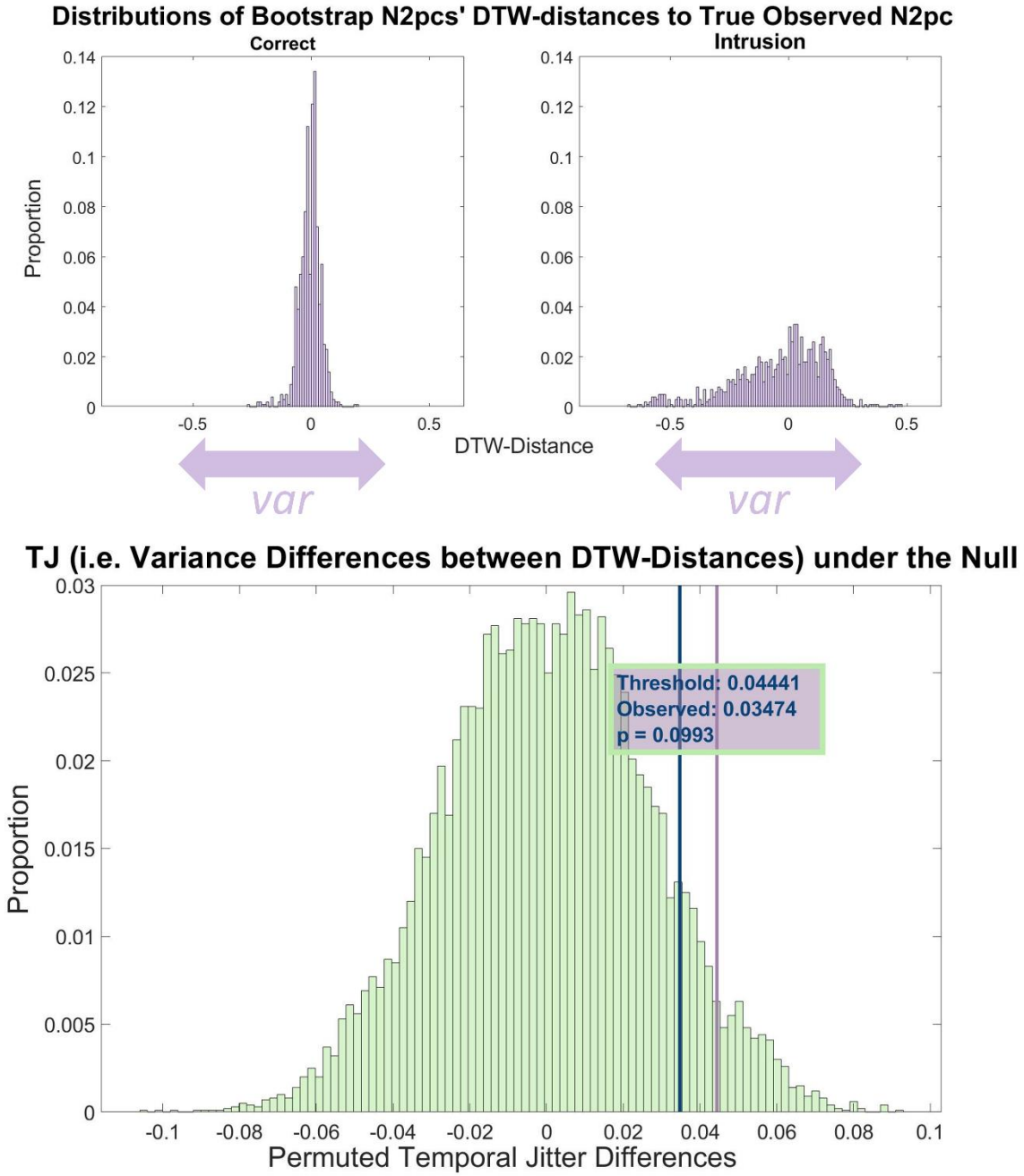


Figure 63. N2pc TJ bootstrap permutation results testing the simple effect of response condition differences in Experiment 1 and assessing the N2pc's onset (150 – 200 ms). Plotting conventions are identical to those of Figure 55. The TJ difference between response conditions in Experiment 1 was found to be non-significant for the N2pc's onset.

Simple Effect of Condition (Exp. 1) – Full N2pc

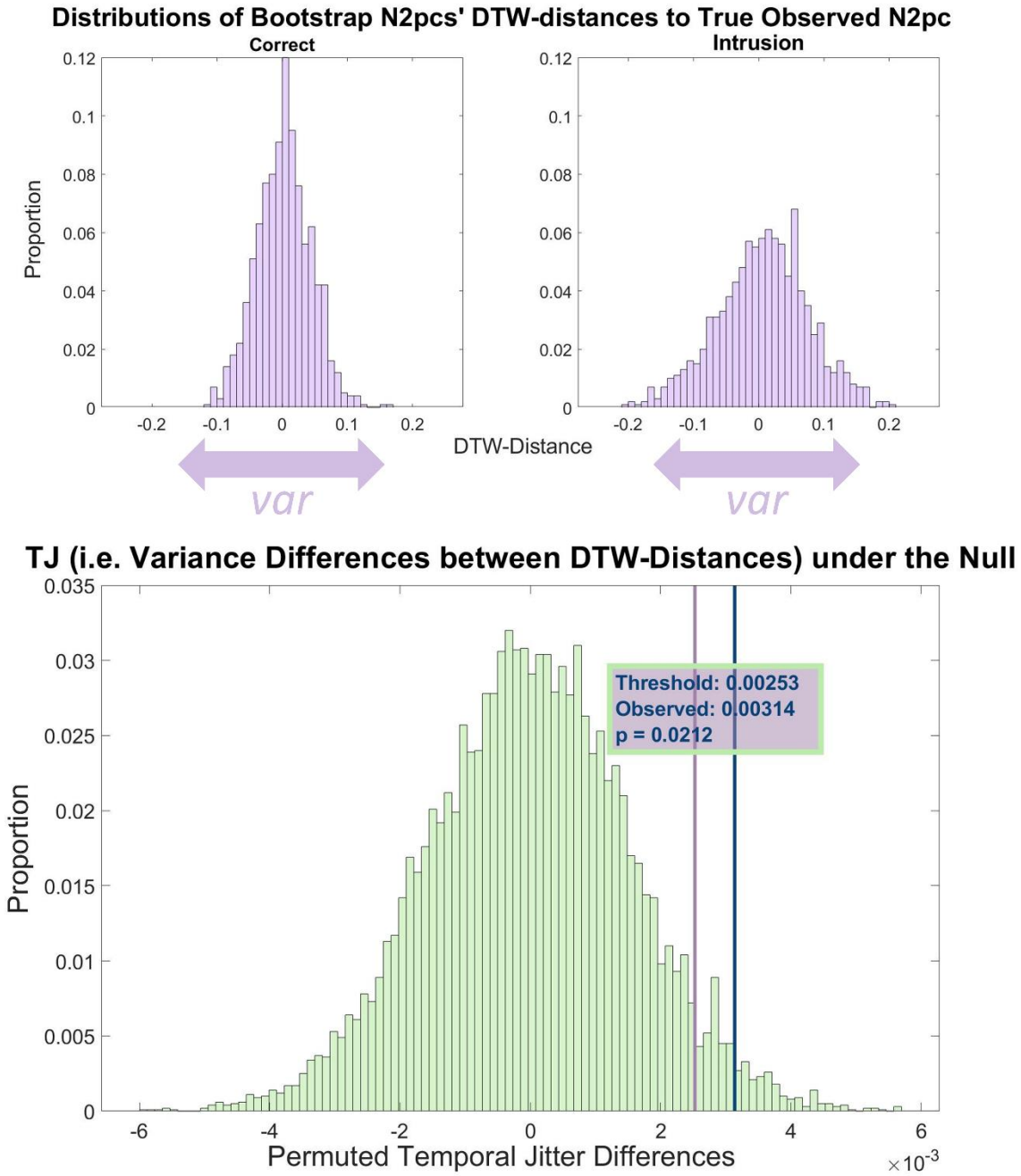


Figure 64. N2pc TJ bootstrap permutation results testing the simple effect of response condition differences in Experiment 1 and assessing the full N2pc (150 – 400 ms) component. Plotting conventions are identical to those of Figure 55. The TJ difference between response conditions in Experiment 1 was found to be statistically significant for the full N2pc.

Simple Effect of Condition (Exp. 2) – N2pc Onset

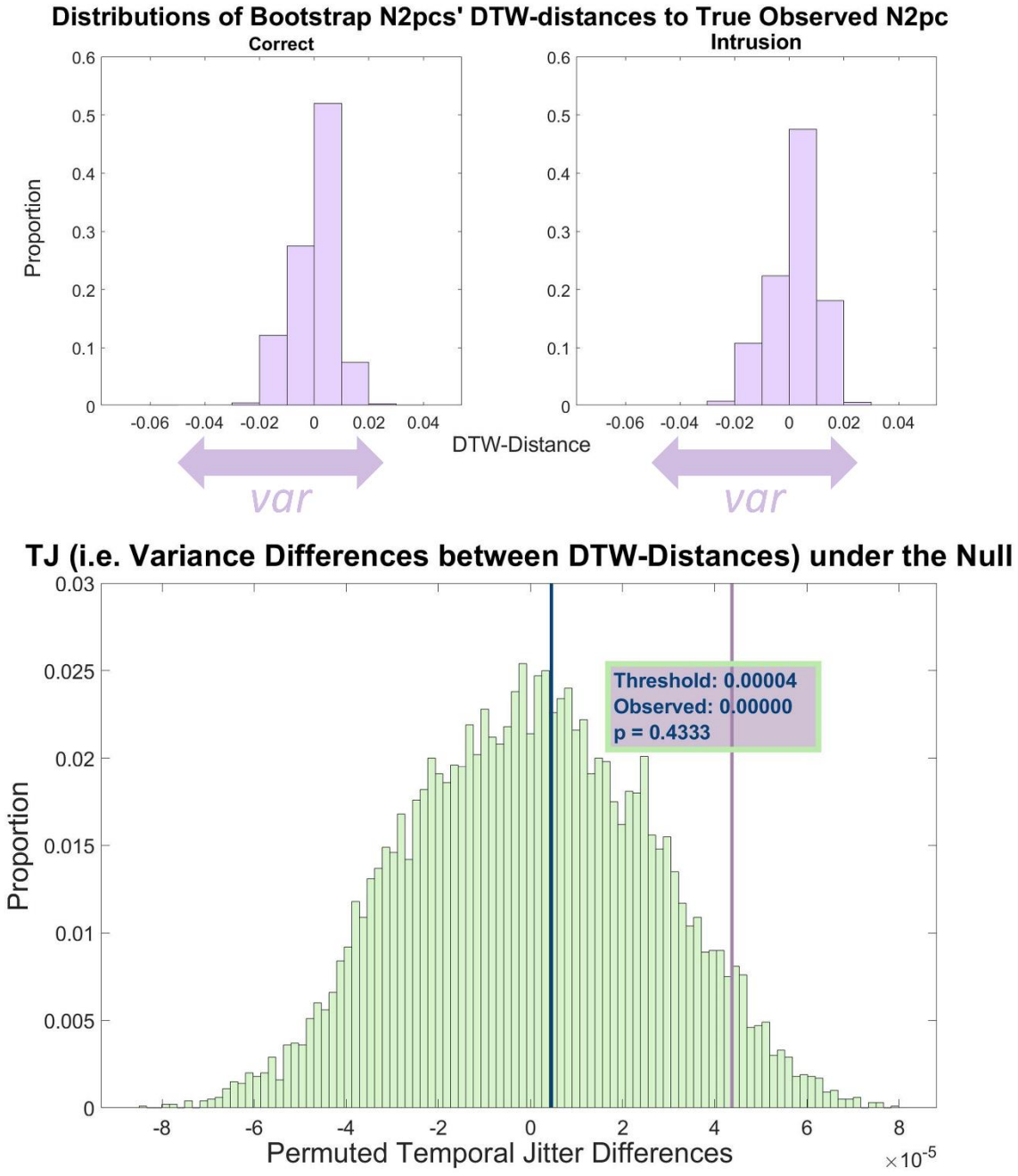


Figure 65. N2pc TJ bootstrap permutation results testing the simple effect of response condition differences in Experiment 2 and assessing the N2pc's onset (150 – 200 ms). Plotting conventions are identical to those of Figure 55. The TJ difference between response conditions in Experiment 2 was found to be non-significant for the N2pc's onset.

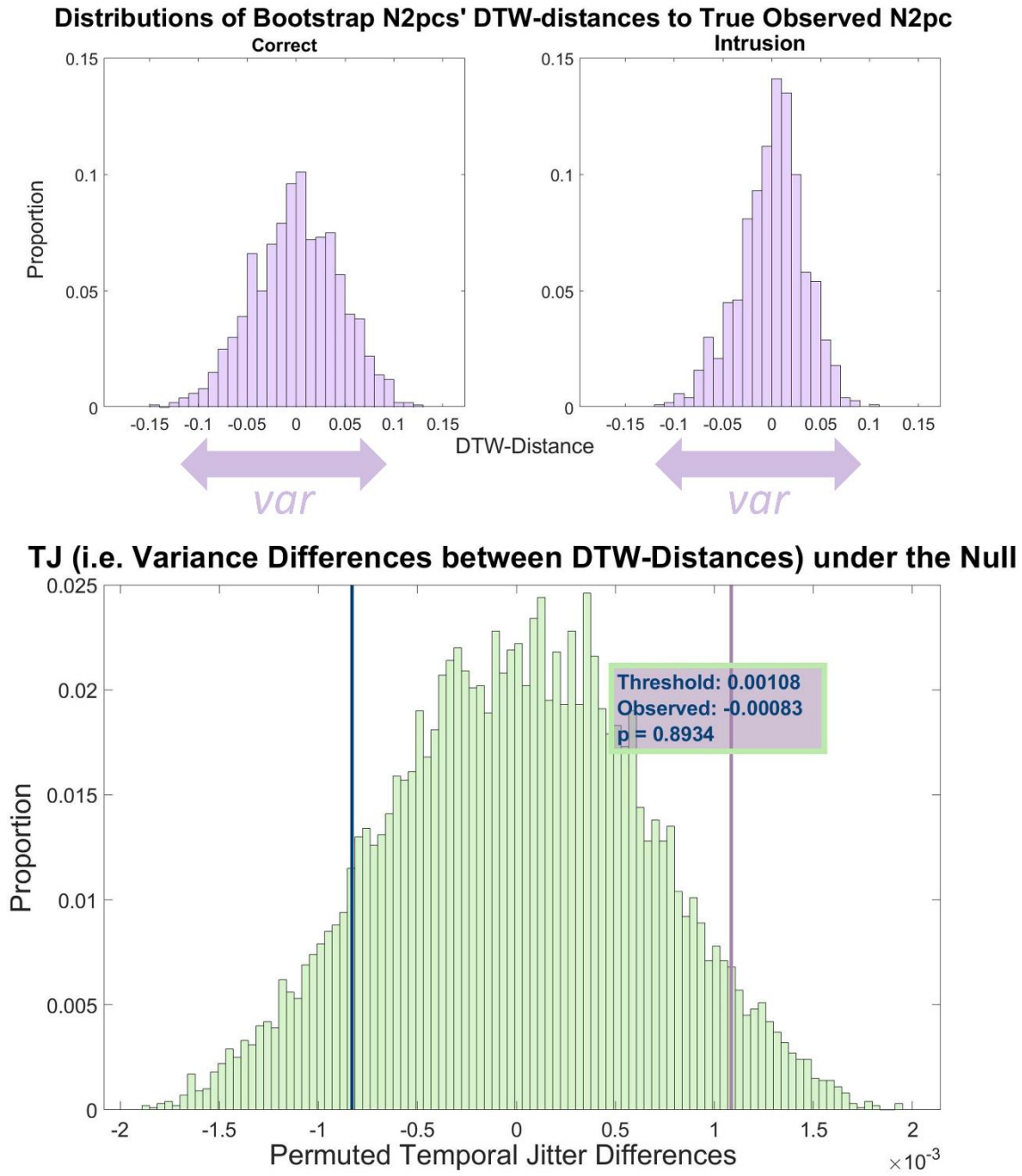


Figure 66. N2pc TJ bootstrap permutation results testing the simple effect of response condition differences in Experiment 2 and assessing the full N2pc (150 – 400 ms) component. Plotting conventions are identical to those of Figure 55. The TJ difference between response conditions in Experiment 2 was found to be non-significant for the full N2pc.

To sum, we have demonstrated significant N2pc TJ differences for the main as well as both simple effects of experiment for the N2pc's onset (150 – 200 ms), suggesting more temporally variable TAE deployment after the less salient key features adopted in Zivony and Eimer's (2020a) Experiment 1. We further revealed a significant main effect between response conditions for the N2pc's onset, suggesting more temporal variability of the N2pc's

onset after intrusion trials. The only significant finding of the full N2pc (150 – 400 ms) was found for the simple effect of response condition in Experiment 1. The latter finding suggested a more temporally variable whole N2pc component after intrusion trials when the key feature is not salient (i.e., in Experiment 1). No significant interaction effects were found.

The results of the previous sections provide evidence in favour of the first part of Hypothesis 3, i.e., that increased key feature salience induces less temporally variable TAE deployment in human cognition. Since the tests of between-experiments effects only resulted in significant results for the N2pc's onset, it is suggested that the impact of key feature salience on the N2pc's TJ affects the onset in particular. Moreover, the two significant effects of response condition suggested that intrusion trials lead to more variable N2pc. It is striking that this is suggested to affect the N2pc's onset when collapsing experiments (i.e., testing the main effect), but the full N2pc in the case of a less salient key feature (i.e., testing the simple effect of Experiment 1). These findings will be elaborated upon later.

2f-ST^2 Model with Gamma Noise and temporally variable TAE-deployment

In this section, we will present the results of running the 2f-ST^2 model in various configurations. It will commence with a brief overview of those model configurations that were already presented in Chapter 4. This overview will be split between model configurations of main and minor interest. In addition, in this overview, we will present the results of the 2f-ST^2 model with response TFL/ type layer filtering and a τK of 9 time-steps (replicating Zivony and Eimer's (2020a) Experiment 2), a configuration not presented in Chapter 4. Importantly, we only include the model's vERP results in this overview for the majority of model configurations, as these were notably affected by the introduction of Gamma Noise. Specifically, the model's response distributions remained comparable to those presented in Chapter 4 for a given configuration, which is why they are not included in the present overview (an example will be presented in the end of this results section to support this claim). All figures that display vERPs will conform to the presentation structure introduced in Chapter 4 and hence display the unsmoothed (i.e., original) and smoothed-difference (e.g., as defined in Chapter 4's glossary (Table 1)) vN2pc and vP3s in panels A and B, respectively.

We will finally present a series of figures specifically aimed at illustrating the extent of the 2f-ST^2 model's replication of the empirical results of Zivony and Eimer's (2020a) Experiments 1 and 2.

vERP Patterns of Model Configurations of Main Interest

Figure 67 - Figure 69 present the vERPs of the model configurations of main interest. Figure 67 displays the vERPs generated from running the 2f-ST^2 model with no task-filtering in the response TFL/ type layer and no fixed delays to either pathway. Thus, Figure 67 corresponds to Figure 24 presented in Chapter 4, which showed the vERPs of this model configuration without Gamma Noise. As expected, the vERPs presented in these two figures are rather comparable, as the impact of Gamma Noise is minimal for low values of τK (0 in this case).

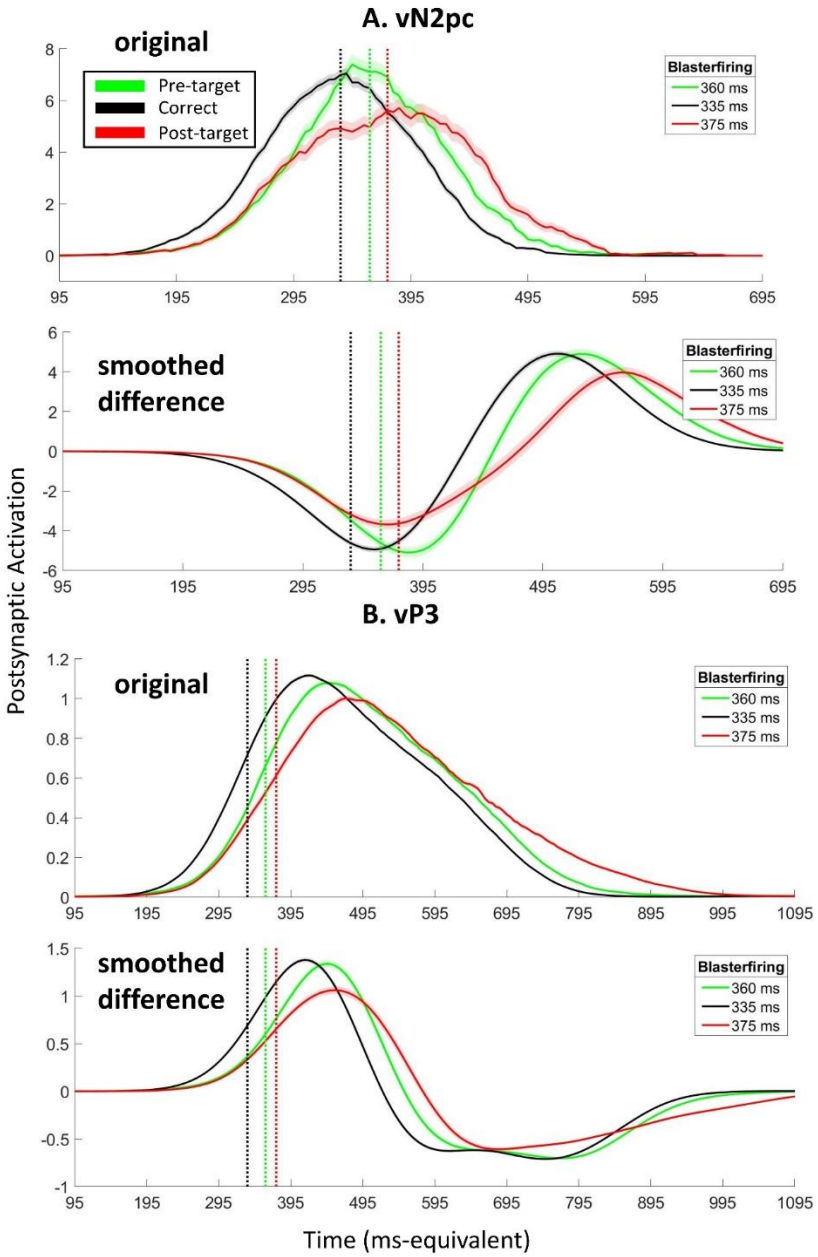


Figure 67. vERPs obtained by running the 2f-ST^2 model with Gamma Noise, no task-filtering in the response TFL/ type layer and no fixed delays to either pathway. Thus, these vERPs correspond to Figure 38, which.showed the vERPs of this model configuration without Gamma Noise. Plotting conventions are identical to those presented in Chapter 4.

Figure 68 displays the vERPs after running the 2f-ST^2 model with task-filtering in the response TFL/ type layer and the implementation of a τK of 24 time-steps. Thus, Figure 68 corresponds to **Figure 32** presented in Chapter 4. The vERPs generated by the Gamma Noise 2f-ST^2 model (Figure 68) exhibit a similar onset- but a later offset-latency to those presented earlier (i.e., in Figure 46). Moreover, the latency-differences between correct and intrusion trials are much more pronounced in Figure 68. This is the case for the vN2pc as well as the vP3. This pattern is again to be expected as the introduction of Gamma Noise has a much more substantial impact on the 2f-ST^2's model dynamics for larger values of τK (24 in this case).

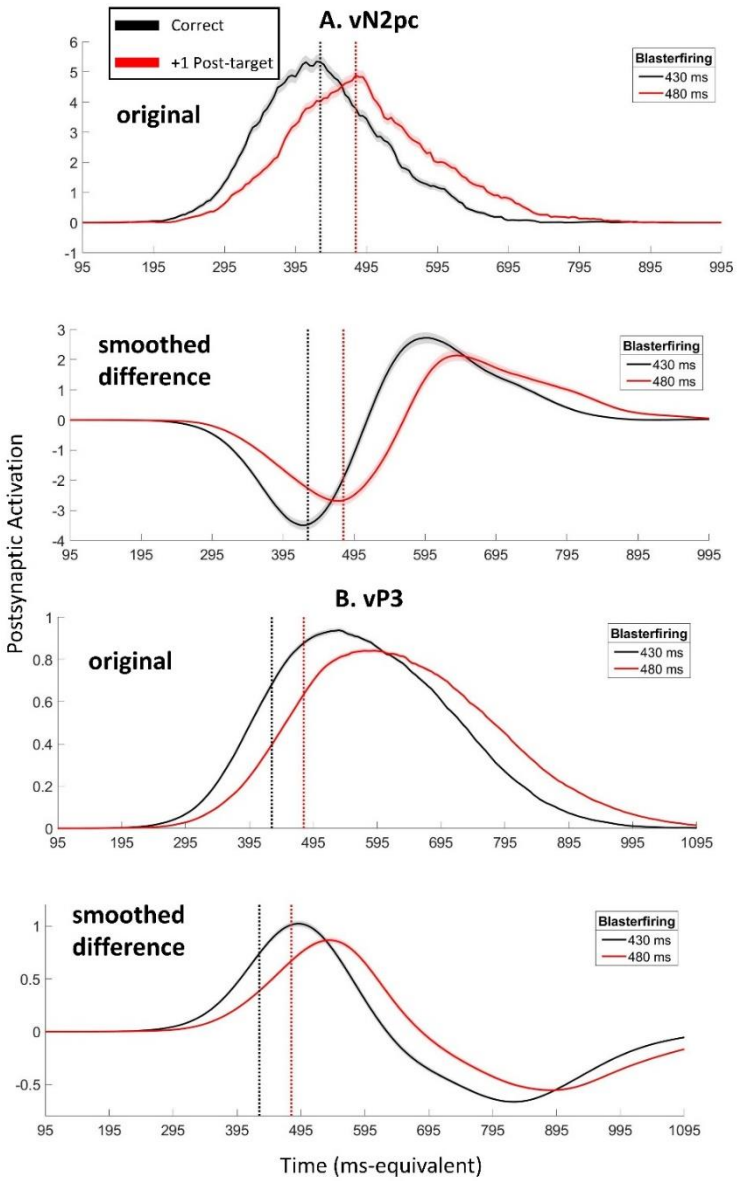


Figure 68. vERPs obtained by running the 2f-ST^2 model with Gamma Noise, task-filtering in the response TFL/ type layer and a τK of 24 time-steps. Thus, these vERPs correspond to Figure 46, which showed the vERPs of this model configuration without Gamma Noise. Plotting conventions are identical to those presented in Chapter 4.

Finally, Figure 69 depicts the vERPs obtained by running the 2f-ST^2 model with Gamma Noise, task-filtering in the response TFL/ type layer as well as a τK of 9 time-steps. This is the configuration used to replicate Zivony and Eimer's (2020a) Experiment 2. As to be expected, the vERPs plotted in Figure 69 exhibit an earlier onset- as well as offset-latency and decreased latency-differences between correct and intrusion trials compared to those generated with a τK of 24 (Figure 68). A more in-depth comparison of these vERPs as well as other aspects of the model's behaviour in addition to a link to the empirical data presented in Zivony and Eimer (2020a) will be presented later on in this section.

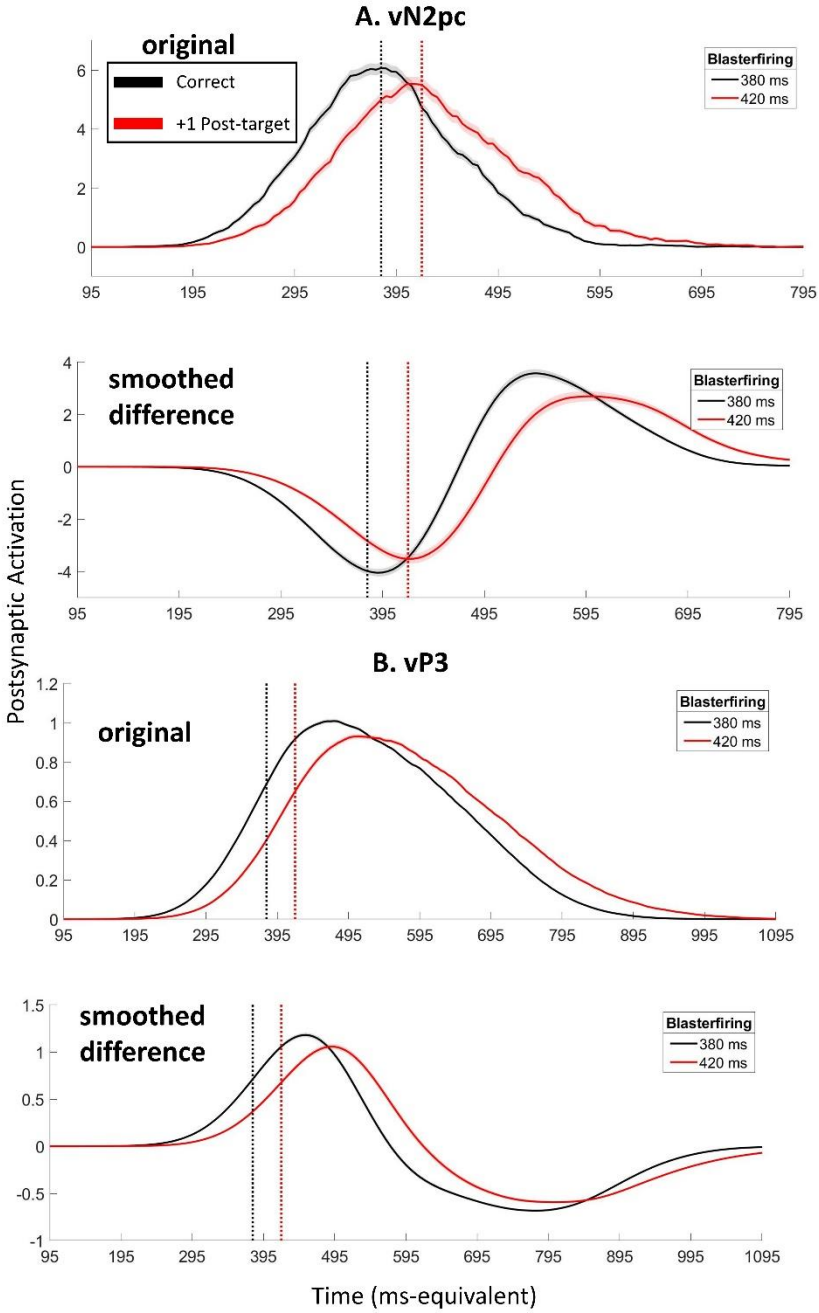


Figure 69. vERPs obtained by running the 2f-ST^2 model with Gamma Noise, task-filtering in the response TFL/ type layer and a τK of 9 time-steps. This model-configuration was used to replicate Zivony and Eimer's (2020a) Experiment 2 and will be presented in detail later on. Plotting conventions are identical to those presented in Chapter 4.

vERP patterns of model configurations of minor interest

For the sake of completeness, we present the model configurations of minor interest previously presented in Chapter 4 next. Figure 70 displays the vERPs after running the 2f-ST^2 model without task-filtering in the response TFL/ type layer and the implementation of a τK of 24 time-steps. Thus, Figure 70 corresponds to Figure 34 presented in Chapter 4. The vERPs generated by the Gamma Noise 2f-ST^2 model (Figure 70) exhibit a similar onset- but a later offset-latency to those presented earlier (i.e., in Figure 34). Additionally, latency-differences in vERPs between response conditions (i.e., correct, +1- and +2- intrusions) are again more pronounced after implementing Gamma Noise to the 2f-ST^2 model.

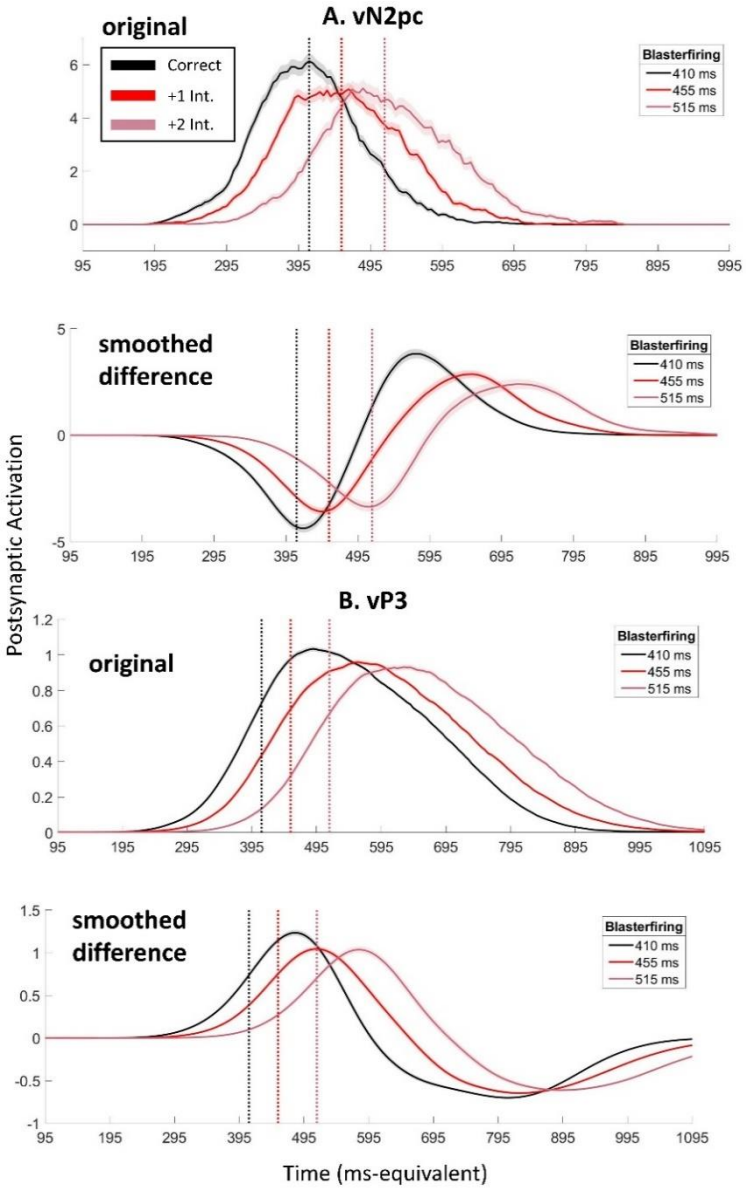


Figure 70. vERPs obtained by running the 2f-ST^2 model with Gamma Noise, without task-filtering in the response TFL/ type layer and a τK of 24 time-steps. These vERPs hence correspond to **Figure 34**, which showed the vERPs of this model configuration without Gamma Noise. Plotting conventions are identical to those presented in Chapter 4.

Figure 71 displays the vERPs after running the 2f-ST^2 model with Gamma Noise, task-filtering in the response TFL/ type layer and no fixed delays to either pathway. Thus, Figure 71 corresponds to Figure 36 presented in Chapter 4. The vERPs in Figure 71 are comparable to those presented in Figure 36. This is again to be expected due to the minor impact the introduction of Gamma Noise has for low values of τK (in this case being 0).

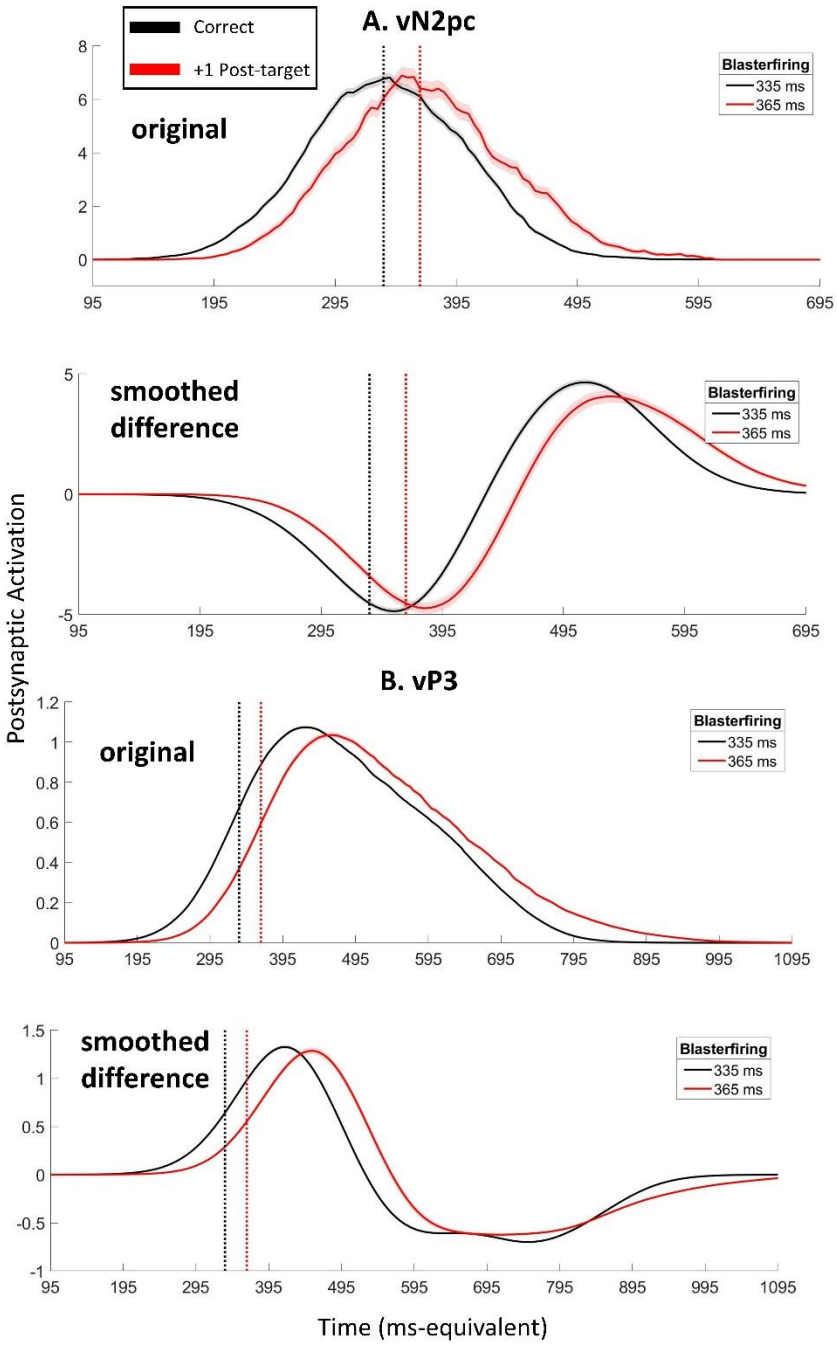


Figure 71. vERPs obtained by running the 2f-ST^2 with Gamma Noise, task-filtering in the response TFL/ type layer and no fixed delays to either pathway. Thus, these vERPs correspond to **Figure 36**, which showed the vERPs of this model configuration without Gamma Noise. Plotting conventions are identical to those presented in Chapter 4.

2f-ST^2 model replicating Zivony and Eimer's (2020a) Experiments 1 & 2

Figure 72 displays the response distributions empirically found by Zivony and Eimer (2020a) in Experiments 1 (Figure 72A) and 2 (Figure 72C). We qualitatively replicate these response distributions with the 2f-ST^2 model, implementing τK values of 24 (Figure 72B) and 9 (Figure 72D) time-steps to simulate Experiments 1 (low key feature salience) and 2 (high key feature salience), respectively. Figure 72B demonstrates that with a distribution of 47.9% correct and 52.1% intrusion responses, we replicate Zivony and Eimer's (2020a) first experiment. The latter (Figure 72A) yielded 34.95% correct, 57.65% intrusion and 7.4% wrong responses (Zivony & Eimer, 2020a). Figure 72D displays the response distribution of the 2f-ST^2 model implementing a τK of 9 time-steps, generating 62.3% correct and 37.7% intrusion responses. These proportions compare well to those found in Zivony and Eimer's (2020a) second experiment: 55.77% correct, 39.5% intrusion and 4.73% wrong responses (Zivony & Eimer, 2020a). The proportion values stated for Zivony and Eimer's (2020a) human data correspond to averages across Experiment 1's A and B runs (i.e., run B being a direct replication of run A) and across grey and digit distractors for Experiment 2. The latter is justifiable, since no differences between distractor type were found in the original study and was further done by the authors themselves before running an ANOVA that revealed a main effect of experiment, implying overall higher accuracy in Experiment 2 than 1 (Zivony & Eimer, 2020a). Furthermore, and as stressed in Chapter 4, the apparent discrepancies in the proportion of correct trials between our model and human data is because we excluded trials in which wrong responses (misses) were given by our model.

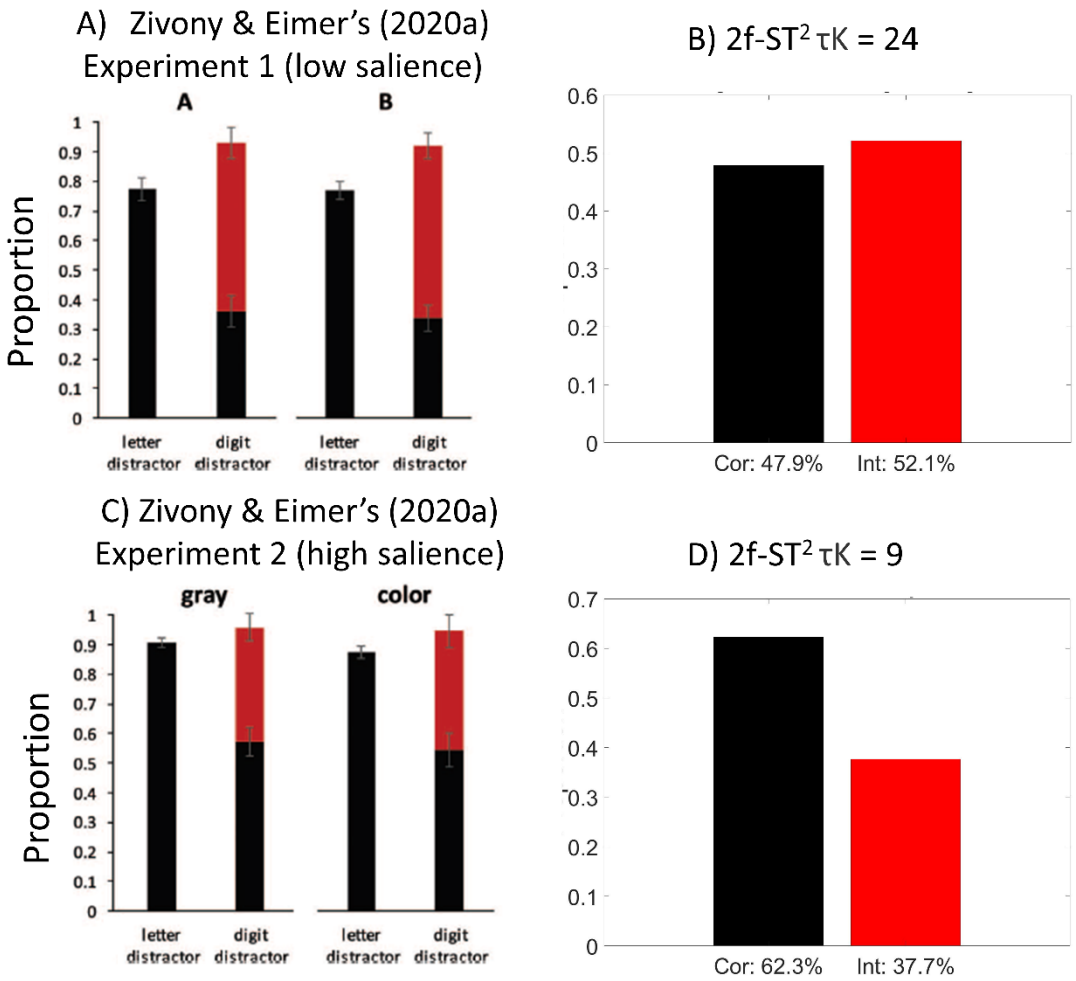


Figure 72. Comparison of response distributions empirically found by Zivony and Eimer (2020a*)* (panels A & C) and those generated by the 2f-ST^2 model, implementing τKs of 24 (panel B) and 9 (panel D) time-steps. The 2f-ST^2 model replicates the accuracy patterns of Experiment 1 (panel A) with a τK of 24 (panel B) and those of Experiment 2 (panel C) with a τK of 9 time-steps. The empirical proportion values mentioned in the text were the average values across Experiments 1 A and B for Experiment 1 and across gray and coloured distractors for Experiment 2.

Figure 73 plots the human ERP components of Zivony and Eimer's (2020a) Experiments 1 (Figure 73 A&B) and 2 (Figure 73 C&D), with black and red lines representing correct and intrusion conditions, respectively. Zivony and Eimer (2020a) ran two between-experiment ANOVAs on the N2pcs of the two experiments, which revealed that the N2pc emerged earlier in Experiment 2 (187 ms) than 1 (225.5 ms) and that N2pc amplitudes were larger in Experiment 2. The amplitude ANOVA furthermore yielded a significant interaction between experiment and response condition, implying that the enhancement of N2pcs for correct compared to intrusion trials was more substantial in Experiment 1 than in Experiment 2. The latter is easily observable in Figure 73, as the N2pc of correct (black) trials shows a much larger amplitude than intrusion (red) trials' N2pc in Experiment 1 (Figure 73A). However, in Experiment 2 (Figure 73) this difference in amplitudes is much weaker.

Linking the human ERP patterns presented in Figure 73 to the results presented earlier in the context of the N2pc TJ analyses, it can be observed that the statistically significant between-experiment latency difference (with Experiment 2 showing an earlier N2pc component) obtained by Zivony and Eimer (2020a) fits well with the significant main effect of the N2pc's onset exhibiting less temporal jitter when key features are very salient (as was the case in Experiment 2, see Figure 55). This is because as ERP components unfold comparably closer to their earliest possible time-frame (as is the case for the second compared to the first experiment), there is less room for temporal variability across trials. We will provide further thoughts on this interaction as well as provide reasons for why the same TJ difference was not observed between experiments when considering the full N2pc time window in the discussion of the present chapter. In addition, acknowledging that a corresponding statistical test is lacking, it can be argued that Figure 73 shows rather small latency-differences between experiments for the P3 component. Therefore, and according to the reasoning provided in this chapter on the interaction between ERP-latency and jitter in general, one would expect the jitter-differences shown for the N2pc to not translate to the P3 component. An analysis probing the latter expectation constitutes an interesting direction for future research.

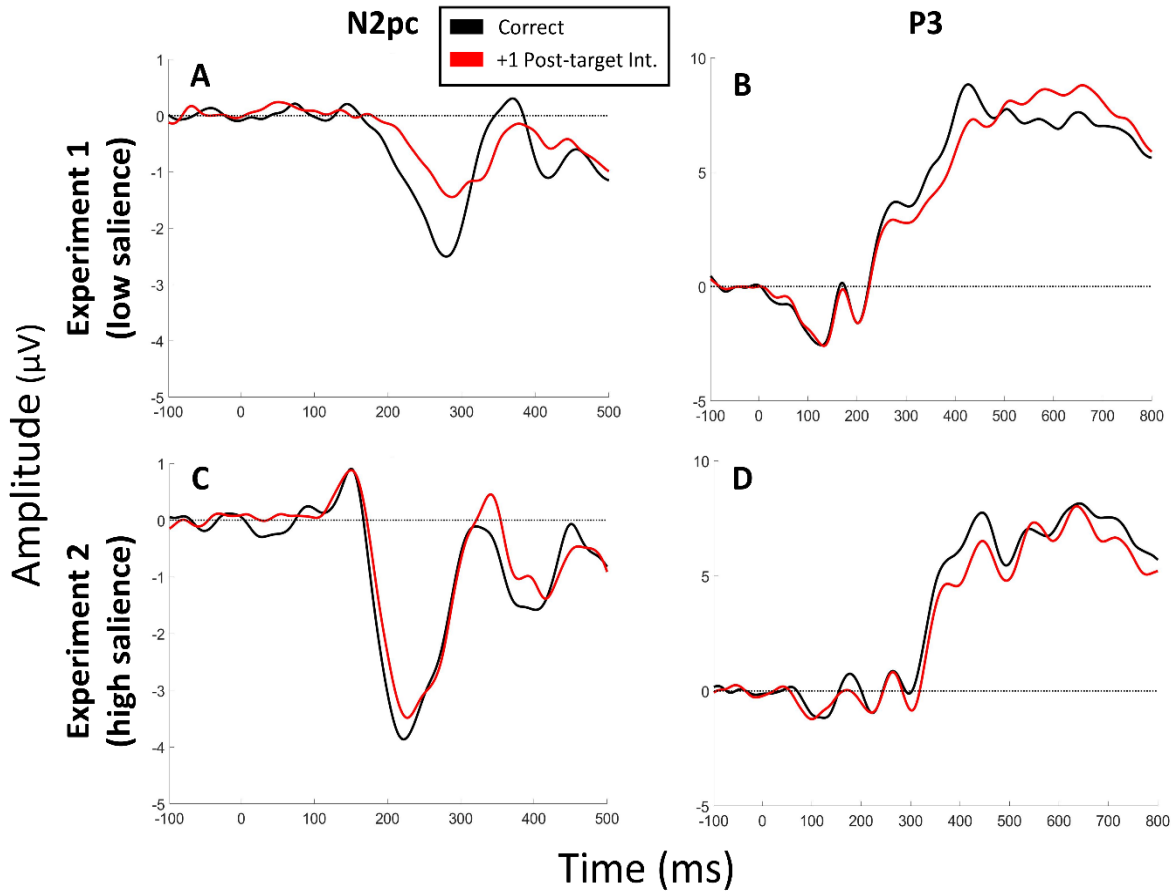


Figure 73. Human ERP Comparison plots. Panels A and C present the N2pc components of Experiments 1 and 2, respectively. Panels B and D present the P3 components of Experiments 1 and 2, respectively. Black and red lines correspond to correct and intrusion trials.

We present comparison plots of the 2f-ST^2 model's vERPs, implementing τK values of 9 and 24 time-steps, replicating Zivony and Eimer's (2020a) Experiments 1 and 2, respectively, in Figure 74 and Figure 75. These vERP plots follow the conventions of previous vERP figures, e.g., those presented in Chapter 4. Hence, black and red vERP traces represent correct and intrusion trials, respectively and vertical dotted lines represent a response condition's average blaster-firing latency. We furthermore again present the original (i.e., unsmoothed, panels A & B) as well as smoothed-difference (panels C & D) vERPs. Figure 74 illustrates that we qualitatively replicate the N2pc patterns found by Zivony and Eimer (2020a) in humans: earlier and larger (i.e., higher amplitude) (v)N2pcs in Experiment 2/τK of 9 (Figure 74 A & C) than 1/τK of 24 (Figure 74 B & D). The 2f-ST^2 model also shows smaller amplitude-differences between correct and intrusion trials with a τK value of 9 than with a τK of 24 time-steps. We therefore also qualitatively replicate the significant interaction effect between experiments and response conditions demonstrated in Zivony and Eimer's (2020a) N2pc-amplitude ANOVA, discussed in the previous paragraph.

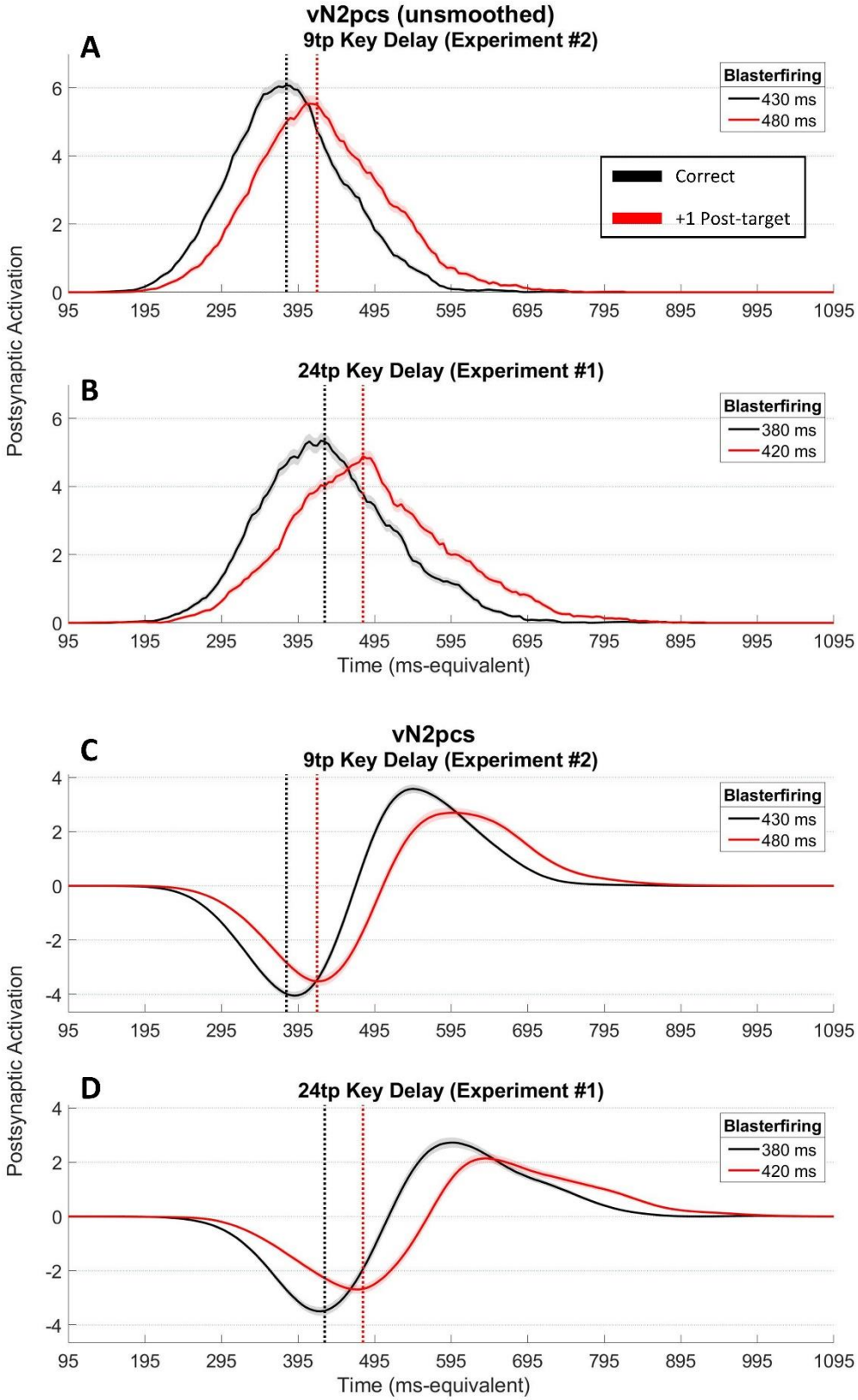


Figure 74. vN2pc Comparison Plots. Panels A and C plot the unsmoothed and smoothed vN2pcs of the 2f-ST^2 model with active task-filtering in the response pathway and a τK value of 9 time-steps. Panels B and D show these for the same model configuration now implementing 24 time-steps as τK. Black and red lines correspond to correct and intrusion trials, respectively and vertical dotted lines indicate average blaster-firing latencies of a given response condition, with respective numerical values being provided in plots' legends. The 2f-ST^2 model qualitatively replicates the human N2pc results shown in Zivony and Eimer's (2020a) Experiments 1 and 2.

Figure 75 displays the vP3s after running the 2f-ST^2 model with τKs of 9 (Figure 75 A & C) and 24 (Figure 75 B & D) time-steps. Again, the vP3s of both response conditions emerged earlier and with larger amplitudes after running the model with a τK of 9 time-steps. This is in line with the patterns shown in Figure 74, which is to be expected because of the direct impact the blaster (vN2pc) has on later model-layers captured in the vP3 (compare Chapter 4 for an in-depth discussion of the relationship between the vN2pc and the vP3). Additionally, the vP3s do not seem to reflect the enhanced amplitude differences between correct and intrusion trials after running the model with a τK of 9 than with 24 time-steps, which were present in the vN2pcs. An empirical study explicitly designed to test human P3 dynamics between response conditions (i.e., correct and intrusions) as well as differences in key feature salience, simulated by τKs of 9 and 24, is still lacking in the literature. Therefore, the results displayed in Figure 75 rather reflect predictions for a representative empirical study.

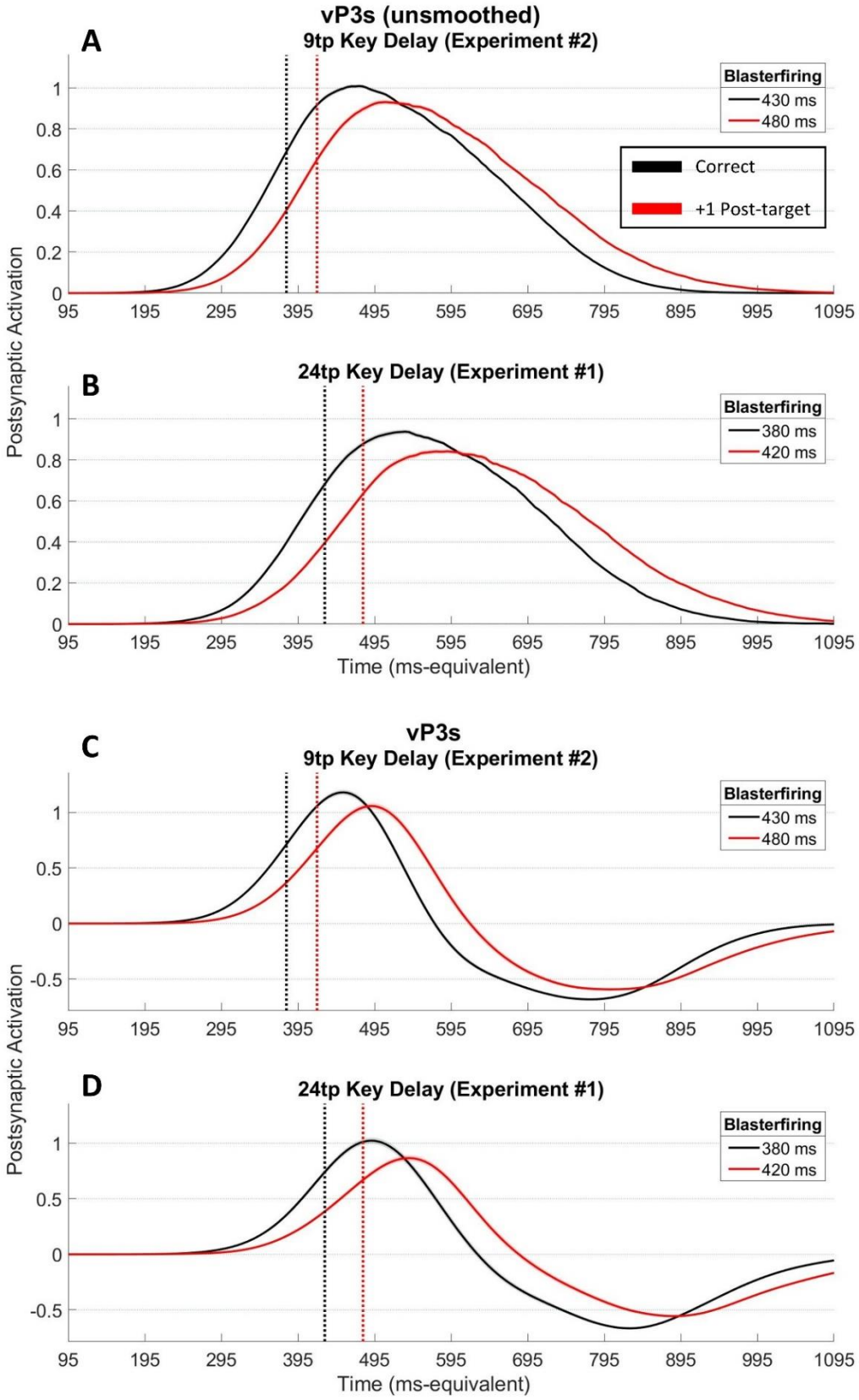


Figure 75. vP3 Comparison Plots. Plotting conventions are identical to those of Figure 74. Since representative human P3 data is still lacking, these results reflect predictions of the 2f-ST^2 model.

Finally, we present the 2f-ST^2 model's blaster-firing latencies in Figure 76 after running the model with a τK of 9 (Figure 76A) and 24 (Figure 76B) time-steps. Blaster-firing latencies specifically reflect the time-point at which a given trial's vN2pc (i.e., the postsynaptic activation of the blaster-output layer) reached 50% of the area under its curve (AUC). The across-trial average of these values has been plotted as vertical lines in vERP plots. The distributions plotted in Figure 76 illustrate the impact of the implementation of Gamma Noise to the 2f-ST^2 model well, as due to the larger value of this parameter with a τK of 24, the blaster fires with more temporal variability, making the distribution in Figure 76B wider than that of Figure 76A (note the difference in x-axis ranges). The latter interaction between τK and variability in blaster-firing latencies represents our implementation of TAE-deployment varying temporally based on key feature salience in human cognition.

Additionally, and strikingly, the long tail of the gamma noise distribution particularly contributes to the intrusion distribution in Figure 76B. This is not surprising, since one would naturally expect that the later the blaster firing, the more likely that one obtains an intrusion. However, an important consequence of this is that intrusions are disproportionately delayed relative to correct responses. This disproportionate delaying counteracts the diminished vERP latency differences when replicating Zivony and Eimer's (2020a) Experiment 1, discussed in the context of the loss of responsiveness pattern in the previous chapter. To stress and as shown in Figure 77, the implementation of Gamma Noise did not reduce the 2f-ST^2 model's efficacy in replicating the counterintuitive RT findings of Botella (1992), since the impact of Gamma Noise is minimal when τK is small (note that τK is set to zero when replicating Botella's (1992) RT findings).

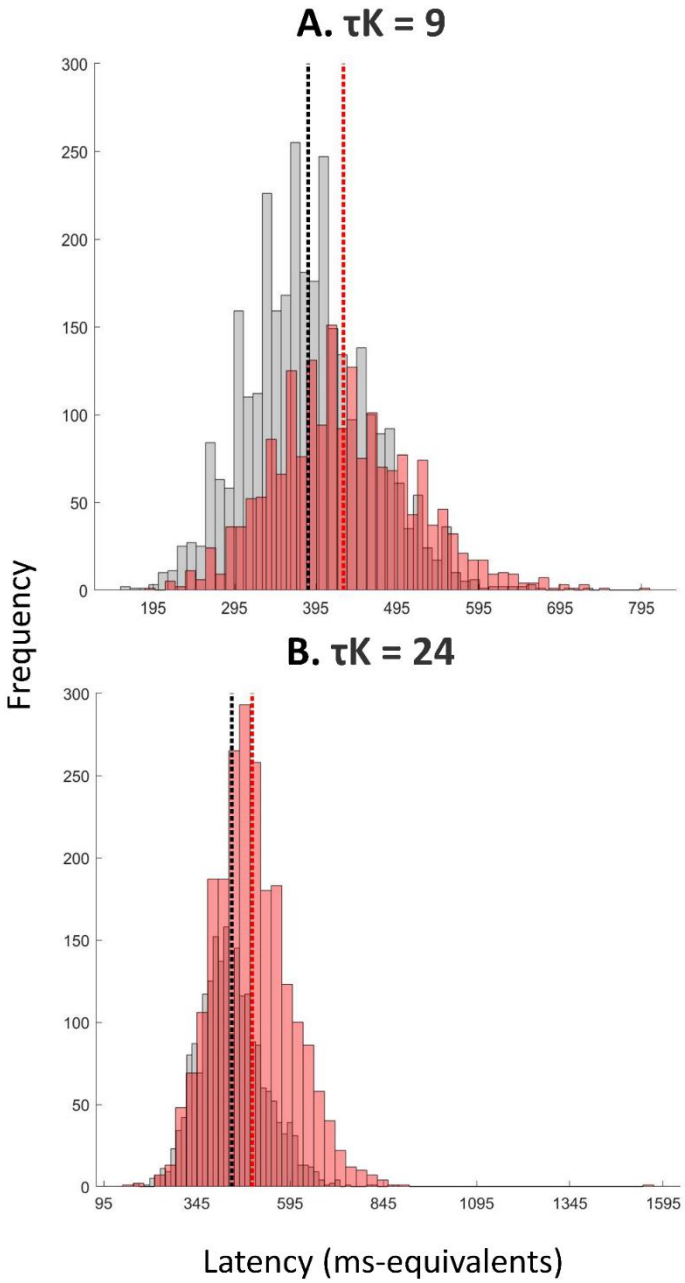


Figure 76. Blaster firing latency-distributions. Panels A and B show the distribution of blaster-firing latencies (defined as the 50% AUC latency of the vN2pc of a given trial) of the 2f-ST^2 model with active task-filtering in the response pathway and τKs of 9 and 24 time-steps, respectively. Across-trial average blaster-firing latencies are indicated by vertical dotted lines and black and red colours correspond to correct and intrusion trials, respectively. Due to the introduction of Gamma Noise, the blaster fires with much higher temporal variability when the model is run with a τK of 24 time-steps. This counteracts the pattern of diminished vERP latency differences between correct and intrusion trials presented in the previous chapter when replicating Zivony and Eimer's (2020a) Experiment 1. Note the difference in x-axis limits between the two panels.

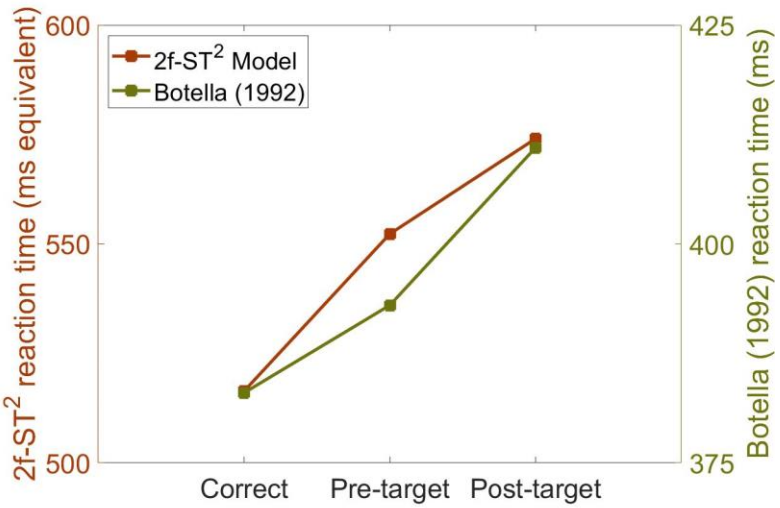


Figure 77. The 2f-ST^2 model with Gamma Noise still replicates Botella's (1992) counterintuitive RT findings. This is because the impact of Gamma Noise is minimal when τK is small, such as being equal to zero in the case of it replicating the experiment of Botella (1992).

To summarise, the version of the 2f-ST^2 model presented in previous chapters, containing synchronised key and response pathway processing, qualitatively replicates all examined empirical findings with the exception of the ERP latency differences between correct and intrusion trials shown by Zivony and Eimer (2020a). The equivalent vERP patterns generated by the 2f-ST^2 model with synchronised key and response pathways, presented in detail in Chapter 5, demonstrated diminished latency differences. As explained in detail previously, we believe this discrepancy to reflect a limitation of our computational model rather than a weak empirical finding. We further believe this limitation of the 2f-ST^2 model presented previously to be due to key and response pathway processing being synchronised within each trial. We demonstrate that implementing Gamma Noise to the 2f-ST^2 model solves the issue of diminished vERP latency differences (compare Figure 74 - Figure 76) when replicating Zivony and Eimer's (2020a) experiments. Our empirical N2pc TJ analyses further provide evidence suggesting that as key features become more salient, the variability with which TAE is deployed decreases. The version of the 2f-ST^2 model presented in the previous two chapters did not model this interaction. Therefore, adding Gamma Noise to the 2f-ST^2 model not only solved the issue of diminished vERP latency differences when responsiveness is lost, it also increased the model's neurophysiological plausibility. As a result, we would argue that the current version of the 2f-ST^2 model (with Gamma Noise) constitutes the most complete version of our computational model.

Linking these findings to this thesis, the previous sections demonstrated that the 2f-ST^2 model with Gamma Noise can model important distractor intrusion phenomena, since it

now considers the temporal variability underlying TAE deployment being modulated by key feature salience (increasing the model's neurophysiological plausibility). Therefore, these sections provide further evidence in favour of Hypothesis 2 of this thesis, i.e., that the 2f-ST^2 model accounts for a broad range of findings obtained with distractor intrusion experiments. Moreover, the Gamma Noise 2f-ST^2 model's behavioural as well as virtual EEG results presented in this section provide evidence in favour of the second part of Hypothesis 3, i.e., that less temporally variable TAE deployment after increased key feature salience can be computationally modelled with the 2f-ST^2 model.

Discussion

Does Key Feature Salience modulate TAE's Temporal Variability?

Hypothesis 3 of this thesis states that increased key feature salience induces less temporally variable TAE deployment in human cognition. The results of the current chapter provide evidence in favour of this hypothesis and specifically suggest that key feature salience modulates the onset latency of TAE deployment. This is suggested by the observation that TJ differences between experiments were shown in the onset of N2pcs, as statistically significant results were obtained for the onset's main and simple effects of experiment (compare Figure 55, Figure 57 & Figure 59) As an illustration, N2pc TJ differences affecting the component's onset specifically is indicated by the Fully Flattened Average (FFA) N2pc components of the two experiments (Figure 78). Figure 78 plots the FFAs of Experiments 1 and 2 in blue and pink, respectively, and displays the complete time-range (-100 – 800 ms) in the top and the time-range of the N2pc (100 – 400 ms) in the bottom panel. Figure 78 indicates that N2pc's offsets are similarly sharp whereas the onset is much sharper in Experiment 2's N2pc.

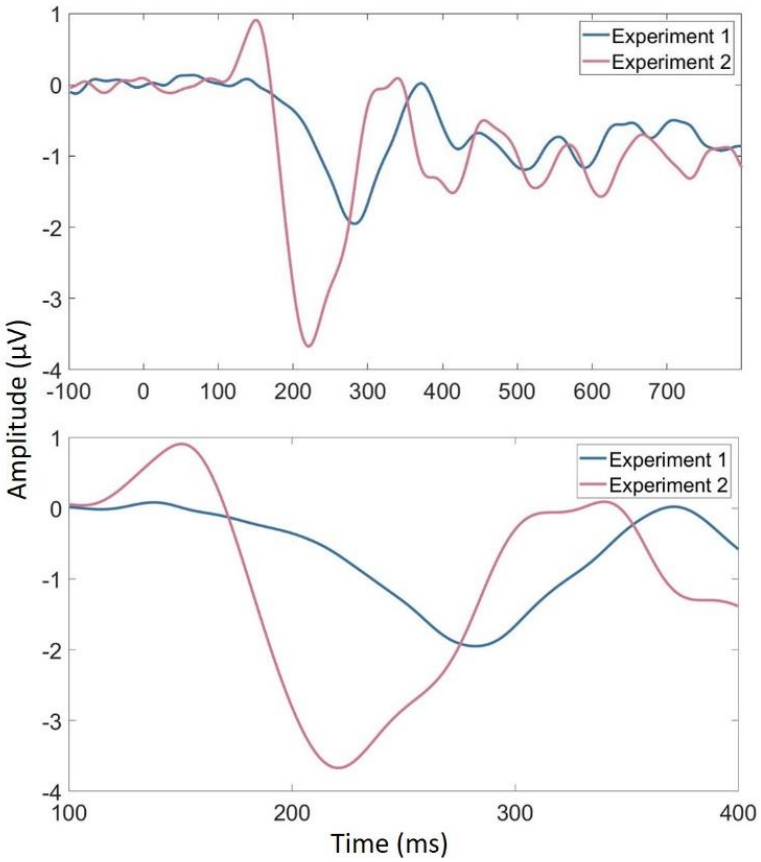


Figure 78. N2pc FFAs. Collapsed all trials within an experiment across experimental conditions (correct & intrusion responses) and subjects. Blue and purple lines correspond to FFAs of Experiments 1 and 2, respectively. The FFA serves as an illustration of key feature salience affecting the N2pc's onset in particular.

N2pc TJ differences being suggested to affect the component's onset in particular is worth further elaboration. One might expect that decreases in N2pc TJ across trials are accompanied with narrower and more symmetric components on the Grand Average level. However, the more salient key feature did not seem to reduce the TJ in Experiment 2's N2pc in this way but instead affected the component's onset in particular. Figure 79 schematically illustrates these two cases, plotting hypothetical single-trial N2pcs in blue and pink lines for Experiments 1 and 2, respectively. In each case and for each experiment, we present hypothetical N2pcs of early, typical, and late trials in top, middle, and bottom rows. Time unfolds from left to right. Case 1 is plotted in the top panel and displays the scenario in which temporal differences between trials shift the whole component laterally. Increased TJ in Experiment 1 is reflected in early and late trials being temporally more distant from the typical trial than is the case for Experiment 2, in which early, typical, and late N2pcs unfold at similar latencies. However, our results suggest that N2pc TJ differences affected the onset in particular, which is schematically illustrated as Case 2 in the bottom panel of Figure 79. In Case 2, the onset-latencies vary substantially between early, typical, and late trials for both experiments' hypothetical N2pcs. However, the components' offsets occur at similar time points in each trial type. In this scenario, late trials are especially interesting to consider, as these would be trials in which the component commences late (i.e., late onset), but ends abruptly, leading to a narrower component compared to early trials, for example, and to similar offset-latencies between trial types. We argue that the characteristic of latency differences between trials affecting the N2pc's onset in particular is generally present in both experiments. However, and importantly, the increased TJ in Experiment 1 would imply that this issue is much more pronounced for experiments with less salient key features. The latter is illustrated in the notably shallow onset of Experiment 1's FFA N2pc (Figure 78, blue line), as onset-latencies in individual trials of this experiment supposedly vary substantially.

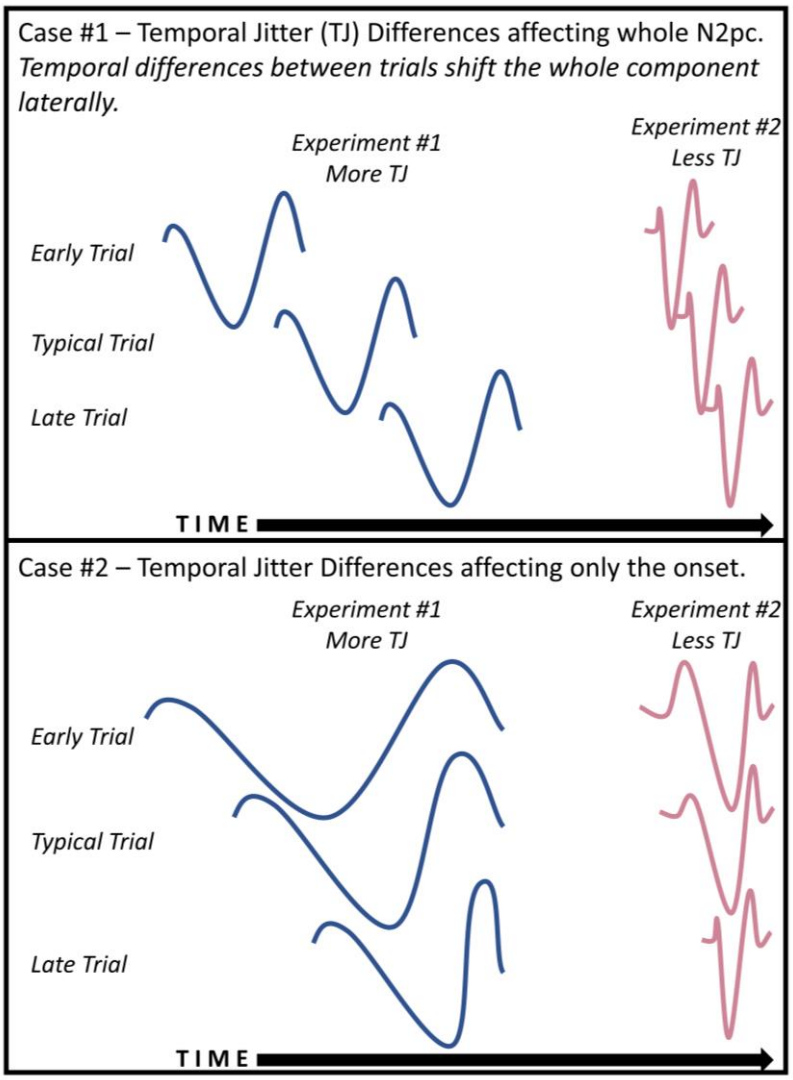


Figure 79. N2pc Temporal Jitter (TJ) differences between Experiments. The diagram shows two cases. In Case 1, TJ differences affect the whole component (i.e., happen laterally), whereas in Case 2 only the onset is affected by jitter-differences. We present hypothetical N2pc waveforms of single trials for Experiments 1 and 2 in left and right columns, respectively. For each experiment and each case, we present a hypothetical early, typical, and late trial's N2pc. In Case 2 the onset-latency varies (and more so for Experiment 1), but the offsets are at a similar point in time in each trial type. The absence of TJ-differences in the offset could be due to a hard-limit in those later time points or a subsequent cognitive process commencing which might affect the N2pc waveforms substantially. Our empirical results suggest that TJ differences induced by key feature salience behave as illustrated in the second case.

The fact that the N2pc TJ differences between experiments do not seem to affect the offset is worth further elaboration. This observation suggests that key feature salience induced differences in the TAE being deployed are very short-lived, which might be due to a hard limit of processing in later time-intervals. To us, the most likely reason for such a hard limit might be that the later time-points of the N2pc (~400 ms) overlap with the beginning of the CDA component (Ikkai et al., 2010). This might be especially relevant since the N2pc and the CDA components are measured at neighbouring recording sites of the EEG. Therefore, one could argue that a secondary cognitive process (reflected in the CDA) might overlap in time with selective attention (reflected in the N2pc), which would make it difficult to see the

isolated offset of the N2pc. In this scenario, there might be N2pc TJ differences in the offsets between experiments, too. However, these would not be reflected in the EEG, since another process would obscure them. We stress the value for future research of pursuing this possibility, since such a pattern would imply that stimulus salience impacts the N2pc and the CDA component in a different manner. Such a finding would be striking, since it would diverge from the cascaded nature of the brain's processing pipeline (Mcclelland, 1979), i.e., that each cognitive process in the sensory pathway only starts when the previous process completes. The correlation between N2pc and P3 latencies demonstrated in Chapter 4 (Figure 30) can be viewed as indicating the cascadic nature of the brain's processing pathway.

With respect to the time-windows tested in the N2pc TJ analyses, the non-significant results of the main and simple effects of experiment testing the full N2pc are in line with the reasoning provided above. This is because the TJ differences in the N2pc's onset were likely not strong enough to lead to larger effects in the analyses testing the full N2pc. If there would also have been large N2pc TJ differences in the offsets between experiments, the experiment effects might also have yielded significant results for the full N2pc time-window. However, and particularly due to the likely start of a subsequent cognitive process overwriting the N2pc's offset, it is understandable that the results testing the experiment effects for the full N2pc were not significant.

N2pc TJ differences between response conditions: We demonstrated two statistically significant patterns when testing N2pc TJ differences between correct and intrusion trials: a significant main effect for the N2pc's onset and a significant simple effect for the full N2pc when only testing the data of Zivony and Eimer's (2020a) Experiment 1. Both significant effects were in the expected direction based on the ERP and vERP patterns presented thus far, i.e., that intrusion trials lead to more variable TAE deployment than correct trials. This pattern was expected since intrusion trials lead to lower amplitude and broader N2pc components (see Figure 73), which is replicated by the 2f-ST^2 model (see Figure 74, for example). Our understanding of this finding, stressed repeatedly in previous chapters, such as in our discussion of the loss of responsiveness phenomenon, is that intrusion percepts are the result of ambiguous trials, in which the target stimulus generally has a weaker representation (a weak key target type in 2f-ST^2 model terms). Therefore, the temporal dynamics with which this ambiguity is resolved varies across trials, depending on the characteristics of stimuli and their timing (e.g., how strong stimulus representations are on a particular trial and what the temporal delay between the target's and neighbouring distractors' response features is in

relation to the target key feature). In contrast, correct trials typically occur after a target stimulus has a strong key feature representation, TAE deployment thus occurs earlier, and the target response feature's representation is strong at the time of TAE deployment. Correct trials are accompanied with less ambiguity and, as a result, with more similar time-courses (as reflected in decreased N2pc TJ). Moreover, the fact that the full N2pc yielded a significant simple effect of response condition for Experiment 1 could suggest that the overall wider N2pc component present in this experiment allowed for TJ differences between conditions to be reflected in the analysis of the full N2pc.

The 2f-ST^2 Model with the Addition of Gamma Noise

In this chapter, we presented the most complete version of the 2f-ST^2 model after implementing Gamma Noise to the model's blaster firing latencies. This version of the 2f-ST^2 model thus refines the cognitive dynamics related to the distractor intrusion phenomenon compared to the version presented in Chapter 4. These analyses provide evidence in favour of Hypotheses 2 and 3 of this thesis, i.e., that the 2f-ST^2 model accounts for a broad range of findings obtained with distractor intrusion experiments as well as that less temporally variable TAE deployment after increased key feature salience can be computationally modelled with the 2f-ST^2 model. Importantly, we implemented Gamma Noise at an early point in the 2f-ST^2 model's processing pipeline, i.e., modulating τK. We acknowledge the cascaded nature of our computational model, leading to early latency or jitter differences translating to later processes of the 2f-ST^2 model. Nonetheless, we stress that the suggested N2pc TJ differences in the component's onset are in line with modelling Gamma Noise to impact processing early. For example, if, for whatever reason, we would have demonstrated N2pc TJ differences between experiments in the component's *offset* only, it would have been more appropriate to add Gamma Noise to a later stage of the 2f-ST^2 model's processing pipeline.

It is worth noting that we did not account for the observation discussed in the previous section that the TJ differences between experiments did not seem to affect the N2pc's offsets. Whilst acknowledging this incompleteness in our implementation of Gamma Noise to the 2f-ST^2 model, we stress that it is unclear whether the N2pcs did not show TJ differences in the offsets because these were absent in the brain or because the signal was overwritten by a subsequent cognitive process, for example that indexed by the CDA component. Therefore, we argue that further empirical work, providing a more complete understanding of TJ differences in the N2pc offsets, is required before potentially modifying the 2f-ST^2 model to reflect N2pc offset dynamics.

To conclude, in this chapter we presented the most thorough 2f-ST^2 model. Compared to the version presented in Chapter 4, the current version additionally simulates that in human selective attention, differences in key feature salience are proposed to lead to differences in the temporal variability of TAE deployment after the detection of task-relevance. We furthermore empirically assessed the latter interaction, contrasting the temporal jitter of the N2pc components of Zivony and Eimer's (2020a) Experiments 1 and 2, which included differently salient key features in their RSVP streams. Our results suggest that more salient key features lead to less temporally variable onset-latencies of the N2pc. The N2pc's offsets were not affected by changes in key feature salience to the same extent. We finally discussed this finding in detail and further elaborate on why our implementation of Gamma Noise to the 2f-ST^2 model conceptually simulates the empirical findings accurately.

Chapter 7 – SVM Decoding of Temporal Event Integration in the Attentional Blink

Introduction

The previous three chapters have investigated the binding of multi-dimensional stimulus features in RSVP paradigms into single percepts in the context of the distractor intrusion phenomenon. In this context, the process of binding stimulus features into single percepts is suggested to occur between 300 – 450 ms after the target stimulus was presented. Evidence in favour of the latter is provided by Botella's (1992) reaction time (RT) data, in which responses, which necessitate the completion of binding, were provided 375 – 425 ms after the target was presented (Figure 23). Moreover, the ERPs of Experiment 1 of Zivony and Eimer (2020a) are further indicative of the proposed time-interval, since the authors found the N2pc and P3 components to peak at around 300 and 400 ms, respectively (Figure 25). Note that we would argue that the P3 component more accurately reflects the time-course of the feature-binding process itself, since the N2pc indexes the deployment of TAE, which precedes the binding process. The 2f-ST^2 model's vERPs are also in line with these empirical findings, generating vP3s that peaked at around 400 – 500 ms (see Figure 34 & Figure 36). To reiterate, the 2f-ST^2 model's vP3s encompass both pathways' TFL/ type layers as well as both binding pools and are thus closely related to the time-course of types (i.e., features in model-terms) being bound to the currently active token. Note that our computational model's vP3s unfolding later than their human equivalents is not surprising, since we know that processes in ST^2 models (Bowman & Wyble, 2007; Chennu et al., 2011; Wyble et al., 2009) occur slower

than equivalent processes in the human brain. This is also reflected in the 2f-ST^2 model's RT values (Figure 23), which qualitatively show the same pattern as the human RTs found by Botella (1992), but are numerically about ~150 ms larger than their human equivalents. In this chapter, we will study temporal event integration, which is another case in which human cognition binds, or integrates, target-relevant stimulus information into a single percept. Further, and even though the current chapter's analyses will investigate the binding of stimuli's Gestalt properties (Von Ehrenfels, 1937; Wagemans et al., 2012) in the context of temporal event integration instead of stimuli's multi-dimensional features (as was the case with distractor intrusions), we will provide evidence in favour of Hypothesis 4 of this thesis, i.e., that the temporal integration of two combinable stimuli into a single perceptual event occurs with temporal and electrophysiological characteristics that are consistent with the distractor intrusion phenomenon and the 2f-ST^2 computational model.

Temporal event integration is a phenomenon in which visual target stimuli presented in quick succession are regularly combined into a single episode of memory when there is a perceptually meaningful way of doing so. The present chapter aims at discovering the temporal and spatial unfolding of temporal event integration, or 'joint tokenization', using multivariate pattern analysis (MVPA) by applying the temporal generalisation method to EEG data. Specifically, we intend to replicate Akyürek et al.'s (2017) main findings on a whole-scalp basis and extend the author's work by analysing all electrodes using MVPA, rather than target ERP components. In more general terms, we plan to resolve the question of exactly when and where the human brain merges two quickly shown visual stimuli into one perceptual event. To this end, we will first provide a detailed introduction to Akyürek et al.'s (2017) experimental paradigm, electrophysiological analysis, and results, before presenting the main methods and results of our MVPA analysis.

Akyürek et al.'s (2017) electrophysiological analysis was conducted on the level of event-related potentials (ERP). Group-level ERPs (or Grand Averages) are computed by first taking the average across-trial EEG signal within subjects and conditions and afterwards collapsing these single-subject ERPs across subjects. Specifically, three ERP components were scrutinized, each argued to reflect distinct cognitive processes: The N2pc, a difference wave, computed by subtracting activity captured by electrodes PO7 and PO8, believed to index attentional processing (Eimer, 1996; Kiss et al., 2008; Luck & Hillyard, 1994); the CDA, another difference wave, calculated at electrodes P7 and P8, is proposed to be a measure of WM load (Ikkai et al., 2010) and the P3 component, measured at Pz, is believed to

reflect WM consolidation (Polich, 2007). Both difference waves are generated as the difference between contralateral (opposite) and ipsilateral electrode sites with respect to the visual field that contained the target stream. Figure 80 - Figure 82, adopted from Akyürek et al. (2017), show these ERP components for the three experimental conditions: Single-target; two-target, integration (henceforth called integration); and two-target, no integration trials are represented as green, red and blue lines, respectively. The shaded areas next to the ERP components mark the within-subject standard error and dashed windows indicate the analysis time-frame of the linear mixed-effects models the authors (Akyürek et al., 2017) used for statistical data analysis, which are furthermore asterisked if a significant difference between conditions was evident within them.

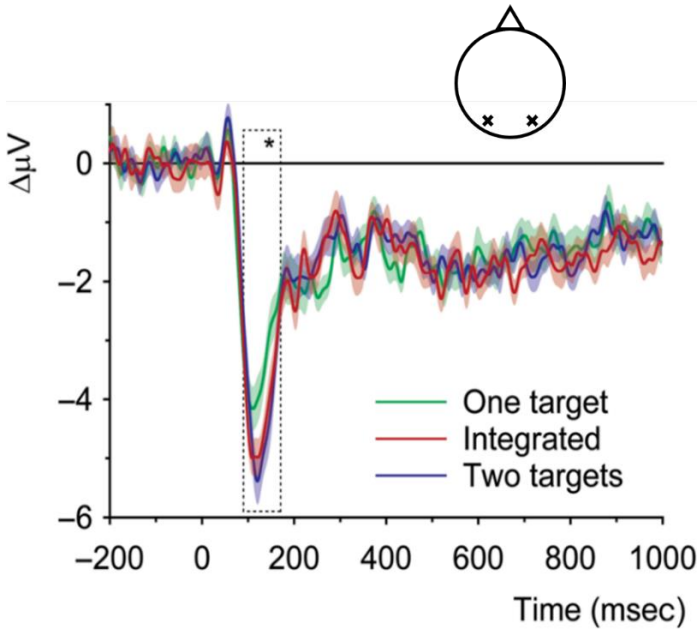


Figure 80. N2pc difference waveforms, measured at PO7/PO8 in micro-volt, as a function of time. Different lines correspond to the experimental conditions of one-target (green), two-target, integration (red) or two-target, no integration (blue). Location of electrode sites on scalp is indicated above ERP timeseries. The dashed window shows the time-frame of statistical analysis, asterisked if significant, and shaded areas indicate within-subject standard errors. Adopted from Akyürek et al. (2017).

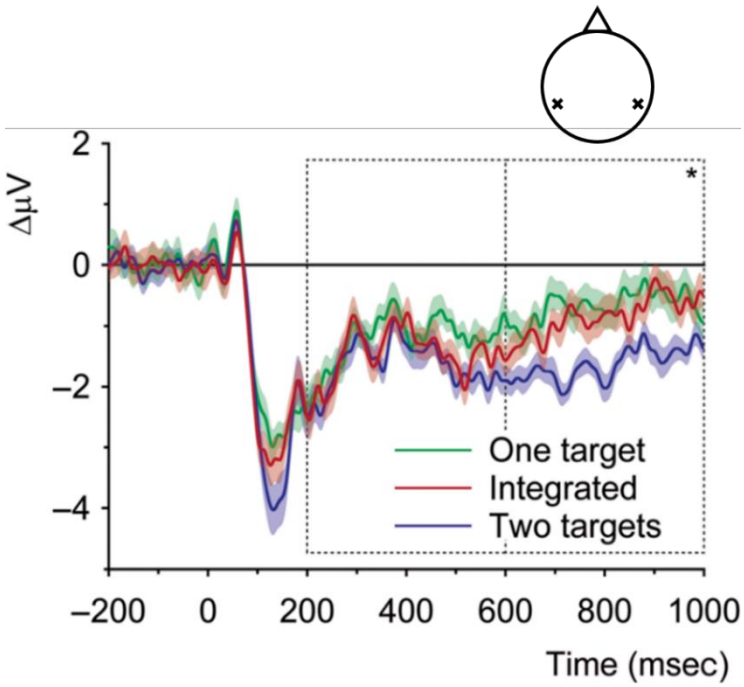


Figure 81. CDA difference waveforms, measured at P7/P8, as a function of time. Plotting conventions are identical to those of Figure 80. Adopted from Akyürek et al. (2017).

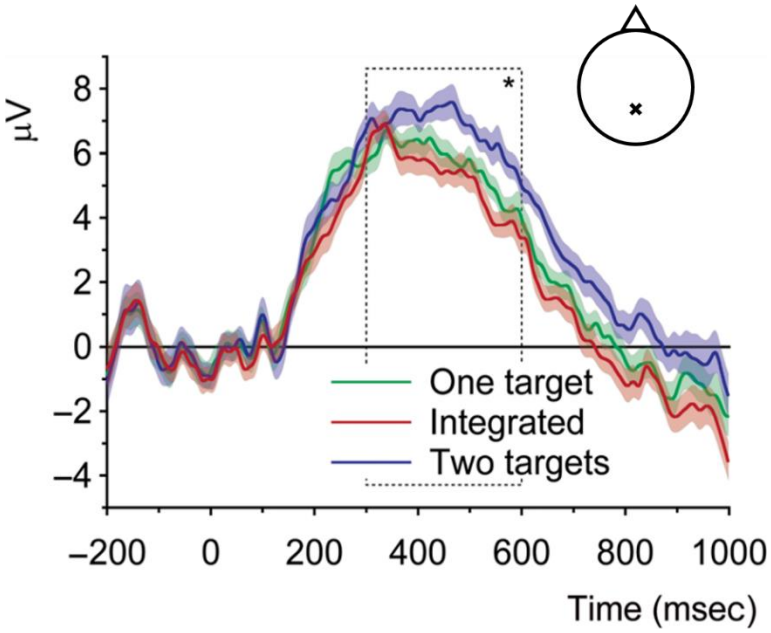


Figure 82. P3 waveforms measured at Pz as a function of time. Plotting conventions are identical to those of *Figure 80*. Adopted from Akyürek et al. (2017).

Figure 80 - Figure 82 illustrate the main findings of Akyürek et al. (2017) concerning the attentional as well as working-memory processes underlying temporal event integration. As can be seen in Figure 80, N2pc amplitude depended on whether participants were presented with one or two target stimuli, being increased for two-target trials. This indicates that early attentional processes required more neuronal activation to process two targets,

irrespective of whether temporal integration occurred or not. This is in stark contrast with the late CDA and the P3 components, as illustrated in Figure 81 & Figure 82, which were substantially reduced in amplitude in the case of a single percept being processed, regardless of the actual number of targets displayed. This observation is pivotal, as it suggests that the process of merging two target stimuli into a single representation alleviates WM load (CDA) and consolidation costs (P3) tremendously, making both akin to the case of a single target being processed, even though two targets were presented in reality. In summary, the authors (Akyürek et al., 2017) main finding thus was that the extent of early attentional brain processes depended on the number of target stimuli presented, whereas later working-memory (WM) related ERP components were modulated solely based on how many target stimuli were perceived, irrespective of how many were actually presented. The latter finding further implied that the process of temporal event integration mitigates the cost of encoding two targets into WM when these can be combined into a single percept in a meaningful way.

These intriguing results motivated us to investigate temporal event integration further. We were particularly interested in investigating the whole-scalp dynamics associated with the phenomenon and in analysing the temporal characteristics of potential brain-pattern differences between experimental conditions. This chapter will hence commence with an introduction of the analysed data, followed by a presentation of our methodological approach. Our results will be presented next with a specific focus on linking them to Akyürek et al.'s (2017) findings. The chapter will conclude with suggestions for future research.

Methods

Akyürek et al.'s (2017) Experimental Paradigm & Data

Akyürek et al. (2017) designed a dual-stream RSVP paradigm and recorded the EEG data of 39 participants. Participants were asked to attend to the RSVP streams and identify target stimuli after each trial. The design of a typical trial is displayed in Figure 83. Each RSVP stream consisted of 16 stimuli, shown for 70 milliseconds (ms), and including an inter-stimulus interval (ISI) of 10 ms. Distractor letter stimuli were sampled from the complete alphabet and appeared randomly with equal probability. The latter was likewise true for target stimuli, which consisted of combinations of one to four corner lines. The same corners were, however, never presented as subsequent targets within two-target trials, as this would have thwarted the possibility of merging their representations into a single percept. Second targets, if present, always followed immediately (lag 1). Two target trials were considered correctly identified irrespective of targets' order. Participants were instructed to guess whenever unsure

and no feedback was provided concerning their performance on identifying targets (Akyürek et al., 2017). The experiment consisted of two-target (70%), one-target (20%) and catch (10%) trials. In catch trials, the fixation dot changed its appearance from an apostrophe (‘) to a quotation mark (“). This happened at sixth or eight position of the RSVP stream and was hypothesized to enhance participants’ central fixation. The order of trials was randomized and only trials in which correct responses (including correct integrations) were provided were included in further analyses.

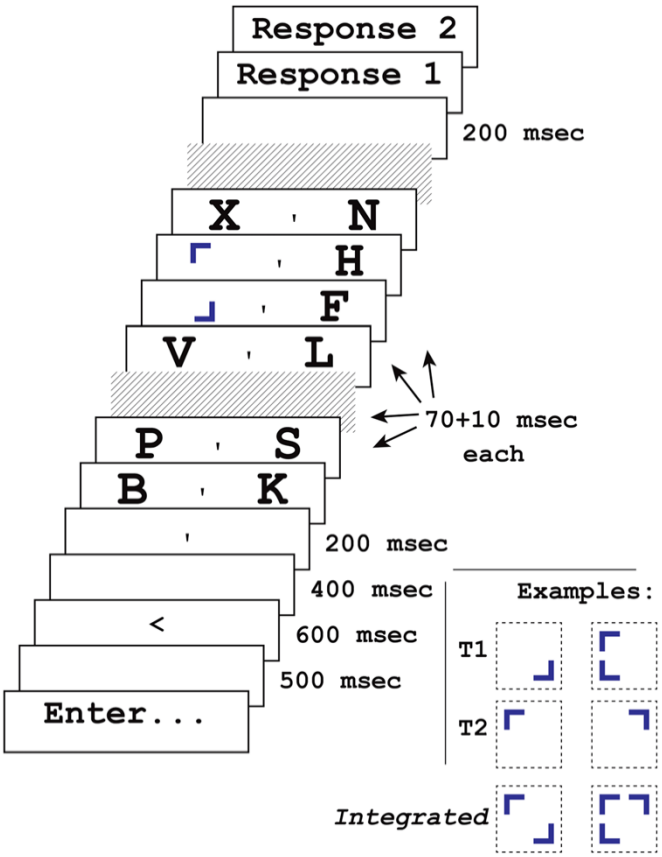


Figure 83. Design of an experimental trial.

Trials were self-initiated and commenced with a cue-stimulus, followed by a fixation dot. Subsequently, the two synchronous RSVP streams started, each comprising 16 stimuli. Alphabetical letters served as distractor- and blue corner lines as target-stimuli. Trials concluded with two response prompts. The bottom right examples illustrate how targets might have been merged in a perceptually meaningful way into a single representation. Adopted from Akyürek et al. (2017).

Electrophysiological (EEG) analysis

The EEG signal was recorded at 500 Hz with a 64 electrode-cap, eye movements were measured via electrooculography (EOG) and an average referencing procedure was used. Subsequent offline pre-processing included bandpass-filtering the signal between 0.1 and 40 Hz and time-locking EEG segments around the onset of the second target, from 200 ms before its onset (referred to as -200 ms hereafter) to 1000 ms post stimulus-onset. Single-target trials

were similarly segmented around the onset of distractor stimuli that followed immediately (lag 1). Furthermore, the pre-stimulus period of -200 ms to 0 ms was used for baseline correction (Akyürek et al., 2017). For a more detailed description of the EEG methodology, see the original paper (Akyürek et al., 2017).

Since Akyürek et al.'s (2017) EEG methodology was specifically tailored to the analysis of the three ERP components presented above, we implemented a modified pre-processing pipeline on the study's raw dataset, which was more appropriate for a whole-scalp analysis. We programmed our additional pre-processing pipeline based on Srivas Chennu's MOHAWK toolbox (https://github.com/srivaschennu/MOHAWK) and using the MATLAB extension EEGLAB.

Our pipeline first filters the raw EEG data before epoching trials, keeping the configuration of the original study (Akyürek et al., 2017), i.e., a 0.1-40 Hz bandpass filter and -200 to 1000 ms epochs. Subsequently, a variance measure is adopted which only suggests artifactual (i.e., bad) channels and trials. Artifacts in EEG analyses refer to cases in which the recorded EEG signal (in a trial or a channel) is comprised by noise, meaning voltages recorded by the electrodes that did not originate from brain activity, but from other sources, such as muscle movement. Bad channels in this context thus means channels that malfunctioned or for which the signal recorded was not as clean as desired. Bad trials refer to trials in which participants moved or blinked with their eyes, which overwrites the brain-activity recoded by electrodes drastically. Channels and trials which were suggested to be bad were inspected manually and removed if they seemed to indeed have been of artifactual nature. At this stage, we adopted a very generous approach and only removed channels and trials if they were clearly due to artifacts, since a secondary manual inspection step was performed later.

We then performed Independent Component Analysis (ICA) on each subject's dataset in order to correct for artifacts, especially those induced by eye-blinks. ICA is a well-accepted computational method in neuroimaging, and signal processing in general. It involves the decomposition of a signal into statistically independent components. In neuroimaging, ICA has been particularly useful for artifact correction since it allows the removal of the signal's artifactual components without the necessity to remove trials altogether. For example, if a subject had strong activation in their jaw muscles during the experiment, this would be reflected in their EEG signal, especially in temporal electrodes that are located close to the jaws. Removing all trials in which this muscular activity strongly affected the EEG recording

might imply the exclusion of most trials of this subject, which naturally is undesirable since the less trials go into a statistical analysis, the lower its statistical power. Instead, one can run an ICA to only remove the component that captured the jaw-muscle artifact, while preserving the subject's trials and, thus, statistical power. In general, ICA decomposes the signal into independent components, which enables the experimenter to remove these components from the signal before transforming it back to the signal's original dimensions. ICA has been used by the neuroimaging community for artifact correction for a substantial amount of time, being adopted for the removal of eye (Mennes et al., 2010) and muscle (Safieddine et al., 2012) artifacts or for real-time artifact correction of an EEG signal while performing fMRI scans (Mayeli et al., 2016). In the current pre-processing pipeline, we specifically focused on removing independent components from subject's data if components demonstrated clear eye-blink, lateral eye-movement, or muscle-activity artifacts. This was done with caution, with components being kept in the data whenever we were in doubt whether they reflected artifacts or not, as the careless removal of independent components can easily remove the brain-signal of interest, too, to some extent.

In the next step, we corrected for missing channels that were removed prior to running the ICA by means of spherical spline interpolation (Perrin et al., 1989), which is a method that estimates a respective channel's activation based on its adjacent channels' activation values. Running ICA on the reduced dataset, i.e., with bad channels being removed, and interpolating these channels after ICA is recommended, since this approach allows the ICA algorithm to better identify, or decompose, the signal into independent components. We subsequently re-referenced the EEG signal to the common average across the scalp, which implies computing the average signal across all EEG electrodes and subtracting this average from each individual electrode at each time-point. Re-referencing to the common average across the scalp reduces the signal's amplitude overall, as it means that at each time-point, the sum of electrodes' activation is equal to zero. However, and importantly, it further leads to channels' signals contributing equally to the overall signal across the scalp, thus making signal amplitudes comparable across different channels. We then applied an amplitude threshold to trials, which rejected trials whenever an amplitude higher than 125 or lower than -125 microvolts was observed at any electrode. ICA supposedly is not too suitable to detect such artifactual amplitudes in trials, since such large amplitudes might only occur for a very brief moment in the signal, thus not constituting sufficient regularity for an independent component to be computed based on them. However, particularly since we used classification algorithms for the present analysis, which are very efficient at using anything in the signal

that allows classes to be differentiated, it was critical to remove trials with such moments of large artifactual amplitudes. Afterwards, we performed a final manual inspection of the dataset and removed any trials that still seemed artifactual in nature.

The final step in our pipeline dealt with the fact that the data we received was split based on whether target stimuli appeared in the left or right half of the experiment's RSVP stream. For the original study, this information was pertinent, as the N2pc and CDA components were calculated as the difference waves between ipsi- and contralateral electrode sites with respect to targets' locations (Akyürek et al., 2017). However, for the present analysis, which performed MVPA on a whole-scalp basis, it was desirable to combine experimental conditions of targets appearing in the left or right visual field in an appropriate manner. Hence, a function was programmed on top of EEGLAB, which exchanged the activation vectors of corresponding lateral electrodes around the central electrode line (Fpz to Iz) for all trials in which the target appeared in the left visual field (LVF). For example, this function placed the LVF data of electrode Fp1, which is recorded in the first row of EEGLAB's data structure, into its second row, which contains Fp2's data, and vice versa. It should be noted that Fp1 and Fp2, like all other electrodes for which this procedure was applied to, are located on the same position on the scalp, one left and the other on the right side. Implementing this step therefore simulated that all targets appeared in the right visual field. It should be noted that this procedure assumes symmetry of the cerebral hemispheres in how visual stimuli are processed. Concretely, it implies that a target presented in the left visual field evokes responses in ipsi- and contralateral cortical areas that are comparable to those following a right visual field target. This is generally assumed to be true based on numerous magnetic resonance imaging (MRI) studies that worked on retinotopic mapping, or how our visual fields map to visual cortex (Bridge, 2011).

Temporal Generalisation

Group-level temporal generalisation analyses were performed on the EEG data of 27 participants to decode differential brain activity over the whole scalp for three contrasts of interest: one-target vs. two-target, one-target vs. integration and two-target vs. integration. We performed temporal generalisation analyses using the MATLAB toolbox MVPA-Light (Treder, 2020).

Classification analyses were first conducted on the single-subject level, which has been recommended as it accounts for differences in participants' brain anatomy (King et al., 2014). To reiterate, for each participant and contrast of interest, 600 different linear SVM

classifiers were trained, one at every time-point of the time-series. For each one of these classifiers, the activation in microvolt across all 64 EEG electrodes at that given time-point of the trial's time-series was used as the training input. Classifiers were thus always trained on 64 input features to best separate two experimental conditions for a contrast of interest (the latter, for example, being single-target (1T) versus two-target, integration (INT) trials) at their respective time-point. The samples used for training and testing were comprised by the experimental trials of a given subject and a given contrast of interest.

Classifiers were then tested on their ability to differentiate, or decode, brain activity measured with EEG at all other time-points (horizontal decoding) as well as at the time-point they were trained at (diagonal decoding). Thus, classifiers generated an output vector of length 600 of area under the curve (AUC, introduced in more depth below) values informing about their efficacy of decoding class-differences in whole-brain activation patterns at all time-points of the time-series. Classifications were performed using 5-fold cross-validation, which describes a procedure that partitions a given dataset into 5 pieces. A classifier is trained on four of those parts and tested on the last. This procedure is repeated five times so that each subset served as the testing set once. The average classification performance over these five runs is finally reported as a measure of how well the classifier was able to distinguish classes. Each within subject and contrast classification was repeated twice and average area under the curve (AUC) values were computed. AUCs are a robust non-parametric measure of classification performance that are preferred to mean accuracy because they do not require assumptions about the data's underlying distribution. AUCs are the result of an empirical receiver operating characteristic curve (ROC) analysis, which can estimate the effect size of a Wilcoxon/Mann-Whitney U Test (Mason & Graham, 2002) between correct and incorrect predictions of the classifier. Therefore, an AUC of 50% coming from a two-class classification problem means that a given prediction is as likely to be correct as incorrect and 100% corresponds to perfect prediction (King et al., 2013). The resulting 600 by 600 AUC maps were finally averaged across participants, which led to the group-level temporal generalisation matrices depicted later.

Importantly, to avoid bias during training, the number of trials per class had to be equal for each individual, within-subject and contrast, classification. Thus, we had to decide whether to under- or oversample trials. The former implies that trials from the majority class are removed from the analysis until that class matches the number of trials of the minority class. The latter describes a bootstrapping procedure, where trials from the minority class are

used a random number of times, until the number of trials of the majority class is matched. This oversampling procedure was implemented within each cross-validation fold, which is required as performing the procedure before cross-validation could introduce dependencies between training and testing set. For example, a given trial could be included in the training set and its bootstrap-copy in the testing set. Such a scenario would wrongly inflate classification performance as it would make it easier for the classifier to differentiate between classes. We performed the analyses using under- as well as oversampling and the respective results were comparable, as can be seen in Figure 84. We decided to incorporate oversampling in our final analyses as it allowed for the preservation of experimental trials. It should further be noted that the TG maps plotted in Figure 84 were generated before implementing our own pre-processing pipeline presented above. These maps thus look slightly different than those we will present in the Results section, because the maps in Figure 84 classified data that contained substantial artifacts. We nonetheless present these maps here, as their purpose is only to illustrate the impact of under- and oversampling trials on resulting TG maps. We moreover decided against running these analyses again on the artifact-free datasets to avoid the computational demands and energy costs associated with doing so.

Even though each of the plots in Figure 84 show quite interesting and distinct classification patterns, naturally, valid conclusions about their meaning cannot be solely based on measures such as accuracies or AUCs. To this end, an accurate and powerful statistical test is required. We hence performed the peak-level as well as the cluster-extent permutation test detailed previously. It is worth noting that we did not consider the period from -200 to 100 ms for either permutation test. The motivation for this is two-fold. First, we had no reason to believe that effects between experimental conditions occurred in this time window, as the earliest differences shown by Akyürek et al. (2017) were reflected in the N2pc component and occurred after 100 ms (Figure 80). Moreover, excluding this time-window increased the test's statistical power.

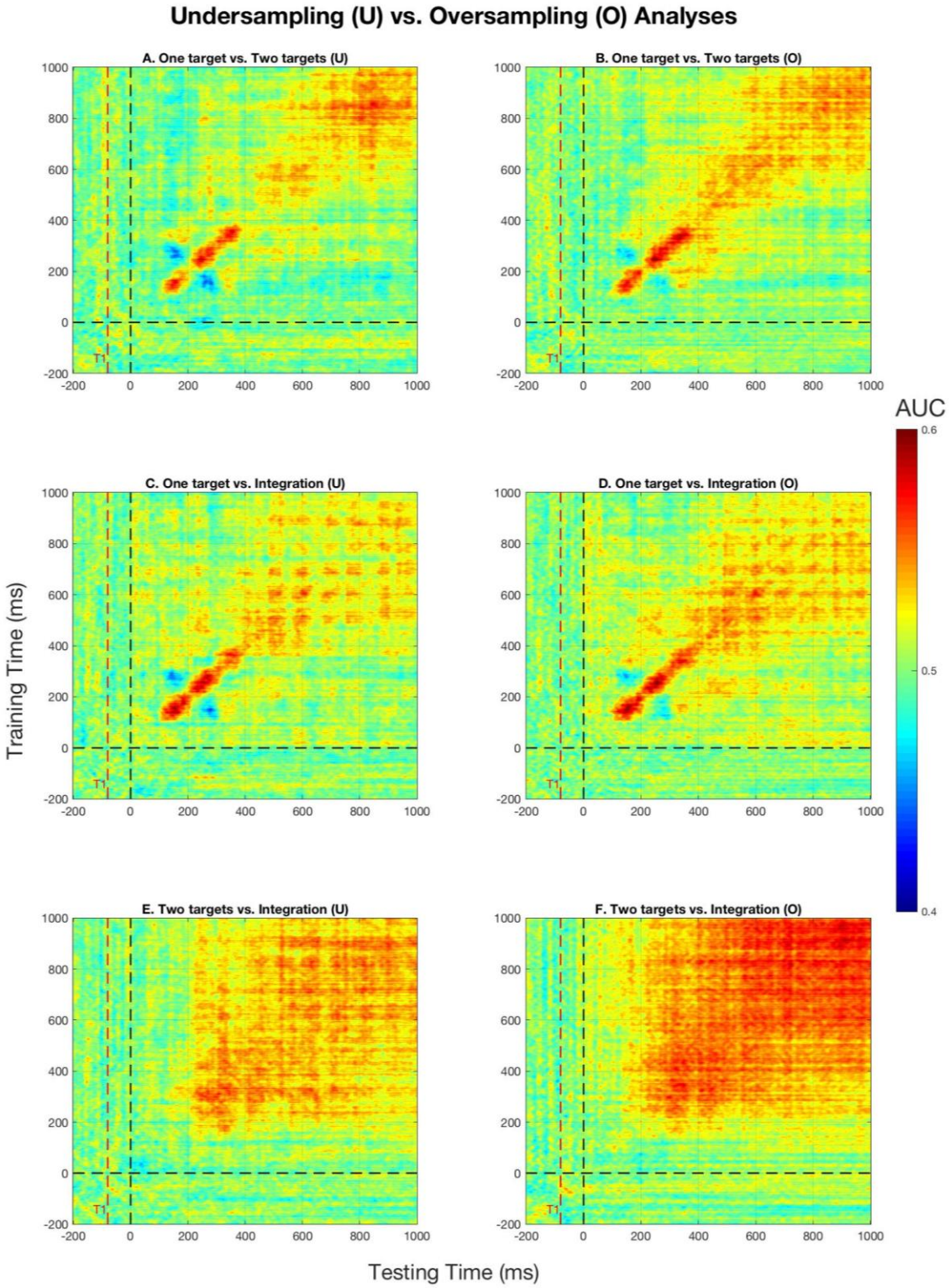


Figure 84. Temporal generalisation after under- and oversampling.

The top row (A & B) displays matrices contrasting one- with two-target trials, the middle row (C & D) shows the contrast of one-target and integration trials and the bottom row's matrices (E & F) depict temporal generalisations of the two-target versus integration contrast. Temporal generalisation analyses were carried out after balancing trials via under- (A, C & E) as well as oversampling (B, D & F).

All trials were segmented around the onset of the second target (time = 0). The onset of the first target is indicated by the vertical red line labelled 'T1'. Classifiers' training and testing times are reflected in y- and x-axes, respectively. Time is measured in milliseconds and negative values correspond to the time-window before the onset of the second target. Classification performance was assessed by area under curve (AUC) as displayed in the colour bar plotted on the right.

Results

Temporal Generalisation

Figure 85 depicts the results of applying TG to our three contrasts of interest: two-target, no integration vs. one-target (1T vs. 2Ts, left); one-target vs. two-target, integration (1T vs. INT, middle); and two-target, no integration vs. two-target, integration (2Ts vs. INT, right). Figure 85 presents the resulting TG maps without the application of any statistical test in Panel A and after running the peak-level and the cluster-extent sign-swap permutation tests in Panels B and C, respectively. The left TG map in Figure 85A (1T vs. 2Ts) hints at an early diagonal classification pattern between approximately 100 and 350 ms as well as a sustained classification pattern from 500 ms onwards. The classification map of 1T vs. INT is revealed in the middle map of Figure 85A. Again, an early, diagonal classification pattern can be observed, which is a bit shorter than that shown for 1T vs. 2Ts (left map) and ranges from roughly 100 to 300 ms. Finally, the right TG map of Figure 85 presents the contrast of 2Ts vs. INT. This map does not show the early classification pattern of the other two maps. Instead, a late classification pattern can be observed, similar to that of the 1T vs. 2Ts map (left). However, the late classification pattern of the 2Ts vs. INT map (right) is of lower amplitude and commences later (~600 ms) than the pattern found for the 1T vs. 2Ts map (left). The results of applying our peak-level permutation test to the three TG maps is shown in Figure 85B. Significant pixels are marked as black dots and were only found for the early classification pattern contained in the 1T vs. 2Ts and 1T vs. INT maps. Figure 85C presents the results of our cluster-extent permutation test. This test revealed one significant AUC cluster in the late classification pattern of the 1T vs. 2Ts map. These observations are overall in line with the original paper's (Akyürek et al., 2017). Finally, Figure 85D plots the main diagonal classification time-series, i.e., AUC values after testing classifiers on data that was recorded at the same time-point each classifier was trained at.

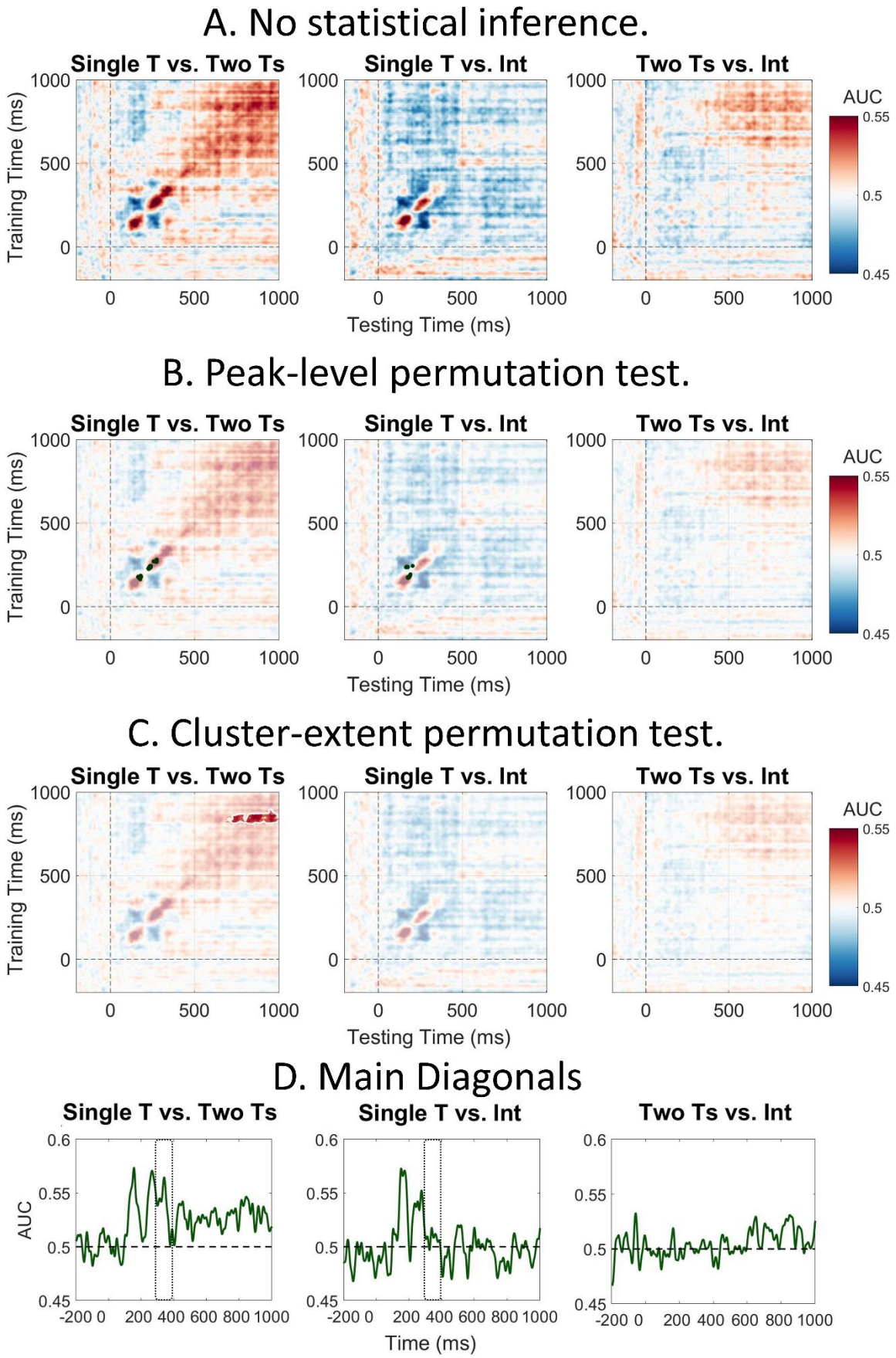


Figure 85. Temporal generalisation maps. The 1T vs. 2Ts, 1T vs. INT and 2Ts vs. INT contrasts are presented in left, middle and right maps. Panel A shows the TG maps without the application of any statistical test. Panels B and C present TG maps as well as the results of our peak-level and cluster-extent permutation tests, respectively. Panel D plots only the main diagonals (i.e., training being equal to testing time). Black rectangles indicate approximate time at which we suppose temporal integration to commence. Time is measured in milliseconds (ms) and classification performance in area-under-curve (AUC).

The early diagonal classification pattern observed whenever single-target trials were compared to two-target trials, irrespective of the presence of temporal integration, supports the notion that only the number of target stimuli presented within a trial matters for attentional processes, which are at work during this time-frame. This fits the N2pc differences shown in Figure 80 (Akyürek et al., 2017). In this context, it is striking that the classification pattern we observed is more than 100 ms longer than the N2pc component, which was observed from 100 to 200 ms. The late classification patterns evident in the 1T vs. 2Ts and 2Ts vs. INT maps complement the original ERP study (Akyürek et al., 2017), too, as the original study demonstrated WM to be affected by temporal integration between 300 – 600 ms (P3) as well as 600 – 1000 ms (CDA). These time-windows coincide neatly with that of our late classification pattern, especially regarding the CDA component. The late classification pattern therefore supports Akyürek et al.'s (2017) claim that later, WM related, cognitive processes depend on how many perceptions need to be encoded into WM, regardless of how many target stimuli really were presented. With respect to narrowing down the time-interval of temporal event integration occurring in the brain, especially the main diagonal time-series plotted in Figure 85D suggest that stimuli are merged into a single percept after 300 ms. We added black rectangles to the main diagonal time-series in Figure 85D's left and middle column, which indicate that classifiers decoded class-differences in the 1T vs. 2Ts, but not the 1T vs. INT contrasts 300 ms after the second target was presented. This failure of classification is hence indicative of whole-brain activation patterns of INT trials approximating those of 1T trials, meaning that integrated percepts evoked a brain response similar to that of a single target during this time-interval. Next, we will present ERP topographies that narrowed down this temporal estimate of event integration occurring further.

ERP difference topography maps

Additionally, ERP difference topography maps (Figure 86 & Figure 87) were generated to ascertain those cortical areas that potentially enabled classifiers to distinguish conditions. For these, within-condition Grand Average ERPs were computed for all electrodes. Then, three difference ERPs were computed for each electrode by subtracting the grand-average ERPs of two conditions at a time (2Ts – 1T, 1T – INT and 2Ts - INT). In both figures, topographical maps were plotted for seven time-points of interest. These plots were generated by computing the average signal across a given time interval. For example, the 600-800 ms topographies of Figure 87, correspond to the average signal between 600 to 800 ms. Also, all trials' data was utilized for these plots without the application of any balancing

procedure to match the number of trials between conditions. We present the contrast topographies for the time-range of 320 to 476 ms in Figure 86, with time progressing from bottom to top in these plots. This particular time-range was chosen as that of main research interest, since our results suggested that temporal integration likely occurs somewhere during this time-window (explained below). Moreover, the topographies of Figure 86 incorporated time-steps of 26 ms and hence show the average activation differences between conditions during respective time-intervals. A few interesting patterns can be observed, which again generally support the work of Akyürek and colleagues (2017).

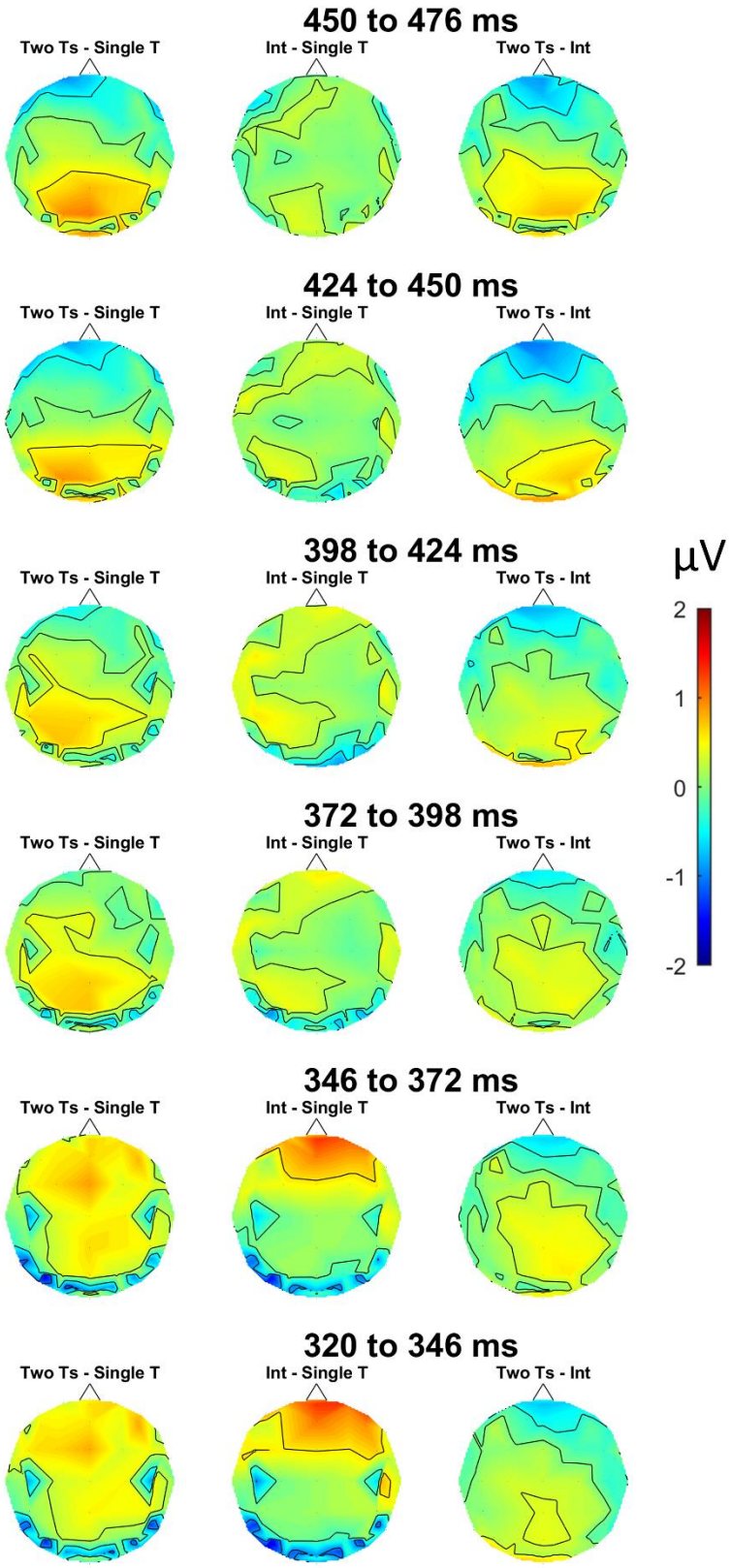


Figure 86. ERP difference topography maps.

The time series evolves from bottom to top, spanning the period 320 to 476 ms. The columns correspond to the three differences of interest: two targets minus one target (left), one target minus integration (middle) and two targets minus integration (right). Difference in ERPs is measured in microvolt, as indicated by the colour bar on the right. Topographies show the average activation across respective time-intervals. The frontal lobe is located on top in each map.

Based on the authors' findings, one should expect the integration (INT) condition to resemble the two-target (2Ts) at early and the one-target (1T) condition at later time-points in these plots. Thus, the 2Ts – 1T topographies (left) should be akin to those of the INT – 1T (middle) topographies at early time points. Comparing the left and middle column of Figure 86, we can confirm this expectation, as the topographies look comparable until 398 ms, which is when they start to differ. This is furthermore the time-point when differences in the 2Ts - INT topographies (right column) start to emerge, which again fits the idea that early in the trial, temporal integration has yet to happen and thus integration trials evoke a similar kind of brain response as perceptions of two separate targets. The second prediction is that the process of temporal integration should be finished at later time points, which should lead to integration trials being comparable to one-target trials in terms of topographical brain response. Therefore, the 2Ts – 1T topographies (Figure 86, left) should resemble the 2Ts – INT topographies (Figure 86, right). This resemblance can be seen in the top two rows of Figure 86, i.e., from 424 ms onwards. In this later time-period, both contrast topographies exhibit a negativity over frontal and a positivity over occipital cortical areas. For the sake of thoroughness, we also present the three contrast topographies across the whole trial (i.e., -200 to 1000 ms) in Figure 87. Figure 87's topographies again support the notion that the brain-activation pattern of INT trials resembles that of 2Ts trials at early and that of 1T trials at later time periods.

It is worth stressing that we did not implement inferential statistical tests on these topographical plots because we wanted to prevent the practice of double-dipping (Kriegeskorte et al., 2009), especially since our time-interval of interest in Figure 86 was inspired by the classification results presented in Figure 85. Even though the topographies thus constitute rather exploratory analyses, we would argue that their amplitude differences during the time-intervals of interest likely are informative. We particularly base this statement on the observation that the relative variability of amplitudes during the baseline window of -200 – 0 ms is low compared to the amplitude variability during the time-intervals of interest (see the original study's ERPs in Figure 80 - Figure 82 as well as the bottom row of Figure 87 for the baseline window).

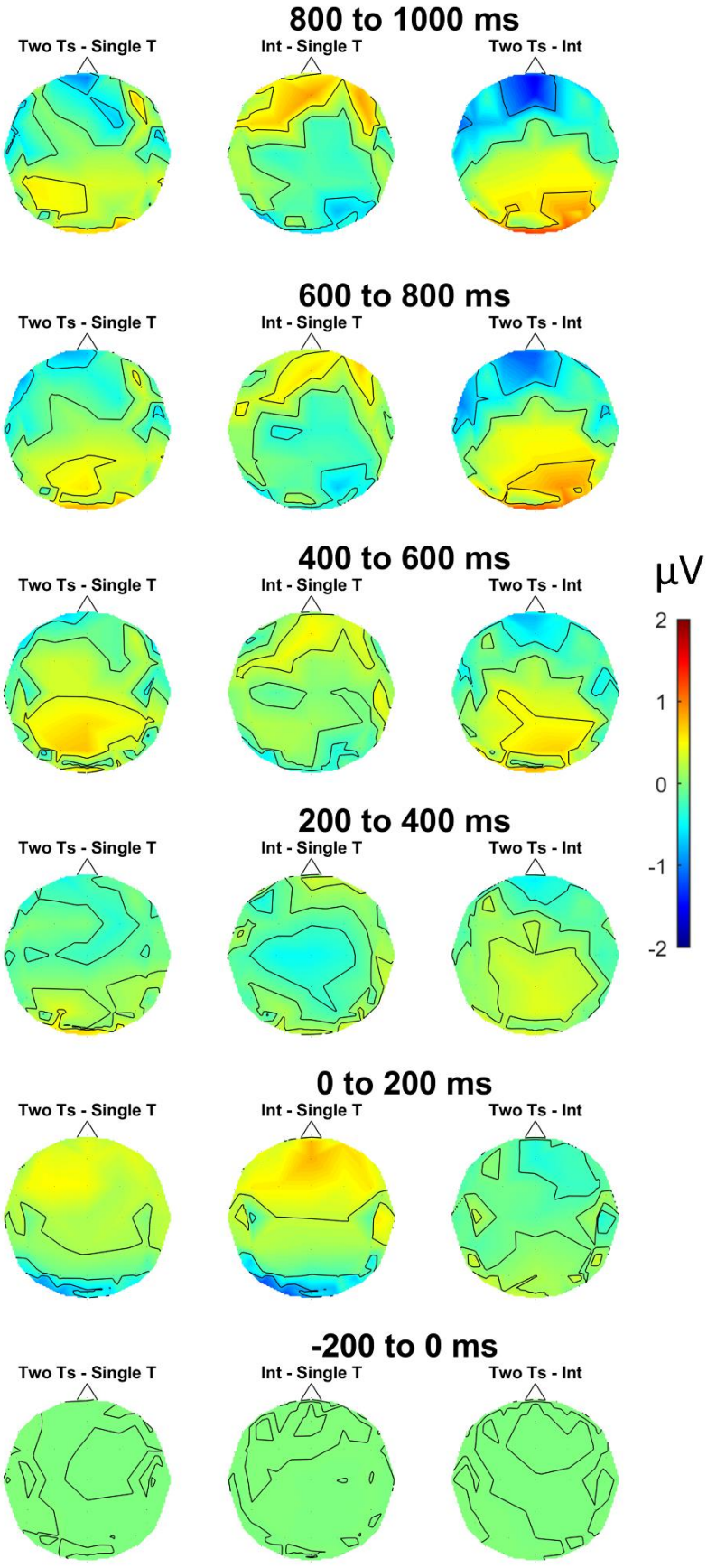


Figure 87. ERP difference topography maps.

The time series evolves from bottom to top, spanning over the period of the whole trial in steps of 200 ms. All other plotting conventions follow those of Figure 86.

Discussion

This chapter introduced a gap in neuroscientific research regarding temporal event integration in the attentional blink and presented how we added to the contemporary state of research by applying MVPA to Akyürek et al.'s (2017) EEG data. Our results replicated the main findings of Akyürek et al. (2017) on a whole-scalp basis. We especially found further support for the author's (Akyürek et al., 2017) claim that early attentional processes are modulated by the number of presented items, whereas later, WM-related processes depend on the number of perceived items. In addition, our temporal generalisation as well as ERP topography results propose temporal event integration to occur approximately between 350 – 450 ms after the second target is presented. Importantly, this is in line with the results concerning feature binding presented in previous chapters. This chapter therefore provides evidence in favour for Hypothesis 4 of this thesis, i.e., that the temporal integration of two combinable stimuli into a single perceptual event occurs with temporal and electrophysiological characteristics that are consistent with the distractor intrusion phenomenon and the 2f-ST^2 computational model.

Nonetheless, we acknowledge the potential for a more accurate analysis of how exactly temporal event integration unfolds temporally as well as spatially in the human brain and therefore provide suggestions for future research next.

Stimulus based classification

In the context of applying MVPA to data from RSVP experiments such as that of Akyürek et al. (2017), a potentially very useful approach would be to train classifiers on individual target stimuli. The underlying idea is to train separate classifiers on single-targets and their complementary two-target trials, staying within a given set of target stimuli. For example, one could train a classifier on trials in which a bottom left and bottom right corner was presented as the first and second target, respectively. A different classifier would then be trained on trials in which a combination of bottom left and right corners was presented as a single target. Both classifiers would finally be tested on their ability to decode trials in which these targets were presented in the context of all three experimental conditions. One could further perform this analysis for each combination of target stimuli individually and in the end assess the average time-series of classification performance over a trial. This should provide a more accurate estimate of temporal integration's temporal dynamics than that presented with the current TG analyses, since different types of target stimuli would no longer be collapsed across. Therefore, differences in brain activity based on different kinds of target stimuli being

presented would be factored out, which should result in a more accurate estimate of the time-course of temporal event integration.

Locating temporal event integration - source localisation and a multivariate alternative

The ERP topographies (Figure 86 & Figure 87) provided a valuable initial understanding about the spatial dynamics of temporal event integration. There are, however, methods that localise brain activity differences more accurately. One such method is source localisation, which computationally aims to find those sources in the brain that led to the observed sensor-level (in this case measured by electrodes) differences. Source localisation uses thousands of virtual dipoles and positions these around a template brain volume in which the sensor level data has also been projected. Different kinds of algorithms are then applied to estimate which combination of dipoles can best explain the sensor-level data. This allows inference about which cortical structures most likely generated the sensor-level observations.

Moreover, due to the fact that our classifiers' weights correspond directly to individual electrodes, it is tempting to generate topographic plots of these weights in order to assess which features, or electrodes, were most useful for the classification task. However, this would not validly contribute to finding out where in the cortex temporal integration occurs, as classifiers trained on EEG data inherently contain signal as well as noise in their weights. Haufe and colleagues (2014) addressed this issue by introducing a way of transforming backward, e.g. classification, models (extracting factors from data) into forward models (expressing data as a function of factors). The authors demonstrate, using simulated experiments, that their approach reliably decouples noise from the signal of interest in classification weights and hereby yields activation values, which are, in contrast to classifier weights, physiologically interpretable (Haufe et al., 2014).

In the current and previous three research chapters, we have analysed two cases of event integration, either binding stimuli's features or Gestalt properties, in rapid stimulus streams. The final research chapter of this thesis will be presented next and show some methodological risks when adopting machine learning algorithms to neuroimaging datasets. It will hence investigate Hypothesis 5 of this thesis, i.e., that machine learning algorithms applied to neuroimaging datasets, particularly temporal generalisation analysis used to determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfitting to hyperparameters) classification results if appropriate preventive measures are not implemented.

Chapter 8 – I Tried a Bunch of Things: The Dangers of unexpected Overfitting in Classification of Brain Data

Introduction

Machine learning has been transforming the approach to data science across a variety of research disciplines. The proper utilisation of computer algorithms to decode (as well as encode, such as in the context of generative models) the signal of interest contained in a dataset has been yielding novel kinds of results, leading to intriguing findings. An illustration of the application of machine learning to probe research questions that have not been able to be asked previously in this manner was provided in the previous chapter. Using the temporal generalisation method, we assessed the temporal extent of whole-scalp pattern differences in the context of Temporal Event Integration. In the context of neuroimaging, probing the temporal extent of whole-scalp pattern differences across conditions was much more challenging prior to the adoption of machine learning, justifying the increasing popularity of such methods in science. We provide metrics in Figure 88 that illustrate the increasing popularity of machine learning in the Neurosciences, specifically. Figure 88 was obtained (on the 29th of September 2021) using Dimensions AI (https://www.dimensions.ai/) and presents the publication count of academic papers that included the keyword "machine learning" in their title or abstract, restricting the search to the research categories of "Neurosciences" and "Cognitive Sciences" and the years to 2012-2021. Moreover, many pieces of computer software have been developed over the past years to enable the easy adoption of machine learning algorithms to neuroimaging data, demonstrating the high demand for the latter by the field. Examples include MVPA-Light (Treder, 2020), which we used for our analyses, the Amsterdam Decoding and Modelling Toolbox (ADAM, (Fahrenfort et al., 2018)), the Neural Decoding Toolbox (http://www.readout.info/), the Decision Decoding ToolBOX (DDTBOX, https://link.springer.com/article/10.1007/s12021-018-9375-z/), CoSMoMVPA (http://cosmomvpa.org/cimec2016_intro.html), PyMVPA (http://haxbylab.dartmouth.edu/publications/HHS+09a.pdf), Princeton MVPA (https://github.com/PrincetonUniversity/princeton-mvpa-toolbox), The Decoding Toolbox (https://sites.google.com/site/tdtdecodingtoolbox/) and PRoNTo (http://www.mlnl.cs.ucl.ac.uk/pronto/index.html).

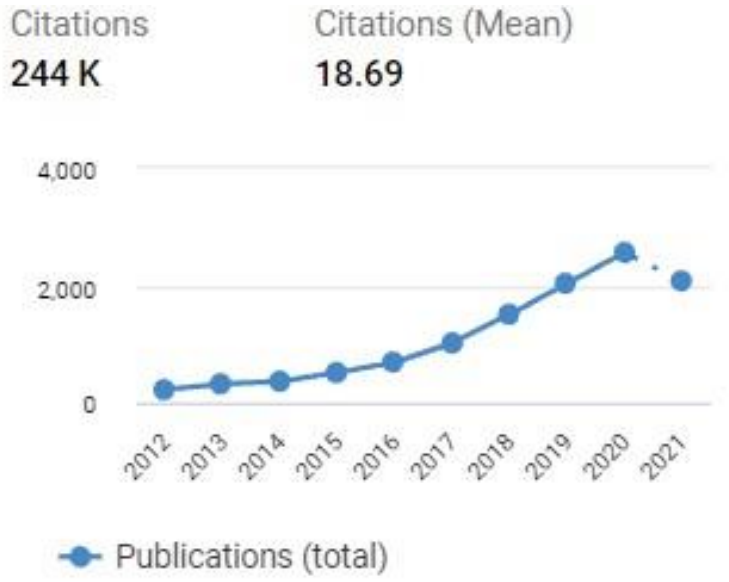


Figure 88. Machine Learning Popularity in Neuroscience from 2012-2021. The plot was obtained using Dimensions AI and shows the publication count of academic papers that included the keyword "machine learning" in their title or abstract.

Threats to External Validity: Overfitting & Overhyping

However, and importantly, with an increase of popularity of a certain methodological approach, there comes an increase in responsibility of the scientific community to soundly adopt the methods on their datasets and, further, to ensure that the generated results are valid and reliable. In the current chapter, we will present a methodological concern regarding machine learning that was thus far mostly overlooked by the Neuroscience community. Concretely, we will introduce the concept of *overhyping* (of classification results to the dataset), a concept that is conceptually very similar to that of *overfitting* (of classification results to the *training* dataset), the latter being arguably the most widely known methodological concern in machine learning in general. Both concerns pertain to the issue of external validity (or, generalisability), i.e., the extent to which classification results obtained with a given dataset are generalisable to the general population from which that dataset was collected. The classic case of overfitting occurs if one trains and tests a classification algorithm on the same dataset. During training, classifiers utilise the dataset's noise (i.e., activation patterns that differentiate classes but are solely due to chance) in addition to its signal (systematic class-differences that are due to the experimental manipulation, for example) to a large extent. Therefore, any decision boundaries that are formed by the algorithm to distinguish classes are also heavily influenced by the dataset's noise. When testing a classifier on the same dataset that was used for training (i.e., the training data), the algorithm performs very well at differentiating classes because the noise with which it was trained is included again in the dataset that is used for testing (i.e., the testing dataset).

Importantly, this also means that the classification accuracy provided will be much higher than if it would have been tested on a *different dataset sampled from the same population* that is unknown to the classifier. Thus, classification models are overfit to the training data if they do not yield results that are generalisable to the population from which a given dataset was sampled. It is hence widespread knowledge that training and testing a classifier should never be done using the same dataset. There are many appropriate measures to prevent overfitting, with the most popular one being cross-validation (CV). The most popular variant of CV is implemented via the repetitive separation of a dataset into a training and testing subset. For example, in 5-fold CV, 80% of the dataset is used for training and the remaining 20% is used for testing. The classification procedure is repeated a total of five times, ensuring that each fifth of the dataset is used once as the testing set. Subsequently, the five classification accuracies that result from the five folds are averaged and provided as the final accuracy value.

Cross-Validation and why it does not prevent Overhyping

Even though CV prevents overfitting to a large extent, we will present cases in which classification results can be inflated despite its usage. Overhyping (i.e., overfitting of hyperparameters) can similarly inflate classification accuracies and threaten the external validity of classification results. While overfitting pertains to *parameters* of the classification algorithms, such as feature weights, overhyping pertains to *hyperparameters* of the analyses. Hyperparameters can be aspects of the analysis pipeline that the experimenter decides on (also known as experimenter degrees of freedom), such as which classification algorithm to use, the extent of classifiers' regularization (i.e., the penalisation of algorithms' complexity), feature selection (e.g., which EEG electrodes to run the analysis on) and many more. Such decisions are called *hyper*-parameters because they are made *on top of an algorithm's internal parameters*. In this chapter we will demonstrate, implementing a rather small search space of hyperparameter configurations, that optimising hyperparameter decisions can lead to inflated classification performance, thereby threatening external validity, *despite the implementation of CV*. CV is regularly believed to prevent inflated estimates of performance because the training and testing datasets are separated in each of the folds. Therefore, the average accuracies on the testing datasets provide unbiased estimates of classification efficacy on out-of-sample datasets (out-of-sample datasets being datasets that originate from the same population as the training data but consist of different samples). Specifically, any parameter-tuning that occurs during training in order to maximise decoding efficacy only transfers

parameters' efficacy to the testing data's *signal*, not its *noise*, as the noise between the training and testing datasets differ, but the underlying signal is shared.

However, and *critically*, even though adopting CV ensures that the noise between training and testing datasets is not shared, CV does not *prevent* overfitting. Overfitting can occur in any machine learning context despite CV. CV, to stress, is a particularly useful tool for obtaining an accurate estimate of the extent of overfitting present. Overfitting, in essence, means fitting a model too strongly on some dataset, and thus being trained too much on the data's noise in addition to its signal, whereas underfitting means fitting a too simple model, thus not capturing all the signal during training. Over- & underfitting are similar to type 1 and type 2 errors in statistics. Regularisation, or complexity penalisation, attempts to balance the likelihood of over- as well as underfitting with classification models, regularising the complexity of a given model with respect to the accuracy it yields. We present an illustration of a model that is underfit, regularised, and overfit in Figure 89A, B, and C, respectively. In these examples a model is fit to a simple two-dimensional dataspace. Figure 89A shows an underfit model because a straight line does not capture the dataset well. Therefore, the model is not trained as well as it could have been, and classification performance will be poor. Figure 89C shows an overfit model, in which the model is fit perfectly to each point of the data. Even though this will provide perfect accuracy *on this dataset*, the model will fail to generalise to a new dataset coming from the same population, thus not providing external validity or generalisability. Finally, Figure 89B illustrates a regularised model, fitting a second order polynomial to the data. This regularised model captures the data better than the underfit model of Figure 89A and not so well that its generalisability is threatened as was the case with the overfit model shown in Figure 89C. The regularised model therefore is the best solution, since it provides a decent fit to the data while generalising to other datasets.

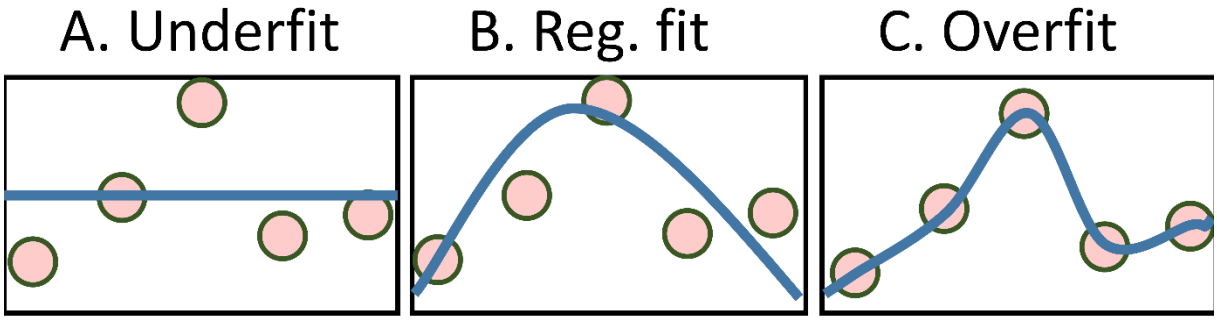


Figure 89. A linear model fit to a simple two-dimensional dataset illustrating underfitting (A), regularised fitting (B), and overfitting (C). The underfit model (A) lacks flexibility and will thus not perform well. The overfit model (C) will perform perfectly on this dataset. However, it will not generalise to other datasets. The regularised model (B) is the best solution, as it will provide decent fit whilst simultaneously generalising to other datasets.

Importantly, overhyping can occur despite CV as well as regularisation because even though these measures ensure that classification accuracy does not reflect the *parameters'* efficacy at decoding the noise contained in the training data, it does not affect the extent to which any *hyperparameter* decisions are based on the dataset's noise. Exploring different hyperparameter configurations based on the classification accuracy they yield, even when adopting CV while doing so, can lead to overhyping, since hyperparameters can be chosen based on which ones decode the dataset's noise best.

For example, a researcher might need to decide on which classification algorithm to use. In this case, the choice of algorithm is a hyperparameter, i.e., a researcher degree of freedom, which has to be decided on top of any *internal parameters* that a given algorithm implements. The researcher could, for example, first run an LDA classifier, then try a logistic regression algorithm and finally implement an SVM. Each algorithm should yield different accuracy values in differentiating the dataset's classes. This example might involve a two-class classification problem, the implementation of 5-fold CV, and the accuracy values could be 52%, 55%, and 59%, for the LDA, logistic regression, and SVM classifiers, respectively. It is tempting for the researcher to conclude that the SVM classifier yielded the highest accuracy score because the algorithm simply decoded the underlying signal best. This thought will be especially tempting if the researcher is conscious about the dangers of overfitting and believes to be protected against these dangers, since they implemented CV. However, in reality, the SVM classifier might not *only* have yielded maximum accuracy across the three candidate algorithms because it decoded the signal best, but, to some extent, the noise, too. The extent to which different hyperparameters lead to different accuracy values because they utilised the dataset's noise, too, can be seen as the extent to which the *hyperparameters are overhyped to the dataset at hand.* If the researcher would collect another dataset from the same population and apply the same three classifiers to decode the same class-contrast, the accuracy values should vary. Hence, the accuracy results obtained by the researcher are not generalisable well to the general population. External validity is threatened.

To demonstrate this issue of overhyping, we will present simulation analyses that resemble the above example of choosing hyperparameters based on maximum classification accuracy (while using CV) for the case of 1000 separate researchers using the temporal generalisation method in this chapter. These analyses will therefore constitute our main piece of evidence in favour of Hypothesis 5 of this thesis, i.e., that machine learning algorithms applied to neuroimaging datasets, particularly temporal generalisation analysis used to

determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfitting to hyperparameters) classification results if appropriate preventive measures are not implemented. We will further explore whether the issue of overhyping is more pronounced for certain classification algorithms. Finally, we will present analyses that investigate the general impact of different hyperparameter configurations on characteristics of temporal generalisation maps, specifically focusing on classification performance and maps' smoothness. It should be mentioned that the main analyses presented in this chapter were published in Neuroscience and Biobehavioural Reviews (Hosseini et al., 2020). Hence, parts of the writing of the current chapter were based on the sections I wrote for that journal paper and for which I had received revisions from my collaborators Howard Bowman and Brad Wyble.

Methods

Data

The simulations presented below were performed on EEG data, which was collected from rapid serial visual presentation (RSVP; Experiment 3 of Callahan-Flintoft et al. (2018)). Subjects viewed a series of changing letters, presented bilaterally, updating at intervals of 150 milliseconds (ms), and were tasked with reporting the one or two digits that would appear on each trial. For this analysis, we selected the trials containing either a single digit, or two digits presented in sequence separated by 600ms and attempted to classify for each trial, whether one or two digits had been presented. However, the trial labels were randomly shuffled within subjects to obscure any actual effect of this manipulation. During each trial, EEG was recorded at 32 electrode sites and according to the standard 10-20 system. It was further bandpass filtered from 0.05 – 100 Hz, originally sampled at 500 Hz and down-sampled offline to 125 Hz for the present analysis. For further details on pre-processing and artifact rejection, see the EEG recordings section of experiment one in the original paper (Callahan-Flintoft et al., 2018). The original study excluded one subject due to an insufficient number of trials after artifact rejection. We decided to exclude an additional subject, choosing the one with the least number of trials, in order to be able to split the data into two equal parts for hyperparameter optimization and lock box, detailed below. The final number of subjects was 24. Finally, the original study divided the data based on the visual hemifield in which the target stimulus was presented and only included trials in which correct responses were provided. We collapsed the data across hemifields and included all trials regardless of accuracy to increase the number of available trials per subject. Using all trials in this way is an experimenter degree of freedom

(i.e., a hyperparameter) that was adopted without looking at the analysis results and thus could not have contributed to over-hyping. Simulations were programmed and analysed in MATLAB 2017b, using functions of the MVPA-Light toolbox (Treder, 2020) for running temporal generalisation analyses.

(Hyper-)Parameter Optimization versus Lock Box

This chapter's most central analysis measures the property of temporal generalisation within an EEG signal, which indicates whether a classifier trained at one point in time relative to stimulus onset is able to classify trial categories at other time points. Such analyses have been used to examine whether memory representations are stable over time in working memory research (e.g., King & Dehaene (2014)).

We ran a series of 1000 independent executions (which we will refer to as iterations below) to measure whether and how often a spurious effect could be obtained if one tested a set of 40 different classifiers on independent random shuffles of a data set. In effect, this is similar to 1000 scientists trying to perform over-hyping on 1000 randomly shuffled copies of the same data set. Each of the 1000 scientists uses cross-validation on 40 different kinds of classifiers and then chooses their best result from the 40.

It needs to be stressed: all analyses were exclusively performed on null data. Hence, any systematic improvements above chance performance must be due to over-hyping. Also, the dataset was randomly split into two equal parts of 12 subjects. One set was for (hyper)parameter optimization (PO) and the other was the lock box (LB) set. The lock box dataset is a portion of the full dataset that is set aside prior to any hyperparameter tuning. In the discussion of this chapter, we will explain why the lock box approach, which entails tuning hyperparameters on one portion of the dataset (the PO set) and accessing classification performance *only* on the dataset's other portion (the LB set), is an effective safeguard against overhyping. Without going into too much detail at this point, we would like to stress that the latter is only the case given that the lock box data is only used *once* to access classification performance. In our simulations, the data were reshuffled into new PO and LB sets at the beginning of every iteration to ensure that any effects were not subject-specific.

For each of the 1000 iterations randomized, PO & LB data sets were created. 40 configurations of classifiers (i.e., 40 different hyperparameter configurations) were used in each iteration to generate temporal generalisation maps to determine which configuration had produced the most desired outcome classifying the random PO data set. The 40 configurations were derived from 4 different classifiers: support vector machines (SVM) with three different kernels (linear, polynomial (order of 2); radial basis function (RBF)) and a linear discriminant

analysis (LDA). Additionally, the extent of regularization was varied through 10 choices for each classifier. For SVMs, the C parameter took values of 0.0001, 0.0007, 0.0059, 0.0464, 0.03593, 2.7825, 21.5443, 166.81, 1291.5496 and 10000. The choice of C values was inspired by (and equal to) the search space of MVPA-Light's default regularization search for SVMs (Treder, 2020). For LDA, candidate lambdas were 1, 0.88, 0.77, 0.66, 0.55, 0.44, 0.33, 0.22, 0.11 and 0. As explained in this chapter's introduction (Figure 89), severe regularisation of algorithms carries the risk of underfitting classification models (compare Figure 89A), whereas too little regularisation allows for more complex models, risking overfitting (Figure 89C). Our candidate lambda values in the case of LDA algorithms were distributed linearly in the space from 0 to 1 and implied less regularisation and more complex models as lambda values *decreased*. In contrast, SVM's C values of the present analysis were distributed non-linearly and meant less regularisation and more complex models as C *increased*. Temporal generalisation analyses were performed using 5-fold cross-validation.

These 40 candidate configurations competed in each of the 1000 iterations of the analysis, which we call *PO Competition* as it represents a competition between hyperparameters to decide the best hyperparameter configuration available. This PO competition was decided using a measure we call classification mass (*C-Mass*), which was computed on group-average temporal generalization maps (i.e., averages of 5-fold cross-validated single-subject maps). C-Mass reflects the average AUC value across the entire temporal generalization map. For each of the 1000 iterations, the hyperparameter configuration that led to maximum C-Mass (i.e., highest map-average AUC value) when classifying the PO data set was selected as the respective winner of the PO competition for that iteration. These winning configurations were then used to assess the degree of over-hyping by comparing them to the LB set.

We implemented this C-Mass criterion after feedback from a reviewer when submitting the corresponding journal paper (Hosseini et al., 2020). Our initial simulations used squared C-Mass, to measure the extent of above- as well as below-chance classification across a temporal generalization map. This was computed by first subtracting chance-level classification (AUC of 0.5) from all AUC values of a given map, then squaring these values and finally taking the average of the entire map. We squared AUC values after subtracting them from chance-level because of a unique and interesting implication of below chance-level (or mis-) classification in temporal generalisation plots. Misclassification can imply that the whole-scalp pattern difference between classes measured at testing time y resemble those that were present at training time x, *but were of the opposite polarity*, which leads to classifiers

making more mistakes than they would by chance. Therefore, one can argue that below-chance classification performance in the context of EEG-data is not necessarily uninteresting noise. To illustrate, we present all results of our PO versus LB analysis, which illustrates overhyping, again using our initial, squared, C-Mass variant in Appendix A. It is worth pointing out that both variants of C-Mass reveal essentially similar results, but we use the unsquared (i.e., map-average AUC value) in the chapter's main body because above-chance classification is the more canonical approach.

TG maps were tested for statistical significance according to the cluster-extent sign-swap permutation test detailed earlier, adopting a one-tailed significance test (only considering positive AUC clusters). We also restricted the tested area to be from approximately 100 milliseconds onwards to resemble real-life usage of this test. However, this does not change the statistics in any way, because tests of maximum statistics are automatically correcting for family-wise errors (FWE) as long as one is adopting the same approach (e.g., restricting the area) for the observed and the (permuted) null data (in this case TG maps).

General Impact of Hyperparameters' Complexity on C-Mass and Maps' Smoothness

We further investigated whether different configurations of hyperparameters *generally* influence characteristics of temporal generalisation maps, specifically focussing on C-Mass and TG maps' smoothness. For these analyses, the same hyperparameter configurations as well as temporal generalisation options were adopted as presented above. However, the dataset was *not divided into two subsets*. The analysis commenced with shuffling trials for all 24 subjects, hence simulating null data. Group-level temporal generalisation maps were computed for all 40 hyperparameter configurations as described above. Subsequently, the data was re-shuffled, and the procedure was repeated 1000 times, which resulted in a total of 40000 group-level temporal generalisation maps (40 for each of the 1000 iterations). Maps' smoothness was assessed by analysing the mean oscillatory frequency of *main diagonal* as well as *horizontal* lines. Figure 90 shows example temporal generalisation maps resulting from simple (panel A, C parameter equal to 0.0001) and complex (panel B, C = 10000) linear SVM classifiers classifying the same null dataset. The map in panel A includes the main diagonal in black, which assesses how well classifiers decoded unknown data measured at the time-point they were trained at (training = testing time). An example horizontal line is included in the map of panel B, which assesses how well a given classifier was able to decode class-differences across all time-points, i.e., the extent of *temporal*

generalisation (fixed training & all testing times). Both maps' main diagonal and horizontal lines' AUC vectors were extracted and plotted as time-series in black and pink, respectively, under each map. The shaded areas in the time-series plots reflect the mass of above and below chance-level (AUC of 0.5) classification. The AUC time series show that the simpler classification yielded a higher C-Mass that is smoother through time. To support the former claim, we will provide exploratory analyses later on. To confirm the latter claim, we performed a Fourier decomposition of these lines, i.e., AUC vectors, which were subsequently analysed as follows.

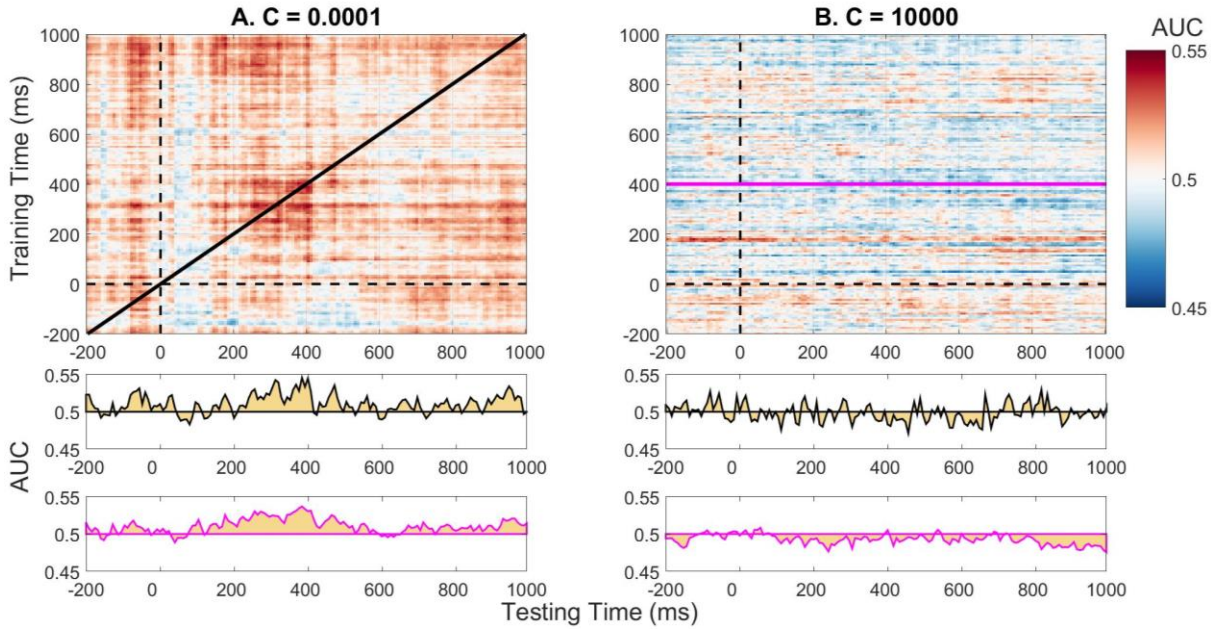


Figure 90. Example temporal generalisation maps (top) selected to show differences in C-Mass and smoothness. Maps depict how simple (panel A) and complex (panel B) linear SVM classifiers solved the same classification problem between two classes of the same null-dataset. Hence, any differences in C-Mass as well as smoothness are only due to selection of hyperparameters. The black diagonal line in panel A shows the 'main diagonal', i.e. vector of AUCs when training = testing time. The pink line in panel B shows a horizontal line, i.e. 'temporal generalisation', or, vector of AUCs with fixed training time (here 400ms) across all testing times. AUCs of maps' main diagonals (middle, black) and horizontal lines (bottom, pink) were extracted and plotted as time-series next to each other for comparison. Note that even though the testing times (x-axes) are identical between main diagonal and horizontal lines, the training times differ (main diagonal (middle): training *is equal to* testing time, but for horizontal line (bottom): training time = 400ms across all testing times). Shaded areas in time-series plots reflect mass of classification above and below chance-level (AUC = 0.5).

The Fourier Transforms of these lines initially included edge-effects. Accordingly, we high-pass filtered lines at 2 Hz prior to computing their Fourier Transform. Figure 91 shows example power spectra that resulted from performing this analysis for the four AUC vectors shown in Figure 90. Each plot shows the original power spectrum in transparent and its smoothed version (convolving with a box car of 10 Hz width) in solid. For comparison, we included a third line (different colour & transparent) in each plot, which shows the smoothed power spectrum of the other model's line of interest (i.e., main diagonals in panels A and B & horizontal lines in panels C and D). Simple and complex classification models are colour coded in green and red, respectively. For example, panels A and B show power spectra of

main diagonal lines after classifying with simple (A) and complex (B) linear SVMs. In panel A, the raw power spectrum of the simple SVM's main diagonal is shown in green and transparent, the smoothed spectrum is green and solid and the smoothed power spectrum of the *complex* SVM's main diagonal is displayed in red and transparent (same line that is red and solid in panel B). This layout is identical for horizontal lines' power spectra (panels C and D). The four solid lines in Figure 91 therefore correspond to the power spectra of their corresponding AUC vectors of Figure 90 (e.g. Figure 91's panel C shows the power spectrum of Figure 90's bottom left AUC vector, i.e. the pink horizontal line of the simple SVM map). Lines' mean frequency was calculated as the expected value across the power spectra and is included in Figure 91 in each plot's top right corner. As anticipated, mean frequencies are lower for the simpler analysis.

We finally fitted respective linear OLS regression models to mean frequencies, with classifier (dummy-coded categorical with linear SVM as reference), regularization value (continuous) and the interactions between them as predictor variables. For computational efficiency, we only considered the first 100 iterations' maps for analyses of smoothness.

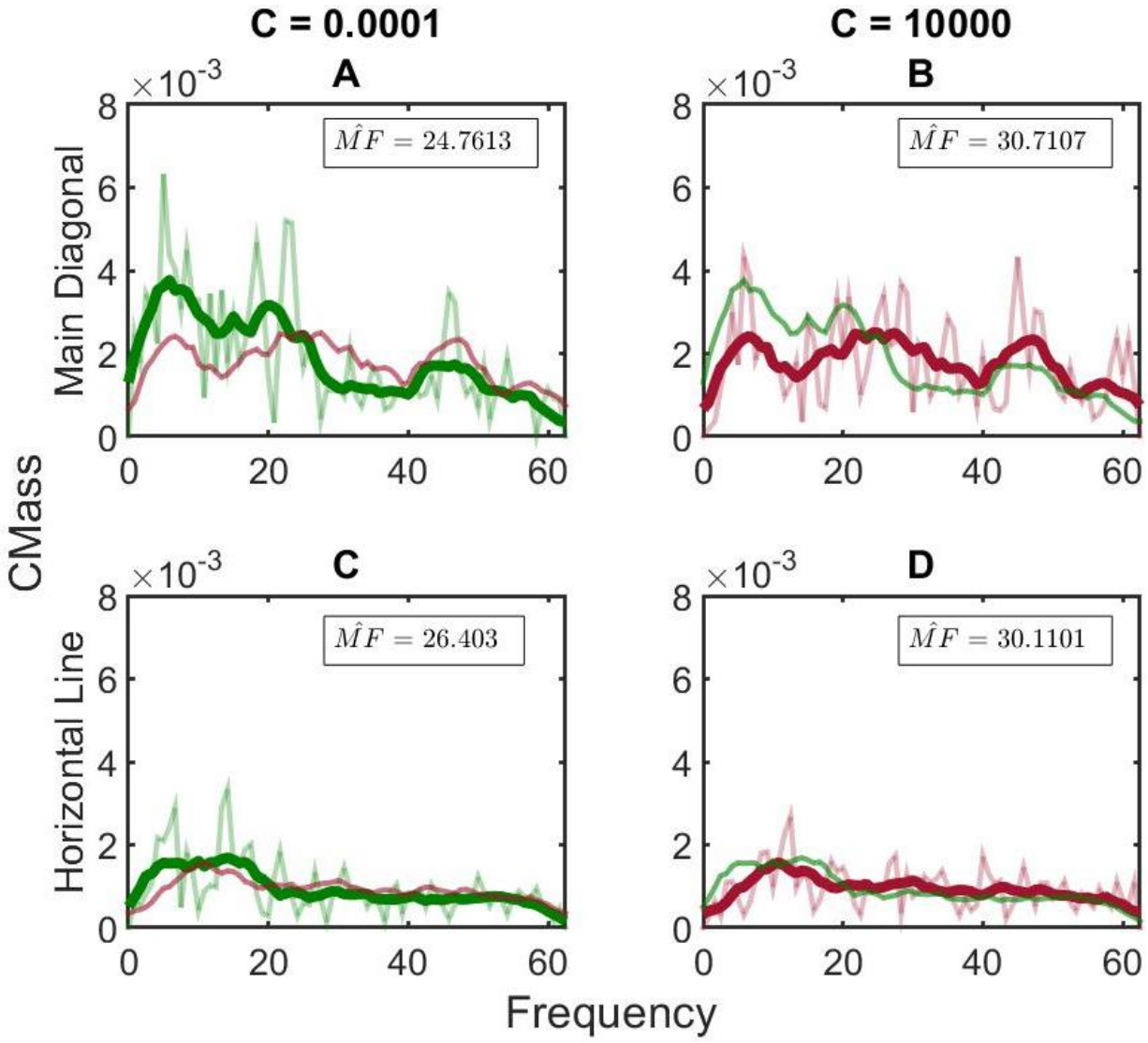


Figure 91. Power spectra after Fourier Transformations of main diagonal (top panels A & B) and horizontal (bottom panels C & D) AUC lines. AUCs were extracted from the temporal generalisation maps shown in Figure 2 (which resulted after simple (left panels A & C) and complex (right panels B & D) SVM models classifying the same data). Each panel includes the original power spectrum in transparent, the smoothed (box-car with width of 10Hz) power spectrum in solid and the other model's smoothed power spectrum in a different colour & transparent for comparison. For example, panel D shows the power spectrum of a horizontal line after complex SVM classification in transparent and red, the smoothed spectrum in solid and red and the smoothed spectrum of the simple SVM horizontal line in transparent & green. Textboxes include lines' expected (i.e. mean) frequency in top right corner. The layout of this figure follows the four AUC vectors shown under temporal generalisation maps in Figure 2.

Results

Systematic Overhyping demonstrated by PO versus LB comparisons

We selected one of the 1000 iterations to demonstrate how manual selection could produce *what appears to be* a theoretically meaningful result in a temporal generalization map for data that was randomly shuffled and subjected to cross validation (Figure 92). Evaluating the efficacy of different hyperparameter configurations on the same dataset can be considered an analogue of an exploratory analysis in which an analyst runs a series of cross-validated pilot analyses and stops on finding one that is theoretically suitable. In the working memory

literature, it is considered important that a classifier is able to decode the condition label after the stimulus has disappeared. In our manually selected case, the winning Parameter Optimisation (PO) map can be regarded theoretically suitable as it appears to exhibit this property whereby the accuracy remains well above chance for a substantial period of time after the target onset (see Appendix B for a randomly selected set of 9 additional iterations). However, this effect is demonstrably spurious since the trial labels were all randomly shuffled. To demonstrate that this observed pattern is due to overhyping and not a general property of our analysis, we also present the results of the same analysis configuration for the companion Lock-box (LB) set as well as of an alternative classifier configuration for the same PO set. It is clear that the pattern observed in the winning PO map does not generalize either to another data set drawn from the same population using the same kernel configuration (the LB set) or to a reanalysis of the same data set with a different configuration (the losing PO set). This is an instance of overhyping (overfitting due to selection of hyperparameters), because the hyperparameters determined with the PO set fitted the noise best compared to the other candidate hyperparameters. As the noise differed in the LB data set, classification performance was overall at a lower level and the pattern of more successful classification at later time points was also disrupted, causing our permutation test to generate significant AUC clusters for the winning PO, but not the LB map. This finding therefore constitutes initial evidence in favour of Hypothesis 5 of this thesis, i.e., that machine learning algorithms applied to neuroimaging datasets, particularly temporal generalisation analysis used to determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfitting to hyperparameters) classification results if appropriate preventive measures are not implemented.

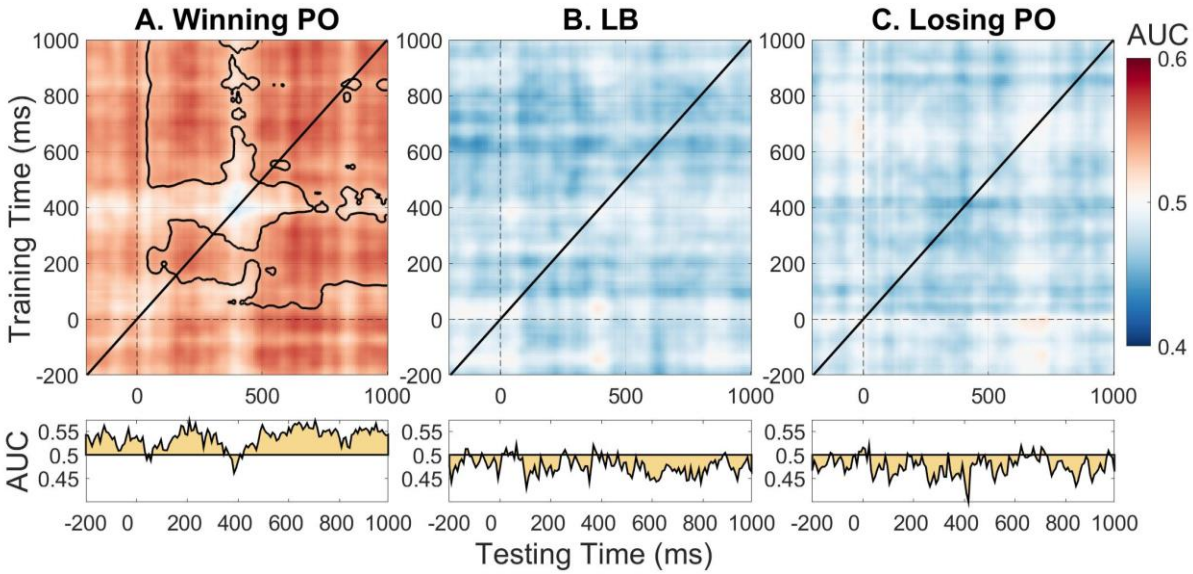


Figure 92. How overhyping manifests in temporal generalisation maps. Maps of a winning optimization set (Winning PO), its corresponding Lock Box (LB) and the worst optimization set (Losing PO), which implemented hyperparameters that led to minimal C-Mass, are plotted with their main diagonal AUC vectors below. Beige areas in AUC time-series plots show divergence from chance-level classification (i.e., AUC of 0.5) in main diagonals. Classification performance was at a higher level for the winning PO compared to both other analyses. A family-wise error correction cluster-extent test was performed (Nichols & Holmes, 2002) for winning PO & LB maps and only showed statistically significant AUC clusters for the PO map. Maps and cluster-boundaries (i.e., matrices determining statistical significance) were 2D-smoothed separately using a boxcar of 40 ms width. This was only done to facilitate visualization and did not affect any analyses, which were all computed prior to smoothing. As all three analyses decoded null-data, any differences in classification performance must be due to the effectiveness of classifiers' hyperparameters (in this case an LDA classifier with a lambda of 1 for winning PO & LB). This is a demonstration of overhyping because these hyperparameters fitted the noise of the PO dataset best, which however differed in the LB dataset and thus led to decreased classification performance for the LB. This map triplet was manually chosen. An additional 9 triplets can be found in the Appendix B.

Figure 93 displays how often candidate hyperparameters were selected based on winning (top row) or losing (bottom row) the PO competition as well as being selected at random (middle row, note that the random analyses were only performed on LB sets). Again, the PO competitions were held between hyperparameters, and winners decoded class-differences found within null-datasets most effectively, whereas losers decoded these differences least effectively. Therefore, any deviations from a uniform distribution must be due to differences in the effectiveness of hyperparameters. As one would expect, the distribution of classification models (left column) is close to being uniform for random analyses, where hyperparameters were selected at random. In contrast, the top and bottom rows of the left column show that LDA classifiers were selected most frequently in both cases, meaning that LDAs both *won and lost* the PO competition the most. Within SVMs, linear models also *won and lost* the most, followed by polynomial models and, finally, RBF ones. Overall, this shows that the more regularized and simple classification models are, the more frequently they *win and lose* the PO competition. This pattern might seem counterintuitive at first, but the reasons underlying it will become apparent when we present how C-Mass behaves for different hyperparameter configurations.

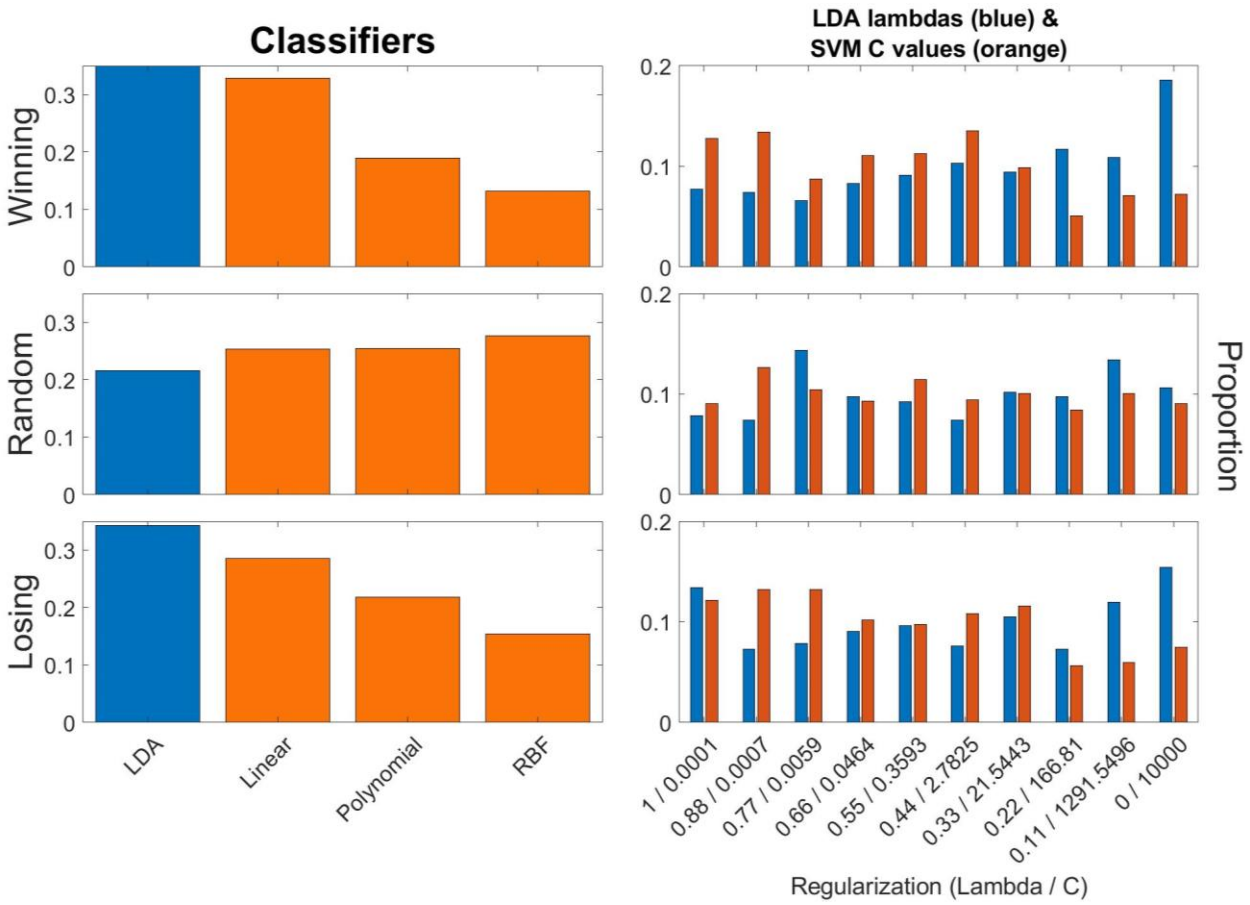


Figure 93. Proportions of hyperparameters selected after winning (top row) or losing (bottom row) the PO competition. The middle row corresponds to randomly selected hyperparameters, which were applied to the LB independent of the PO competition. The left column depicts distributions of classification models (blue = LDA, orange = different SVMs) and the right column shows how regularization parameters (lambda (blue) for LDA and C (orange) for SVM classifiers) were distributed. Note that regularization decreases from left to right (classification models become increasingly complex).

The main results of the present analysis are displayed in Figure 94, which depicts how C-Mass values were distributed for the PO and LB analyses. Distributions of C-Mass for LB analyses are coloured in red, PO distributions in blue (left column) and distributions of the within-iteration differences between PO and LB C-Mass (calculated as C-Mass (PO) – C-Mass (LB) for each iteration, or re-shuffling of class-labels) are green. The coloured vertical lines indicate a distribution's median value and the coloured rectangles around the vertical lines indicate interquartile ranges. The black vertical line in the right column of Figure 94 reflects a difference in C-Mass between PO and LB of zero, i.e., no difference in the effectiveness of hyperparameters based on which datasets (PO or LB) they were applied to. To stress, all C-Mass distributions of Figure 94 resulted from temporal generalisation analyses on *null-data*. Therefore, importantly, any differences between C-Mass distributions of PO and LB analyses must necessarily be due to differences in the effectiveness of hyperparameter configurations to decode class-differences contained in the EEG signal.

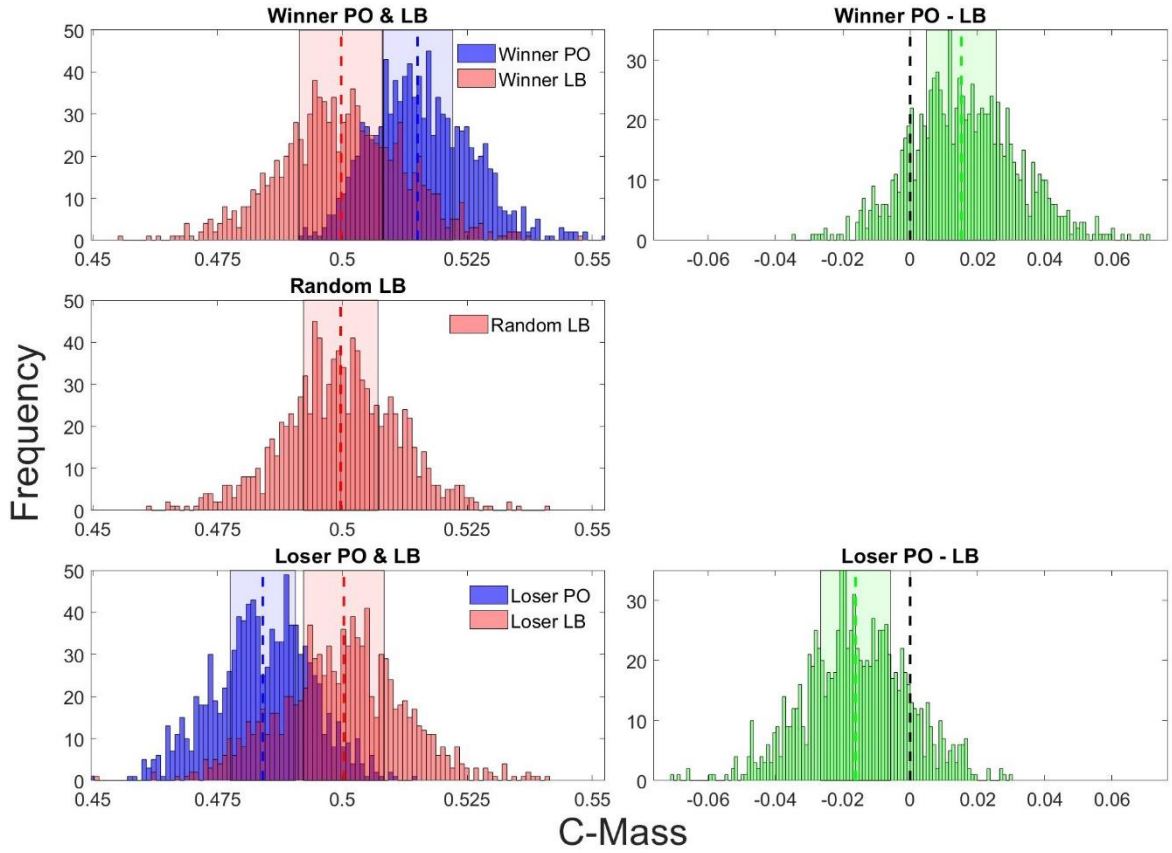


Figure 94. C-Mass distributions of PO (blue) and LB (red) maps (left column), as well as their within-iteration difference (green, right column). The top row shows C-Mass results for the PO & LB maps that incorporated *winning* hyperparameters, the bottom row shows distributions incorporating *losing* hyperparameters and the middle row shows the distribution of LB C-Mass after choosing hyperparameters *randomly*. The coloured vertical lines indicate distributions' median value and the rectangles surrounding these lines indicate the interquartile ranges. The black vertical lines in the right column's plots indicates a PO – LB difference of zero (i.e., no over- or underhyping).

The main focus of this analysis is the top left panel of Figure 94: the distributions of PO & LB C-Mass for winning hyperparameters of the PO competition clearly demonstrate overhyping of classification results. If overhyping was absent, these distributions should sit on top of one another. However, the PO C-Mass distribution has a higher mean (0.516), median (0.515) and smaller variance (0.0001) compared to the LB C-Mass distribution (mean: 0.5, median: 0.5, variance: 0.0002). The top right panel of Figure 94 illustrates how the *within-iteration* differences in C-Mass between PO and LB were distributed. This distribution should be centred around zero if no overhyping was observed (i.e., a given set of hyperparameters leading to similar success in decoding between-class differences for PO as well as LB null-data). The observed mean (0.016) and median (0.015) of the difference distribution was positive, implying higher C-Mass in PO maps. Permutation tests, which were based on randomly determining the direction of subtraction between PO & LB C-Mass to generate a distribution of PO-LB C-Mass differences under the null, confirmed that both values were significantly different from zero ($p < .001$), which provides evidence for overhyping and thereby for Hypothesis 5 of this thesis, i.e., that machine learning algorithms applied to

neuroimaging datasets, particularly temporal generalisation analysis used to determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfitting to hyperparameters) classification results if appropriate preventive measures are not implemented.

We acknowledge that p-values can become uninformative in the context of simulations, where very large, simulated samples can be run. As discussed in Friston (2012), the fallacy of classical inference states that once the sample size is sufficiently large, p-values become trivial as the smallest effects suffice for significance. To be more precise, for a two-sided test, there is a sample size N for any non-zero experimental effect, no matter how small, that will make the p-value significant. We nonetheless chose a large sample size because, in contrast to p-values, standardised measures of effect size become more accurate with increasing sample size. This was recently illustrated in a neuropsychology context (compare Lorca-Puls et al. (2018), especially figures 4 & 5).

Distributions of C-Mass in losing PO & LB as well as random LB maps were mainly created to enable comparisons to the case of winning PO & LB, which shows overhyping. Two important observations can be made. First, the LB C-Mass distributions came out being similar, independent of whether they incorporated winning, losing, or random hyperparameters. This means that using a Lock Box to assess classification performance after choosing hyperparameters with the PO set is a valid approach and the resulting C-Mass is free of any overhyping effects. The second observation lies with the unrealistic case we show in the bottom row of Figure 94, losing PO & LB C-Mass distributions. This case shows what happens if one would choose their classifier and its hyperparameters based on which option performed *worst.* Even though this criterion might never be used to choose hyperparameters, it is worth noting that the C-Mass of losing PO maps was *lower* than that of LB maps, meaning that the PO *underperformed* the LB (Figure 94, bottom left panel) in this case. It should be stressed that these patterns resulted from a search space of just 40 hyperparameter options out of which the best (and worst) configurations were defined using the PO set and which were then forwarded to the LB set. It is valid to assume that these results, as well as the problem of overhyping classification analyses, would only be more severe for larger search spaces.

PO versus LB by classifier

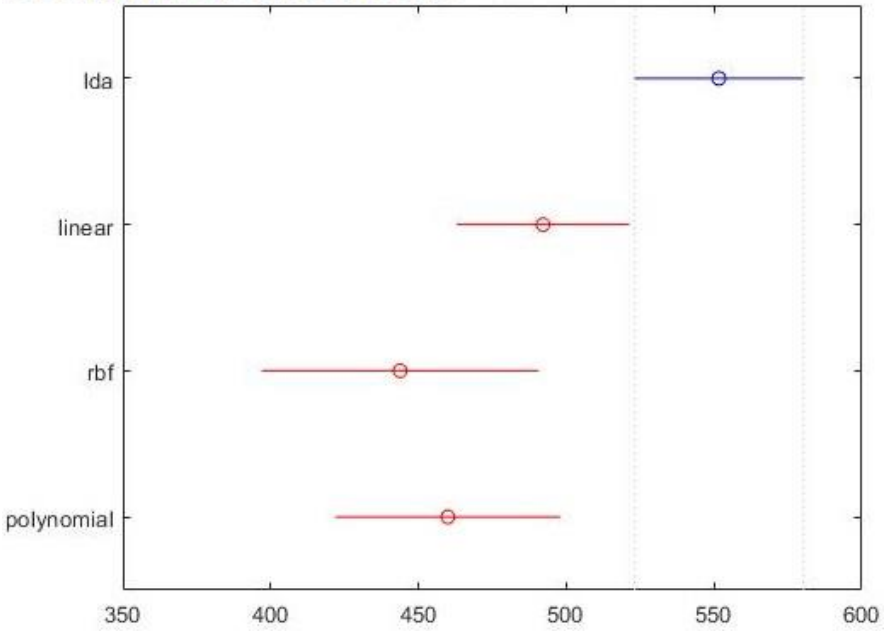
We further assessed how vulnerable the different classifiers were to overhyping. Across all classifiers, the median difference in C-Mass between winning PO and LB was

positive and significantly different from zero after performing the permutation test introduced above (linear SVM: median = 0.015, n = 329; polynomial SVM: median = 0.013, n = 189; RBF SVM: median = 0.013, n = 132; LDA: median = 0.018, n = 350). Classifier-specific equivalents of Figure 94 (PO & LB C-Mass distributions) are provided in Appendix C to further illustrate that overhyping is problematic across all classifiers tested. We furthermore investigated whether overhyping was more pronounced for certain classifiers by conducting a Kruskal-Wallis test (due to non-normality of C-Mass values). The Kruskal-Wallis test is a non-parametric alternative to a one-way ANOVA, which we used for assessing the degree of overhyping across different classifiers because the C-Mass distributions did not meet parametric assumptions. It was computed on classifier-separate C-Mass difference distributions between PO and LB (compare right panels of Figure 94). Thus, it tested the null hypothesis that the PO – LB C-Mass differences of our four groups (i.e., classifiers) originated from the same distribution. This null hypothesis was rejected (χ^2 (3) = 20.07, $p < .001$), as can be seen in the ANOVA Table below (Figure 95A). Following up on this result, we further tested pair-wise mean-rank differences between classifiers using Tukey's honest significant difference criterion to correct for multiple comparisons. The results of these tests are provided in Figure 95B & C and showed that: 1) LDA had a significantly higher mean rank than all other classifiers (i.e., most overhyping); 2) across SVMs, all differences were non-significant, but linear SVMs had the highest mean rank, followed by polynomial and RFB SVMs. These results thus provide evidence that overhyping is especially problematic for LDA classification.

A. ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.67382e+06	3	557940.8	20.07	0.0002
Error	8.16594e+07	996	81987.4		
Total	8.33333e+07	999			

B. Rank Distributions



C. Testing mean ranks between classifiers

*** GROUP 1: lda***

*** GROUP 2: linear***

*** GROUP 3: rbf***

*** GROUP 4: polynomial***

*** Group 1,	Group 2,	Lower Bnd,	Estimate,	Upper Bnd,	P***
1.0000	2.0000	2.5351	59.5120	116.4889	0.0366
1.0000	3.0000	32.0778	107.8654	183.6531	0.0015
1.0000	4.0000	24.7058	91.6828	158.6597	0.0025
2.0000	3.0000	-28.0937	48.3534	124.8006	0.3645
2.0000	4.0000	-35.5516	32.1708	99.8931	0.6138
3.0000	4.0000	-100.3475	-16.1827	67.9821	0.9605

Figure 95. Results of Kruskal-Wallis test conducted on PO – LB C-Mass distributions to assess the degree of overhyping between our four classifiers. Panel A shows the ANOVA table, which indicates a significant main effect of classifier. Panel B plots the rank distributions for each classifier, which indicates that mean ranks (i.e. degree of overhyping) of LDA > linear SVMs > polynomial SVMs > RBF SVM. Panel C shows the results of statistically testing all pair-wise mean-rank differences. Three out of six tests yielded significant p-values after correcting for multiple comparisons using Tukey's honest significant difference criterion, indicating that the extent of overhyping was significantly larger for LDA classifiers compared to all SVM ones.

C-Mass and Smoothness

The impact of classifiers' hyperparameters on C-Mass after temporal generalisation of null-data was assessed with boxplots (Figure 96). These indicated that C-Mass varies less and is overall at a lower level for SVMs if C is higher and thus models are more complex. For the case of LDA, changing lambda seemed to have little impact on C-Mass. Figure 96 demonstrates why simple models *won and lost* the PO competition the most, as was shown in Figure 93. This is because the C-Mass of simple classification models varied much more than for complex models. Therefore, simple classification models more frequently resulted in more extreme C-Mass values, hence resulting in them *winning and losing* the PO competition the most.

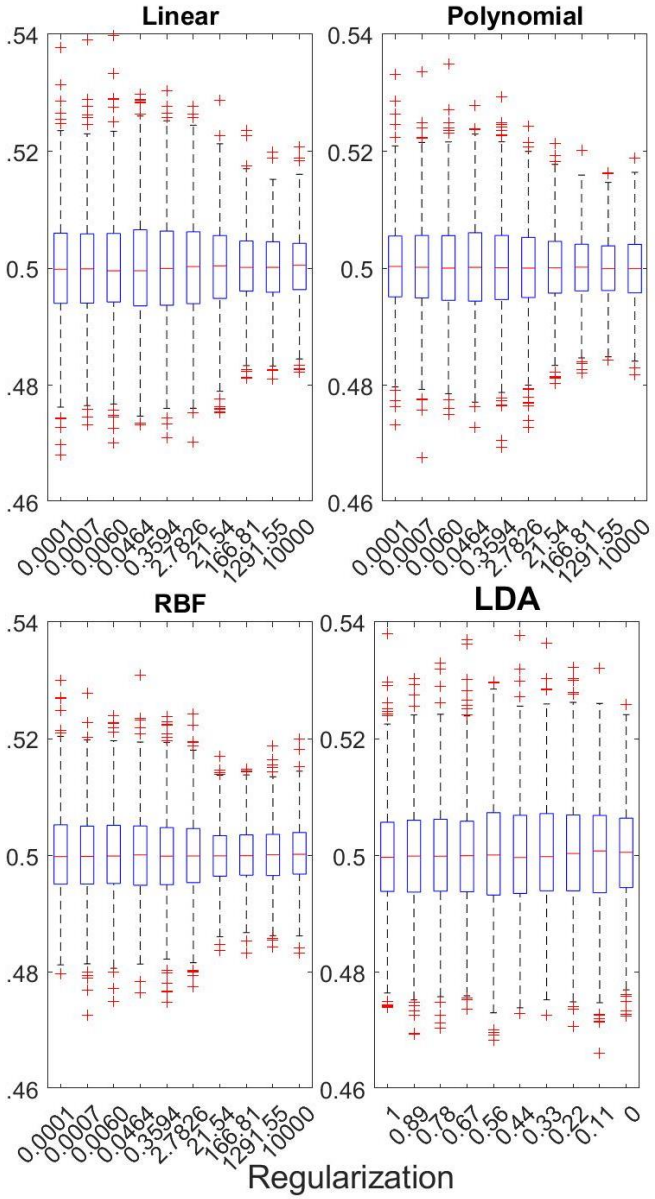


Figure 96. Boxplots of C-Mass separated by classifier. Red horizontal lines indicate the medians.

The relationship between classifiers' hyperparameters and the smoothness of the resulting temporal generalisation maps, looking at 100 iterations, was assessed next. The rationale behind this was that, due to their flexibility, complex classification models should fit the noise in addition to the signal contained in the EEG measured at the time-point they are trained at. We expected this to differ for simple models, which should not fit the noise as well. Hence, the ability to generalise to other time-points' data should be inferior for complex models, which therefore results in decreased C-Mass (Figure 96), because C-Mass assesses classification performance *across the whole map*. Models' complexity should also impact maps' smoothness, as smoother data can be better characterized by simpler models, such as our strongly regularized linear SVM. The boxplots displayed in Figure 97 indicate that, indeed, as regularization decreases, and hence classification models get increasingly complex, the maps' mean frequency increases accordingly (i.e., maps are less smooth). Looking into this relationship further, we assessed the across-iteration second-level mean frequency (i.e., mean of mean frequencies) for each classifier-regularization combination. The surface plots displayed in Figure 98 (horizontal lines) and Figure 99 (main diagonal lines) confirm the observation of increasingly complex models leading to less smooth maps.

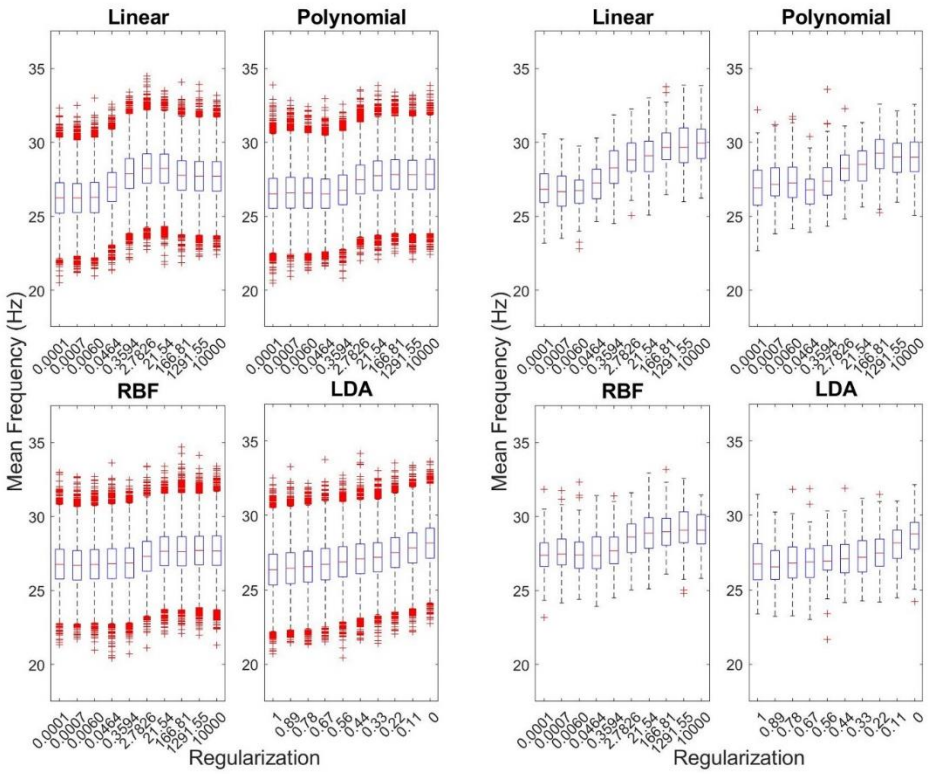


Figure 97. Boxplots of mean frequencies of the horizontal lines (left) and main diagonals (right) analysis separate for each classifier.

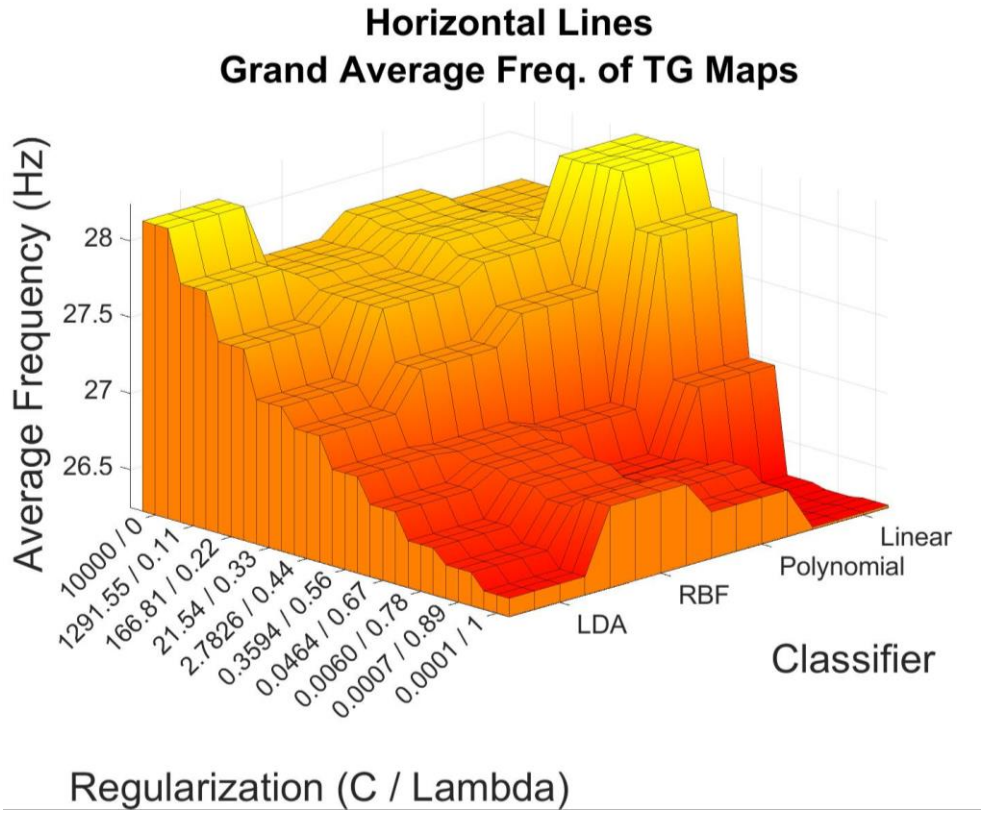


Figure 98. Surface plot of grand-average mean frequency for each classifier-regularization parameter combination of horizontal lines analysis.

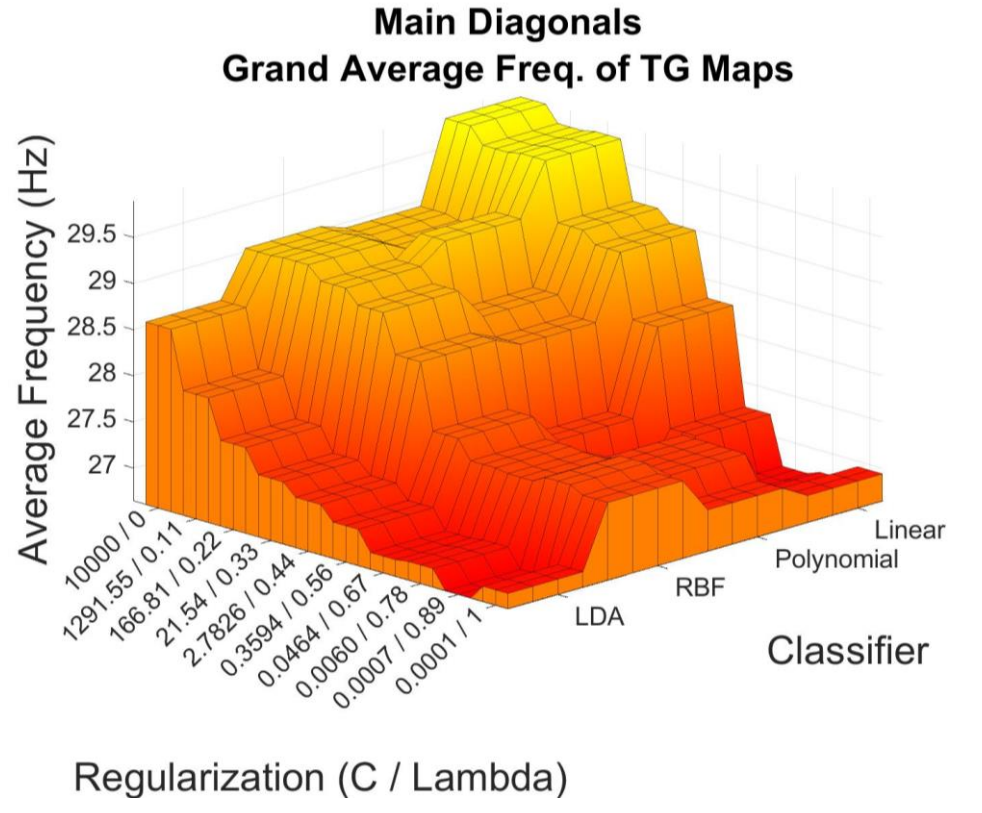


Figure 99. Surface plot of grand-average mean frequency for each classifier-regularization parameter combination of main diagonal analysis.

Finally, OLS regression models (with classifier, regularization & their interactions as predictors) were fitted to mean frequencies of horizontal and main diagonal lines, separately. The models' R^2 values were 12.3% (horizontal lines) and 16.1% (main diagonals). Further inference was conducted on the regression model of main diagonal lines via the unstandardized beta coefficients in connection with main effects' semi-partial correlation coefficients. The beta coefficients of the main effect of different classifiers were all non-significant (sr^2 = 1.9%). The effect of regularization did, however, yield a significant and positive beta of 0.27, implying an increase in mean frequency with decreasing complexity regularization (sr^2 = 13.4%). The interaction terms were all negative, but only the one contrasting linear SVMs with LDA classifiers was significant (b = - 0.115, t(3992) = -4.5, p < 0.0001), suggesting that changes in regularization had a stronger effect on mean frequencies after linear SVM classification compared to LDAs. Interactions are illustrated in Figure 100 & Figure 101. Again, the patterns presented in this paragraph are all results of the analysis of null-data, which lacks any experimental effect. Therefore, the differences in smoothness as well as C-Mass we show here are necessarily due to the selection of hyperparameters. Finally, Figure 90 illustrates how these differences in smoothness and C-Mass may appear in temporal generalisation maps.

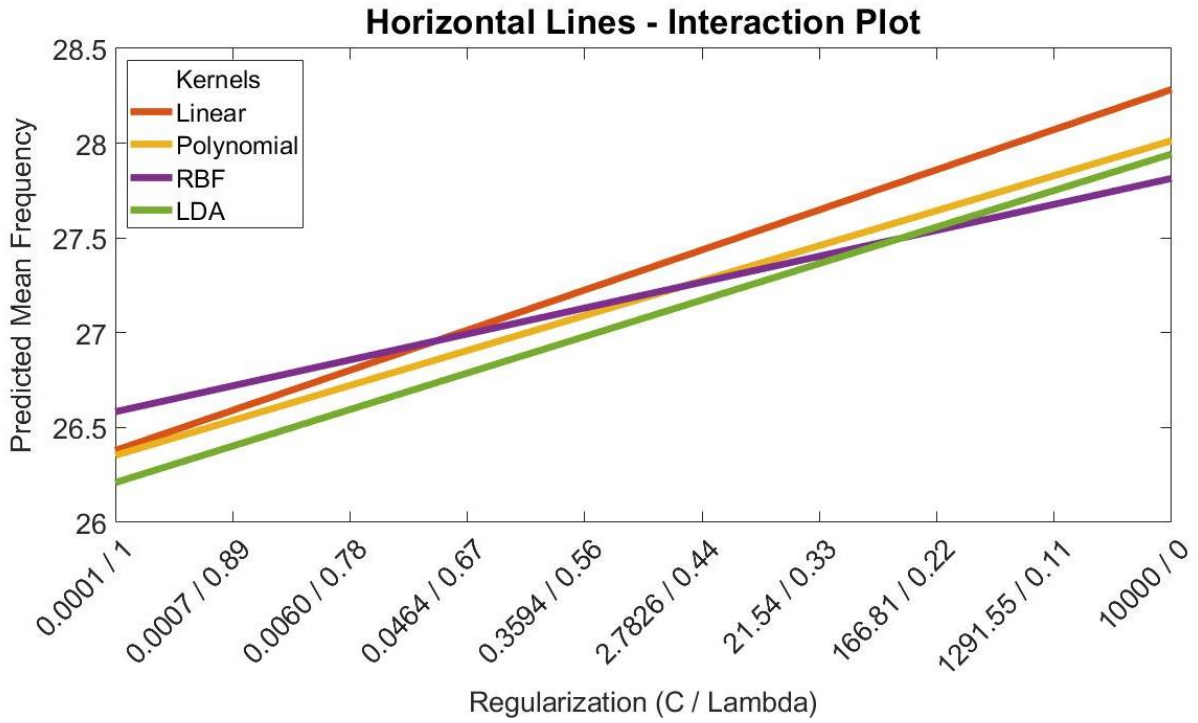


Figure 100. Horizontal lines. Predicted mean frequency for each classifier-regularization parameter combination.

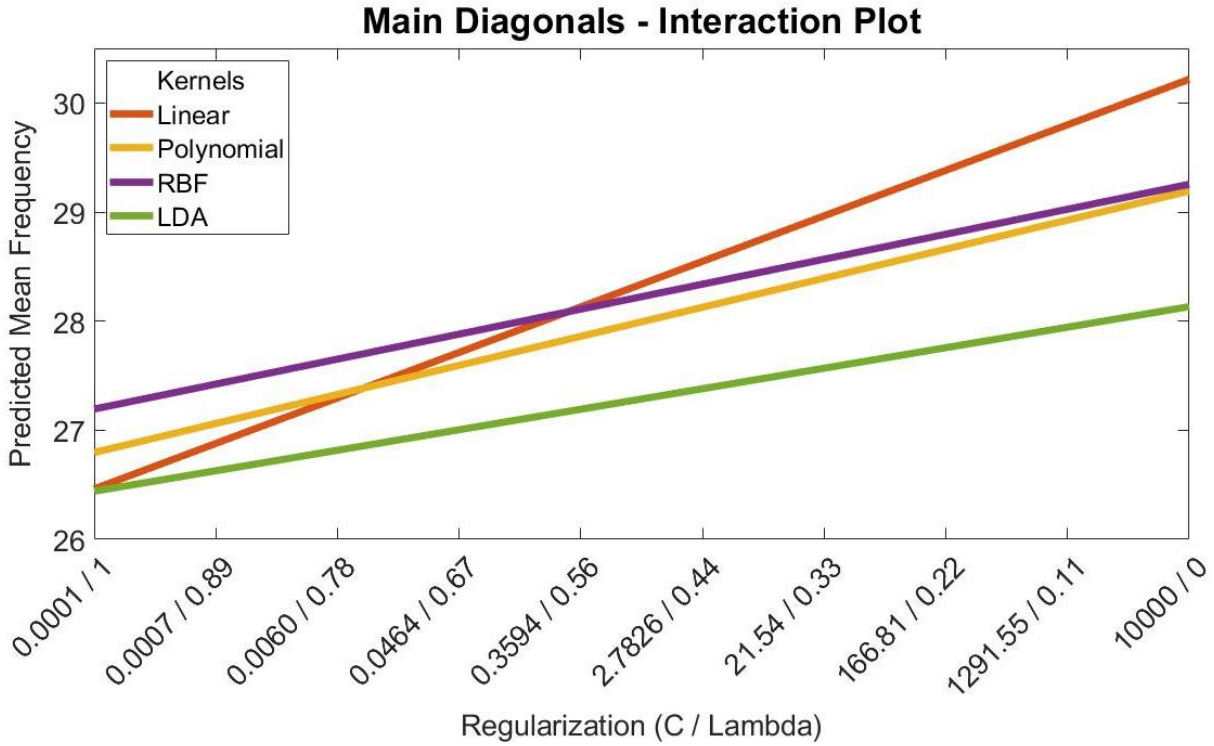


Figure 101. Main diagonal lines. Predicted mean frequency for each classifier-regularization parameter combination.

Discussion

In this chapter we presented a series of simulation analyses that demonstrated *over-hyping* (of classification results), a concept akin to that of overfitting, with the latter being one of the fundamental methodological concerns in machine learning in general. Importantly, we demonstrated, implementing a rather small search space of 40 hyperparameter configurations (4 classification kernels with 10 levels of complexity regularization each), that grounding hyperparameter decisions on maximised classification accuracy can lead to inflated results, threatening external validity, *despite the implementation of cross-validation*. We thereby presented evidence in favour of Hypothesis 5 of this thesis, i.e., that machine learning algorithms applied to neuroimaging datasets, particularly temporal generalisation analysis used to determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfitting to hyperparameters) classification results if appropriate preventive measures are not implemented.

In summary, our main demonstration used null data, the method of temporal generalisation (TG), a measure we call C-Mass to assess classification performance, and simulations that resembled 1000 separate experimenters choosing classifiers' hyperparameters based on which configuration yielded maximum classification accuracy. We demonstrated overhyping (Figure 94) showing a systematic increase in C-Mass when hyperparameter decisions were based on maximising C-Mass compared to when the same hyperparameter configurations were forwarded to a different dataset (the Lock Box). Most critically, classification analyses were always conducted after shuffling the datasets' class-labels to simulate null data, i.e., data in which no differences between classes (in this case response conditions) were due to systematic differences, such as arising from experimental manipulations. Therefore, any deviation from a classification accuracy of 0.5 (chance level) was, in a statistical sense, solely due to chance. Besides these main ((hyper)parameter (PO) versus lock box (LB)) analyses, we demonstrated that overhyping was problematic across all four classification algorithms of interest (however, being most severe for Linear Discriminant Analysis (LDA) classifiers). Moreover, we investigated the distribution of hyperparameters that led to maximal and minimal C-Mass and showed that, compared to complex models, more simple classification models frequently led to minimal *as well as* maximal C-Mass. We additionally provided an explanation for this rather unexpected pattern of results, demonstrating that simple models yield more variable classification accuracies and thus lead to more extreme C-Mass values in general. Additionally, we provided further PO versus LB

analyses in Appendix A, which demonstrated overhyping using a (squared) C-Mass variant in which below- as well as above-chance classification accuracy was considered desirable. We finally presented a series of analyses that suggested that more simple classification models yield smoother TG maps. Importantly, these results do not imply that one should avoid utilising maximised classification performance as the criterion for determining hyperparameters of a machine learning analysis. We instead provide a series of protective measures that allow the latter procedure without running the risk of overhyping.

Safeguards against Overhyping

The first approach that can prevent overhyping is *pre-registration* (Nosek et al., 2018). Pre-registration makes overhyping unlikely because it includes the submission of the complete analysis pipeline prior to conducting any analysis steps. In this case, standard n-fold cross-validation suffices as a protective measure against overfitting as well as overhyping. This is because all analysis steps, including those involving machine learning decisions, such as which algorithm to utilise, are decided on a-priori. Therefore, no decisions about hyperparameters will be based on maximising classification performance and, thus, one does not need to worry about overhyped results due to any hyperparameter choices that were made this way. However, the latter statement already implies the main drawback of pre-registration, namely, that hyperparameters cannot be tuned based on the classification problem at hand. The major advantage of pre-registration, on the other hand, is that the complete dataset can be used. This differs for the second approach we suggest for minimising overhyping, *the lock box approach*.

The lock box is a portion of the dataset that is set aside at the very beginning, i.e., immediately after data collection concludes. The other portion of the dataset is called the hyperparameter optimisation set. The latter can be freely used for any optimisation steps in the analysis pipeline. This can involve tuning machine learning hyperparameters, such as which algorithm to use or how strongly to penalise model complexity, as well as any neuroimaging pre-processing steps (e.g., filtering, artifact correction/rejection, channel/voxel selection). Naturally, one still needs to implement precautions against standard overfitting, such as cross-validation, for example, when exploring the performance of different classification algorithms on the optimisation dataset. Only after *all* methodological decisions are made (importantly, this involves machine learning *as well as* data pre-processing steps), is the lock box dataset is used again. Crucially, the analysis pipeline is applied to the lock box dataset *once* and whatever accuracy results are obtained will be presented or published. It is

fundamental for the lock box approach that the lock box dataset is only used once to assess classification performance after all decisions about the analysis pipeline are finalised. If a researcher would like to change any analysis decisions after accessing the lock box dataset, a secondary lock box is required to assess classification performance after those secondary decisions are made. Even though the lock box approach is a good safeguard against overhyping, it does come at the cost of requiring a portion of the dataset to be set aside. This might lead to decreased statistical power, especially in the context of neuroimaging where datasets are typically expensive to collect. The following approach does not suffer from this drawback, whilst simultaneously being very effective at minimising overhyping, which could explain why it thus far was the most popular choice by studies that cited our corresponding journal paper (Hosseini et al., 2020).

Nested cross-validation (Cawley & Talbot, 2010; Stone, 1974) can be regarded as an extension of the lock box approach and consists of an inner and an outer cross-validation (CV) loop. The outer CV loop is similar to multiple lock boxes, meaning that it is comprised by separate hold-out (i.e., lock box) and inner (hyperparameter) optimisation datasets. In this context, each hold-out set corresponds to a different hyperparameter configuration, for example involving the four classification algorithms of this chapter (LDA & SVMs with three different kernels). *Within each outer CV loop,* an inner CV loop is performed as described previously, which may be used for optimising parameters, or, more generally, analysis decisions, of that given algorithm. For example, one may commence with the first outer CV loop utilising the LDA classifier. One quarter of the data is used at the lock box of this given outer CV loop and set aside (i.e., the hold-out set). Analysis choices are then made using the typical k-fold inner CV approach. Once these are finalised for the LDA classifier, one applies the algorithm *once* to the current lock box (or, hold-out) set and records classification performance. This is repeated for the three SVM classifiers, with the other three quarters of the full dataset being the outer lock box (hold-out) dataset once. Therefore, each quarter of the full dataset was the hold-out set for the outer CV loop once, similar to standard k-fold CV. Finally, one computes and reports the average classification accuracy across the four outer CV loop. Nested CV is a neat method for preventing overhyping whilst being able to utilise the full dataset. However, challenges concerning the interpretability of results may arise due to outer CV folds' accuracy scores being averaged. If, for example, an EEG or fMRI region of interest (ROI) was the hyperparameter that was distinct in different outer CV folds, it will be difficult to understand which ROI contributed to what extent to the final, averaged accuracy value.

The final approach we propose to limit and diagnose overhyping is *the blind analysis approach*. In a blind analysis, the dataset's class labels are randomised or hidden prior to any optimisation steps, thus preventing overhyping, since hyperparameters cannot be tuned based on any desired outcome (e.g., high classification accuracy). One way of conducting a blind analysis is to randomly label classes (such as we did in the current chapter to simulate null data) and to then add some form of artificial signal to the data. This allows the optimisation to be based on maximising classification efficacy on decoding this artificial signal. After deciding on hyperparameters using this artificial signal and scrambled dataset, the signal is removed, the true class labels are incorporated and the classifier, implementing the hyperparameters of choice, is applied to the research question of interest. A similar method is the orthogonal contrast (Bowman et al., 2020; Brooks et al., 2017), in which a contrast that is of no interest to the research question at hand is used for the optimisation of hyperparameters. For example, in a dual-stream attentional blink (AB) RSVP paradigm in which the research question of interest would be to differentiate trials based on whether the second target was reported correctly or not, the optimisation phase could be based on maximising classification efficacy in differentiating whether targets were presented in the left or right RSVP stream. Such blind analyses carry the advantage that the complete dataset can be used. However, as the optimisation is not based on the research question of interest, the choice of hyperparameters is not guided by the variable one is attempting to decode. This may thus lead to suboptimal hyperparameter choices.

The recommendation of one or another safeguard measure against overhyping depends on the context. Pre-registration is recommended in cases in which the analysis procedure has been determined a-priori. The lock box approach is especially suited in cases in which hyperparameters are tuned manually and are of interest for their subsequent interpretation and the dataset at hand is comparatively large. Nested cross-validation, on the other hand, is suited whenever hyperparameters are determined automatically and interpretability of hyperparameters is not a concern. Finally, blind analyses are recommended if the data has variability that is orthogonal to the main goal of decoding (e.g., to determine a suitable time-window or brain region of analysis). For a more thorough discussion of overhyping, the preventive measures against it, as well as a concrete example of nested CV, see the corresponding journal article (Hosseini et al., 2020).

Chapter 9 – Discussion

In this thesis, we have investigated perceptual event integration in rapid stimulus streams, pursuing the main goal of providing new insight about the neuronal mechanisms underlying two cases of our visual perceptual system reaching its limits: the distractor intrusion phenomenon and temporal event integration. With respect to our analysis of the distractor intrusion phenomenon presented in Chapters 4-6, most of our work has revolved around the 2f-ST^2 computational model. In this context, we have provided evidence in favour of the hypothesis that the binding of multi-dimensional stimulus features into a single percept depends on the timing of transient attentional enhancement (TAE). Further, we have shown that the 2f-ST^2 model accounts for a broad range of empirical findings obtained with distractor intrusion experiments, ranging from behavioural to electrophysiological results obtained with humans. Chapter 5 presented the loss of responsiveness phenomenon of the 2f-ST^2 model, which provided insight into how TAE impacts the feature-binding process. It was concretely suggested that TAE-timing is centred around the target stimulus's temporal position in the RSVP stream. In Chapter 6, we presented a series of empirical analyses supporting the hypothesis that increased key feature salience induces less temporally variable TAE deployment in the context of distractor intrusions. This interaction between key feature salience and the temporal variability underlying TAE was furthermore successfully modelled with the 2f-ST^2 model, yielding, as we argue, the most plausible, complete and accurate version of our computational model.

Our work on distractor intrusion errors suggested event integration to occur at 300 – 450 ms in the context of feature-binding. Expanding on this finding in Chapter 7, we investigated whether the temporal integration of two combinable stimuli into a single perceptual event occurs with temporal and electrophysiological characteristics that are consistent with the distractor intrusion phenomenon and the 2f-ST^2 model. Temporal generalisation as well as topographical ERP analyses have provided evidence in favour of this hypothesis. Finally, Chapter 8 has provided a series of simulation analyses that illustrated an important risk when adopting machine learning algorithms for the analysis of neuroimaging datasets, such as those presented in Chapter 7. In this context, we have provided evidence in favour of the hypothesis that machine learning algorithms applied to neuroimaging datasets, particularly temporal generalisation analysis used to determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfit to hyperparameters) classification results if appropriate preventive measures are not

implemented. Next, we will discuss the findings obtained with the 2f-ST^2 model, specifically comparing our computational model and how it accounts for distractor intrusions to the other, previously introduced, models and theoretical accounts. We emphasise the 2f-ST^2 model in this discussion since our computational model, to us, constitutes the main contribution of this thesis. We will subsequently provide a set of concluding remarks.

The 2f-ST^2 computational model and how it contributes to a landscape of theoretical accounts

Vul and colleagues (Vul et al., 2009) express, appropriately, that cognitive experiments regularly yield graded distributions of subjects' responses, which often tempts authors to conclude that the process of selecting a response *itself* is inherently graded in nature. Such graded (i.e. non-uniform) response distributions are often found in experiments of distractor intrusions (Botella et al., 2001; Botella & Eriksen, 1992), being centred around correct reports of target identity. Vul et al. (2009) further emphasize and examine the possibility of the response selection process being all-or-none *within trials* instead and that the graded response distributions often shown result from *across trial* variability. Based on two experiments in which subjects were asked to guess the identity of the target stimulus multiple times (i.e., first guess, second guess, etc), the authors propose a contiguous-graded selection account. According to this account, in their experiment, multiple responses are sampled to varying degrees in each trial. Importantly, across trials, this sampling procedure is always based on the *same distribution* of responses. The authors formulated the latter after observing that irrespective of the temporal position of the target in the RSVP stream as well as the identity of guess one, for example being the target or an (e.g., -2 or +1) intruder, responses provided as second guesses always followed the same distribution (Vul et al., 2009). These empirical results can be evaluated in the framework of the 2f-ST^2 model as follows. According to the 2f-ST^2, after each trial only *one* key and *one* response type are bound to *one* token (called *tokenization*). This implies that only one item is encoded into each episode of WM, as tokenization takes places serially in our model. Thus, this approach seems to stand against the account of Vul and colleagues (2009), which states that multiple responses are available in a graded manner at once. The hypothetical case of asking our model to guess multiple times, which is the empirical paradigm used in Vul et al.'s (2009) experiments, could, though, still yield the pattern that the authors found: a graded response distribution *centred around the first guess*. We expect this pattern because even though only one item is encoded into WM, other items' representations, or types, are still processed on a

pre-encoding level and, importantly, are affected by the TAE (the blaster's enhancement) *to differing degrees.* Whichever item was bound to the token, and hence encoded into WM, had its activation level substantially increased by the blaster. It then follows that items in temporal proximity to the reported item received a similarly substantial enhancement to their representation and would thus be more likely to be reported as subsequent guesses compared to other, temporally more distant, distractors. This is particularly well in line with an observation Vul and colleagues (Vul et al., 2009) emphasize themselves, namely that the distribution of the second guess *was not centred* around the position of *the target item*, but of the item reported as *the first guess*. We stress that in the framework of the 2f-ST^2 model, the amplitude of the lateral inhibition present in the response pathway would be critical for whether and to what extent second guesses can be provided. If the competition caused by the lateral inhibition of response types would be set up to be particularly strong, only the winning type would be provided as the first guess and subsequent guesses would be given at random. This would be expected since the winning type would inhibit competing types so strongly that their representations would not be available. However, if the lateral inhibition should be weak, we would expect the distribution of the second guess to be centred around that of the first guess for the reasons provided above. Importantly, we argue that in experiments, such as those ran by Vul et al. (2009), in which participants know that they will be required to provide multiple guesses, the brain might configure the competition between response feature representations (i.e., types in the 2f-ST^2 model) to be weaker compared to experiments that only require one response. A similar thought was provided by Zivony and Eimer (2020b), who stress the importance of task instructions, stating that a higher encoding threshold might be implemented in a top-down manner by the brain if only one response has to be provided.

Another important account is the model of Botella and colleagues (Botella et al., 2001). We would argue that 2f-ST^2 offers a more parsimonious, yet complete account than the Botella et al.'s (2001) model. Botella et al.'s (2001) model does not specify the exact neural mechanisms underlying its successful focusing and sophisticated guessing routes, which are fundamental to its architecture. In contrast, the 2f-ST^2 model comprehensively describes how the distractor intrusion phenomenon is realised on a neural level. The latter is achieved via one key and one response type, which are enhanced by the blaster, being bound to one token in each episode of WM encoding. Critically, the 2f-ST^2 model never has any information about which stimulus is the correct response in a given trial but instead achieves correct reports, intrusion errors as well as misses with the same underlying architecture. As a result, correct responses are effectively just fortunate conjunctions, and different reports are always

only a matter of which types are bound (or not bound in the case of misses) to which token. Thus, even though the 2f-ST^2 model realizes the distractor intrusion phenomenon in more detail than Botella et al.'s (2001) model, the mechanism allowing for all varieties of reports is in itself more parsimonious than that proposed by Botella et al. (2001). Importantly, 2f-ST^2 is able to account for a set of counter-intuitive RT data (Botella, 1992), which originally inspired the dual-route model-architecture of Botella et al.'s (2001) model. We thus provide support for parallel models of selective attention, such as the (informal) code coordination model (Keele & Neill, 1978), claimed to be incompatible with this specific pattern of RT data (Botella et al., 2001).

However, there are ways in which Botella et al.'s (2001) and 2f-ST^2 models are similar. Firstly, Module K is triggered by the relevant key feature in Botella et al.'s (2001) model, which resembles the task-filter implemented in 2f-ST^2 model's key pathway. In Botella et al.'s (2001) model this mechanism defines the critical time, t_c, whereas in the 2f-ST^2 model it leads to the blaster circuit being activated and enhancing activation levels of later stage 1 layers of both pathways. Even though these two mechanisms are different, there is a degree of conceptual similarity between t_c and the blaster, as once the blaster is fired, a critical time-window emerges in which the currently most active key and response types are ultimately bound to the currently available token. Secondly, Module R continuously makes responses available in Botella et al.'s (2001) model, which is close to the behaviour of later stage 1 layers of the 2f-ST^2 model's response pathway. Further, in Botella et al.'s (2001) model, the magnitudes of item representations in Module R at t_c determine the likelihood of a given item in the critical set (-2 to +2 intruder as well as correct item) being reported. The concept of item magnitude determining the likelihood of report is again akin to the 2f-ST^2 model, in which the activation strengths (specifically the membrane potentials) of response TFL/ type neurons determine which response type will be bound to the token and hence encoded into WM. Importantly, the Module R mechanism in Botella et al.'s (2001) model only explains the results of sophisticated guessing, as the successful focusing route exclusively leads to correct reports. As mentioned earlier, this is fundamentally different to the 2f-ST^2 model.

As indicated before, we would argue that the 2f-ST^2 model is closely related to the theoretical account proposed in Zivony and Eimer's work (2020b, 2020a). Our model is further in line with most empirical patterns demonstrated in the authors' experiments (Zivony & Eimer, 2020b, 2020a). The initial temporal variability account provided in the authors'

earlier paper (Zivony & Eimer, 2020a) conceptually explains intrusion errors exactly how the 2f-ST^2 model does. Specifically, the common idea is that intrusion errors emerge due to temporal variability underlying a TAE mechanism. In Zivony and Eimer's (2020a) account, the item with the strongest representation *at the end of the TAE's amplification time interval* is bound into WM. This idea is notably similar to that presented in Botella et al.'s (2001) sophisticated guessing mechanism, in which the item with the largest magnitude *at the critical time, t_c,* is encoded. In the 2f-ST^2 model, the propagation of information through the model's architecture is always a matter of neurons crossing their membrane potential thresholds. The latter applies to the connection between the response TFL/ type layer and the corresponding response binding pool, too, which ultimately leads to a given response type being bound to a token and encoded into WM. However, in the 2f-ST^2 model, which response type is bound to the token is *not exclusively* a matter of the blaster's temporal variability, as, for example, stated in Zivony and Eimer's (2020a) original account. We further implemented variability in stimuli's input strengths, which, similar to the blaster's boost varying between trials, also affects the likelihood and ease of response TFL/ type neurons crossing the threshold to the binding pool. Moreover, the idea of fast and slow attentional engagement presented in Zivony and Eimer (2020a) implies that salient key features, such as colour, which are easily detected by our attentional system, should lead to the TAE occurring swiftly compared to when key features are less salient, such as an annulus surrounding the target item. This, in turn, affects the likelihood of correct reports and post-target intrusion errors, making correct reports more likely as key feature detection is made easier, as shown in the three experiments presented in Zivony and Eimer (2020a). This interaction between the TAE and report likelihood of the correct and intruder item is also in line with the 2f-ST^2 model. Saliency of key features can directly be modulated via τK, the delay added to key pathway processing, which always shifts the response distribution of the model accordingly.

The authors' additional empirical results (Zivony & Eimer, 2020b) inspired a modification of their original account. An essential question in this paper was where in the processing pipeline the competition between the target and intruder item takes place. The 2f-ST^2 model locates this competition in the response TFL/ type layer using lateral inhibition between neurons. This is in line with Zivony and Eimer's (2020b) proposition that the target-intruder competition should take place at a pre-encoding stage, as the response TFL/ type layer is the last parallel layer of the 2f-ST^2 model's stage 1 and hence pre-encoding in nature. The authors' modified account furthermore specifically states that the target-intruder competition does not necessarily block entry of one item into WM but decreases the

likelihood thereof substantially (compared to when no intruder is present). Whether one or both items are encoded into WM is based on whether an encoding threshold is crossed. If both items cross this threshold, both will be encoded. This concept is very similar to that of *joint tokenization* in the original ST^2-model (Bowman & Wyble, 2007). Joint tokenization in the original ST^2 accounts for lag 1 sparing in attentional blink experiments, in which report accuracy of two targets is high if they are presented in immediate succession in RSVP, but drops significantly if a distractor item, which creates a temporal offset between targets, is presented in-between. Interestingly, the deficiency in performance due to distractors being presented in-between targets was shown to be based on time elapsed after the onset of the first target, rather than on the number of distractor stimuli presented (Bowman & Wyble, 2007). The original ST^2 models this phenomenon via two target types occasionally binding to the same token and hence being both encoded into WM (see Figure 1D in Simione et al. (2017)).

Moreover, Zivony and Eimer (2020b) proposing a decreased likelihood of either the target or the intruder being reported when an intruder is presented in the RSVP stream, due to the target-intruder competition this creates, compared to when only the target is presented (baseline) is compatible with the 2f-ST^2 model as well. As described above, the target-intruder competition is modelled via lateral inhibition in the response TFL/ type layer. This inhibitory interaction between response TFL/ type neurons means that more excitatory input is required for a given neuron to cross its membrane potential threshold to excite its corresponding binding pool unit. It therefore follows that if only one item would meet task-demands in the response dimension, such as the item being the only digit in a stream of letters, as was the case in Zivony and Eimer's (2020b, 2020a) paradigms, baseline report accuracy of our model would be substantially higher, too, since no lateral inhibition in the response TFL/ type layer is affecting the target neuron of that layer. Nonetheless, one key mechanism proposed in the author's modified account is not modelled in the 2f-ST^2 model, namely that the target as well as the intruder are both encoded into WM at times. In modelling terms, the above introduced concept of joint tokenization existent in the original ST^2 (Bowman & Wyble, 2007) has not been implemented in 2f-ST^2, rather we have implemented lateral inhibition, which induces winner-take-all dynamics. Zivony and Eimer (2020b) propose that especially when the TAE is deployed fast, the intruder's representation should regularly be strong enough to cross the encoding threshold and hence be encoded into WM in addition to the target. They base this proposition on their fourth experiment (Zivony & Eimer, 2020b), in which an intruder was shown after the target *in every trial*. Compared to their first experiment, in which targets were regularly presented without a subsequent intruder,

participants were much more likely to report both items in their two responses. This observation further inspired the authors to propose that the encoding threshold should be flexible in human cognition based on different top-down control settings (Zivony & Eimer, 2020b). The 2f-ST^2 model was specifically designed to model distractor intrusion experiments, in which the task instruction is always to report one single target item per trial, hence the strong lateral-inhibition. However, strategic reduction of the strength of lateral-inhibition in the final layer of the response pathway, could be investigated to model these joint-binding effects. Indeed, and as introduced above in the context of Vul et al.'s (2009) experiments, Zivony and Eimer (2020b) acknowledge the significance of different task-instructions themselves whilst elaborating on their findings in light of models of the attentional blink and selective attention in general.

Chapter 10 – Conclusion

This thesis investigated the neuronal mechanisms and processes underlying perceptual event integration in rapid stimulus (i.e., RSVP) streams. To this end, the following core hypotheses were studied in the research chapters 4-8.

1 During rapid stimulus streams, the binding of multi-dimensional stimulus features into one percept depends on the timing of transient attentional enhancement (TAE).

Evidence for Hypothesis 1 was provided in Chapters 4-6. Particularly the empirical finding of correlated N2pc & P3 ERP component latencies, key feature salience and τK respectively modulating empirical and virtual response distributions in the distractor intrusion context, the central role of the blaster in the 2f-ST^2 model and the loss of responsiveness phenomenon supported this hypothesis.

2 The 2-feature Simultaneous Type, Serial Token (2f-ST^2) model accounts for a broad range of findings obtained with distractor intrusion experiments, suggesting that correct and intrusion reports are *qualitatively* the same, implying an interesting contrast to the dual-route model of Botella et al. (2001).

Hypothesis 2 was supported in Chapters 4-6. The 2f-ST^2 model qualitatively replicated empirical response distributions, reaction time data, and electrophysiological findings across two different empirical paradigms and a range of key and response feature salience configurations. We stress that the 2f-ST^2 model accounts for this broad range of findings without the necessity of treating correct and intrusion percepts in a different manner, thus providing a more parsimonious account to that of Botella et al. (2001).

3 Increased key feature salience induces less temporally variable TAE deployment in human cognition, which can further be computationally modelled with the 2f-ST^2 model.

Hypothesis 3 was supported in Chapter 6. Our N2pc temporal jitter (TJ) analyses suggested that key feature salience particularly affects the onset-latency of TAE deployment. We also added Gamma Noise to the 2f-ST^2 model and demonstrated that it successfully accounts for the interaction between key feature salience and the temporal variability underlying TAE deployment. The addition of Gamma Noise to the 2f-ST^2 model further resolved a main limitation of the model presented in Chapters 4 & 5, i.e., that vERP latency differences were diminished when replicating Zivony and Eimer's (Zivony & Eimer, 2020a) experimental paradigm. We therefore argue that the 2f-ST^2

model presented in Chapter 6 constitutes the most complete version of our computational model.

4 The temporal integration of two combinable stimuli into a single perceptual event occurs with temporal and electrophysiological characteristics that are consistent with the distractor intrusion phenomenon and the 2f-ST^2 computational model.

Evidence for Hypothesis 4 was provided in Chapter 7, applying a machine learning approach to the electrophysiological data of Akyürek et al. (2017). Our results suggested that the temporal integration of two combinable stimuli according to Gestalt properties (Von Ehrenfels, 1937; Wagemans et al., 2012) into a single perceptual event occurs between 350 – 450 ms after the second target is presented. This time-frame is well in line with those found in the context of the distractor intrusion phenomenon and the 2f-ST^2 model.

5 Machine learning algorithms applied to neuroimaging datasets, particularly temporal generalisation analysis used to determine temporal electrophysiological characteristics, carry the easily observable and dire risk of overhyped (i.e., overfitting to hyperparameters) classification results if appropriate preventive measures are not implemented.

Hypothesis 5 was supported in Chapter 8. Presenting a series of simulation analyses, Chapter 8 demonstrated how easily classification results can be overhyped, thereby seriously threatening analyses' external validity, if preventive measures are lacking. In this context, we explained why the practice of cross-validation regularly does not sufficiently prevent overhyping and suggested a number of practices that effectively diagnose or limit the risks associated with overhyping.

To summarise, we provided a series of empirical, electrophysiological, and computational contributions, elucidating the cognitive dynamics underlying perceptual event integration in the visual domain. Placing a particular focus on rapid stimulus streams, we provided novel insight about two related cases of human perception integrating visual information, either binding multi-dimensional stimulus features or two separately presented but combinable stimuli into a single perceptual event.

Appendix Material

Appendix A – PO versus LB analyses with original C-Mass measure

As explained in Chapter 8, we initially ran the (hyper-)parameter (PO) versus lock box (LB) simulations with a squared C-Mass measure to measure the extent of above- as well as below-chance classification across a temporal generalization map. This C-Mass measure was computed by first subtracting chance-level classification (AUC of 0.5) from all AUC values of a given map, then squaring these values and finally taking the average of the entire map. The current appendix presents the equivalents of Figure 92 - Figure 96 in Figure 102 - Figure 106 in addition to equivalents of Appendix C's Figure 112 - Figure 115 in Figure 107 - Figure 110. The latter figures provide equivalents of the PO versus LB C-Mass distributions plotted in Figure 94 for each classifier and the different C-Mass measures separately and will be introduced in detail in Appendix C.

Figure 102 depicts an example map-triplet (similar to Figure 92), which shows larger below-chance classification in the Winning PO than in the LB. In these analyses, we also adopted a two-tailed variant of the cluster-extent permutation test introduced in the general methods section, in which above- as well as below-chance AUC-clusters could be formed, hence the statistically significant below-chance cluster in the map of Figure 102A. Figure 103 presents the bar plots that illustrate how winning, random and losing hyperparameters were distributed across the total of 1038 iterations we ran for these analyses. For the current analyses, we ran slightly more than the 1000 iterations presented in Chapter 8 because of differences in how parallel computing was utilized to run the simulations. Figure 103 can be regarded the equivalent of Figure 93. Furthermore, a pattern of systematic overhyping due to kernel selection similar to that presented in Figure 94 was found for the analyses implementing our initial C-Mass measure, too, as is shown in Figure 104. These results indicate overhyping with our initial C-Mass, since the PO C-Mass distribution has a higher mean (0.00081 / 8.1098*e-04), median (0.00071, 7.0974*e-04) and larger variance (1.3977*e-07) compared to the LB C-Mass distribution (mean: 0.000367 / 3.6650*e-04, median: 0.000281 / 2.8111*e-04), variance: 6.2168*e-08). The right column of Figure 104's top row illustrates how the *within-iteration* differences in C-Mass between PO and LB were distributed. This distribution should be centred around zero if no overhyping was observed (i.e., a given set of hyperparameters leading to similar success in decoding between-class differences for PO as well as LB null-data). This differs from the distributions in Figure 104's left column, which cannot be centred around zero (because zero is the lower limit of these

distributions), because the difference distributions being centred would reflect that the PO distribution sits on top of the LB distribution, which would imply no overhyping. The observed mean (0.000445 / 4.4448*e-04) and median (0.000391 / 3.9104*e-04) were positive, implying higher C-Mass in PO maps. The permutation test confirmed that both values were significantly different from zero ($p < 0.001$), which provides evidence for overhyping.

We again assessed overhyping and the degree thereof for each classification algorithm separately. Across all classifiers, the median difference in C-Mass between winning PO and LB was positive and significantly different from zero after performing the permutation test introduced in Chapter 8 (linear SVM: median = 0.00035 / 3.5331*e-04, n = 299; polynomial SVM: median = 0.00036 / 3.5839*e-04, n = 176; RBF SVM: median = 0.000239/ 2.3902*e-04, n = 110; LDA: median = 0.000462/ 4.6182*e-04, n = 453). We present the results of the Kruskal-Wallis test in Figure 105. Similar to the results presented in Chapter 8 (Figure 95), this analysis again provided evidence suggesting that overhyping is most pronounced after LDA classification. Figure 106 is the equivalent of Figure 96 and again suggests that more regularized (i.e., simpler) classification models lead to more variable C-Mass values, which provides an explanation for the pattern of such simple models winning and losing the PO competition more regularly (which, as shown in Figure 103, was demonstrated with this initial C-Mass measure, too).

Finally, Figure 107 - Figure 110 present the C-Mass distribution plots (similar to those plotted in Figure 94 & Figure 104) for each classification algorithm separately. A detailed introduction of these figures will be provided in Appendix C.

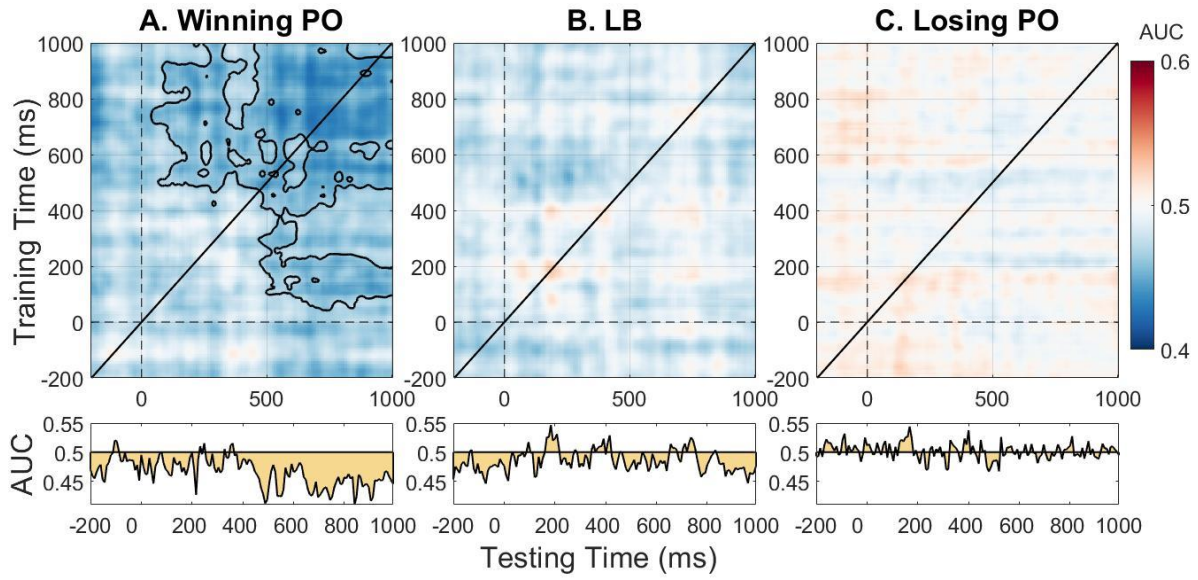


Figure 102. Example TG Map Triplet of initial C-Mass measure. Panel A illustrates the case of overhyping after increased misclassification of the testing dataset, which implies that the testing dataset was comprised of a brain activity pattern that was akin to the one the classifier was trained on, but of the opposing polarity. The Lock Box (Panel B) as well as the Losing PO (Panel C) did not demonstrate significant AUC patterns. Plotting conventions are identical to those of Figure 92.

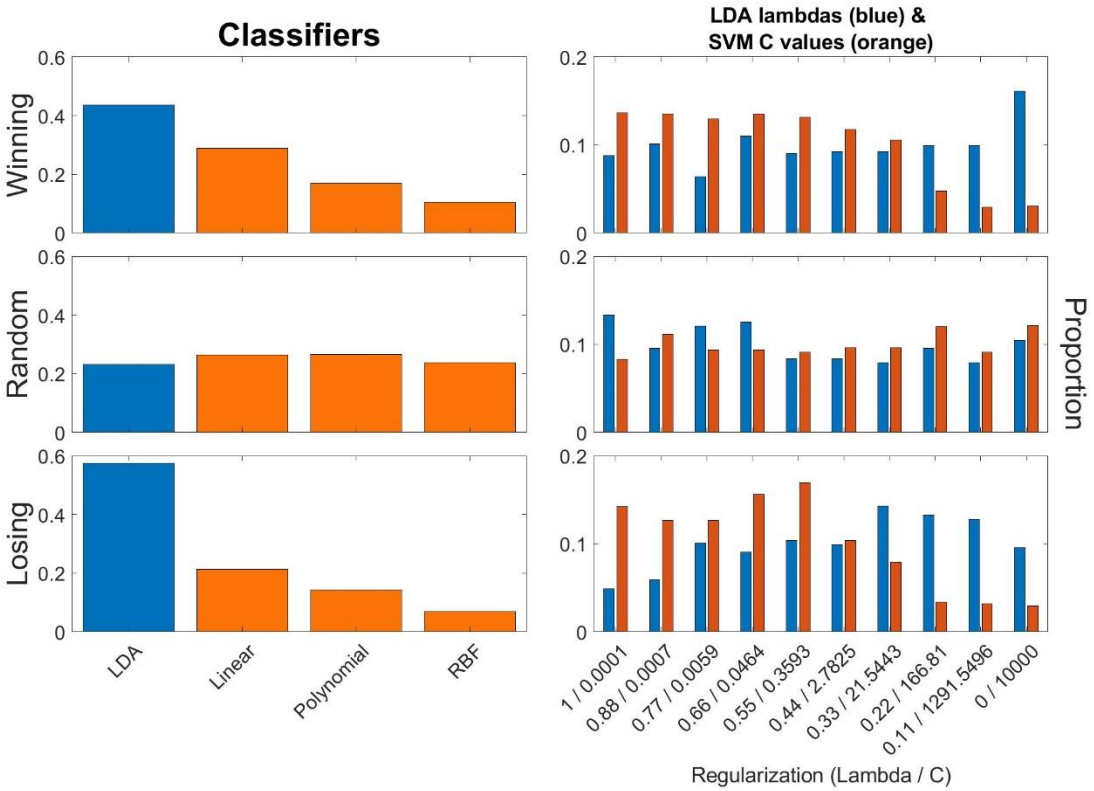


Figure 103. Barplots illustrating distributions of winners & losers of PO competition with initial C-Mass measure. Plotting conventions are identical to those shown in Figure 93.

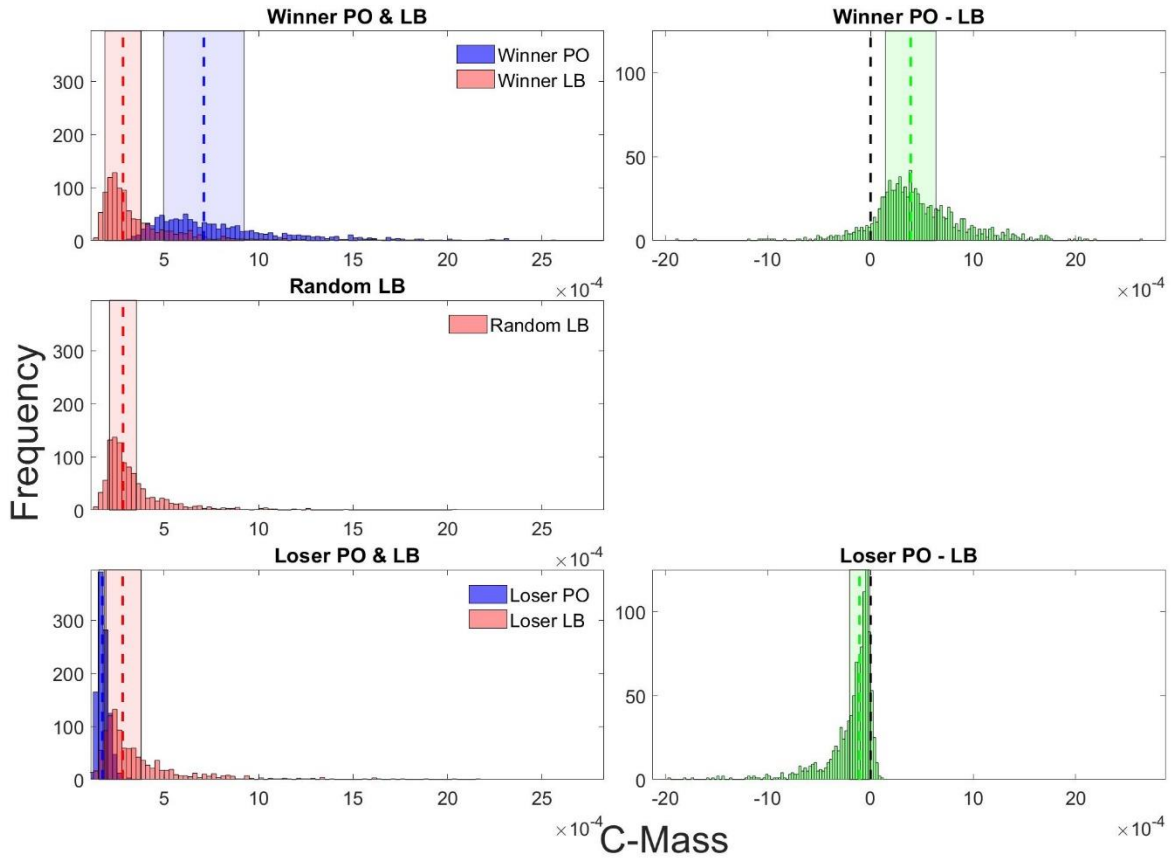
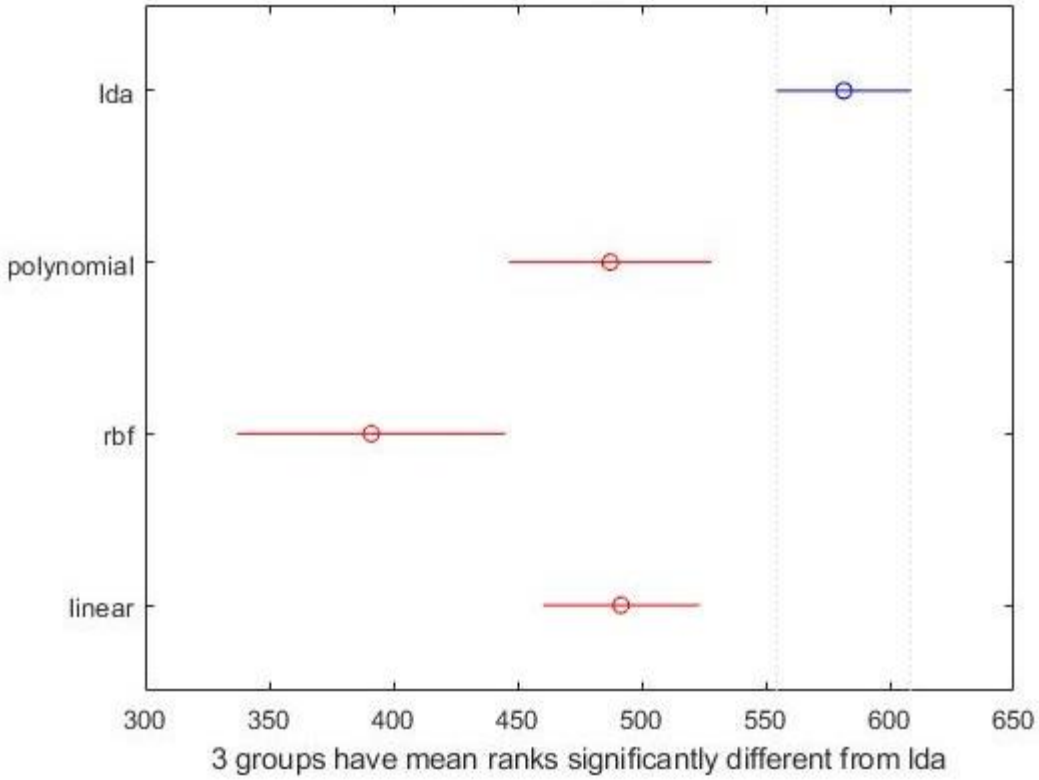


Figure 104. C-Mass distributions across all iterations for initial C-Mass measure. Plotting conventions are identical to those shown Figure 94.

A. ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	3.96694e+06	3	1322313.9	44.14	1.40987e-09
Error	8.92319e+07	1034	86297.8		
Total	9.31988e+07	1037			

B. Rank Distributions



C. Testing mean ranks between classifiers

```
*** GROUP 1: lda***

*** GROUP 2: polynomial***

*** GROUP 3: rbf***

*** GROUP 4: linear***

*** Group 1, Group 2, Lower Bnd, Estimate, Upper Bnd, P***
    1.0000    2.0000   25.6994   94.1072  162.5149    0.0023
    1.0000    3.0000  108.5509  190.4151  272.2793    0.0000
    1.0000    4.0000   32.3623   89.7487  147.1352    0.0003
    2.0000    3.0000    2.6994   96.3080  189.9165    0.0409
    2.0000    4.0000  -77.5296   -4.3584   68.8128    0.9987
    3.0000    4.0000 -186.5509 -100.6664  -14.7819    0.0139
```

Figure 105. Kruskall Wallis test of C-Masses across different classifiers for initial C-Mass measure. Similar to the results shown in Figure 95, overhyping was also most severe for LDA classifiers.

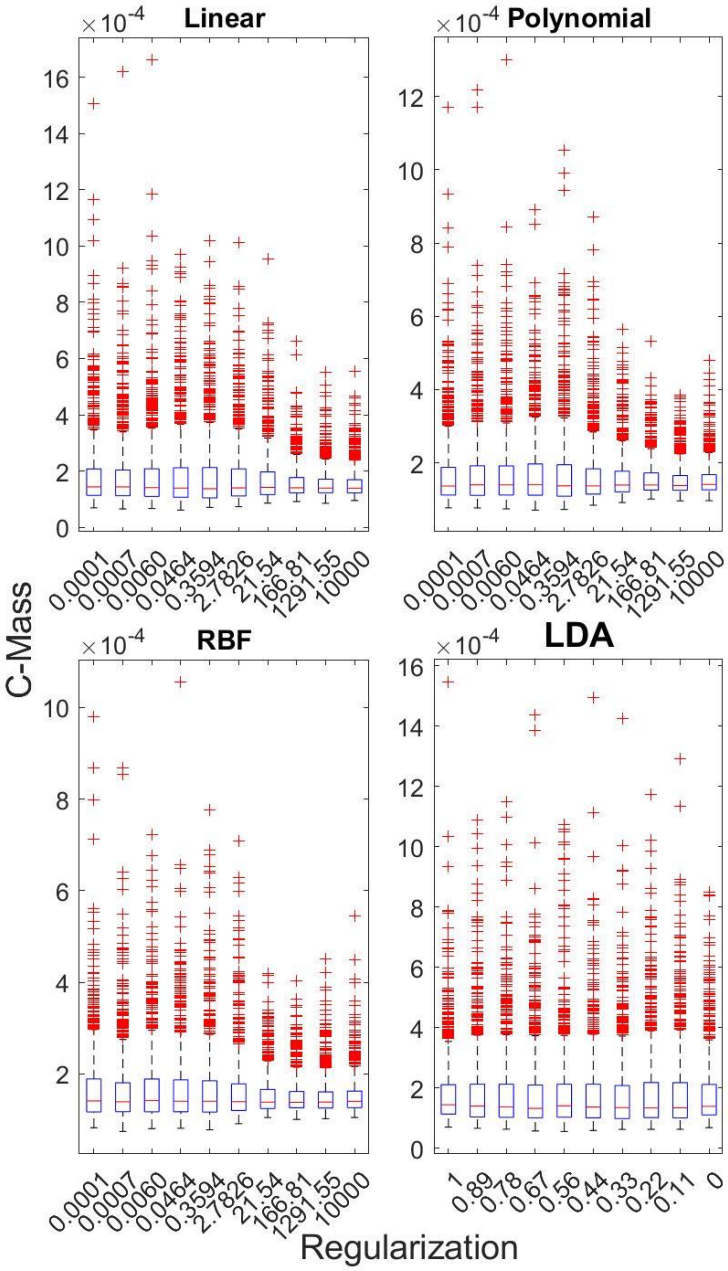


Figure 106. C-Mass boxplots of initial C-Mass measure across different classifiers and regularization values. Plotting conventions are identical to those presented in Figure 96.

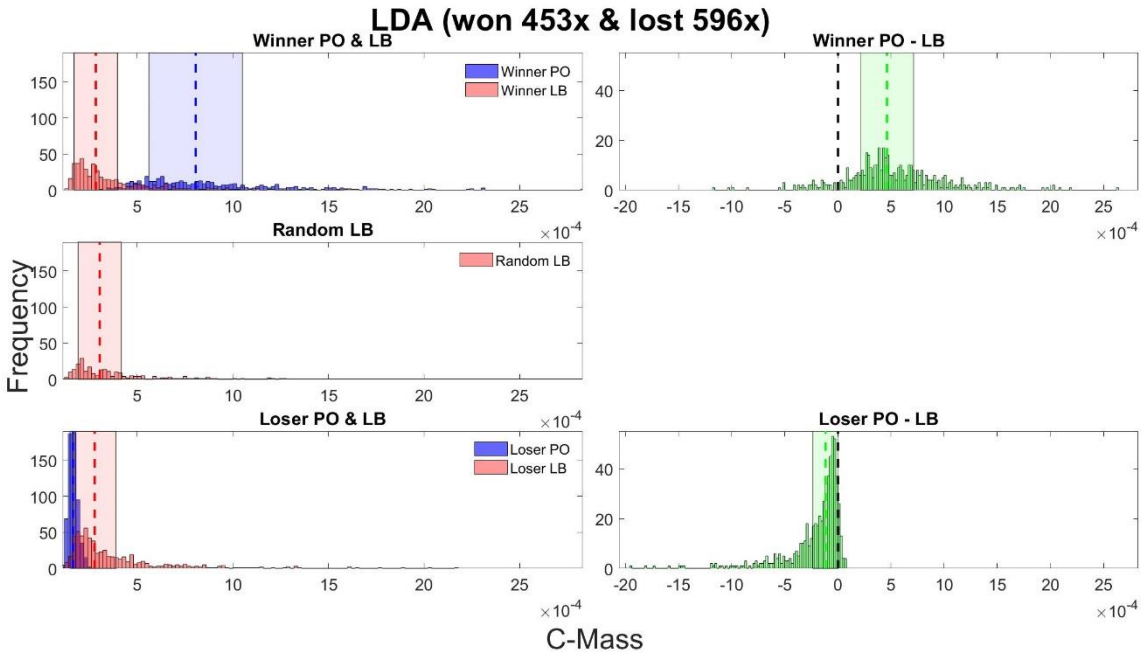


Figure 107. C-Mass distributions for the LDA classifier implementing our initial C-Mass measure. Plotting conventions are identical to those of Figure 94.

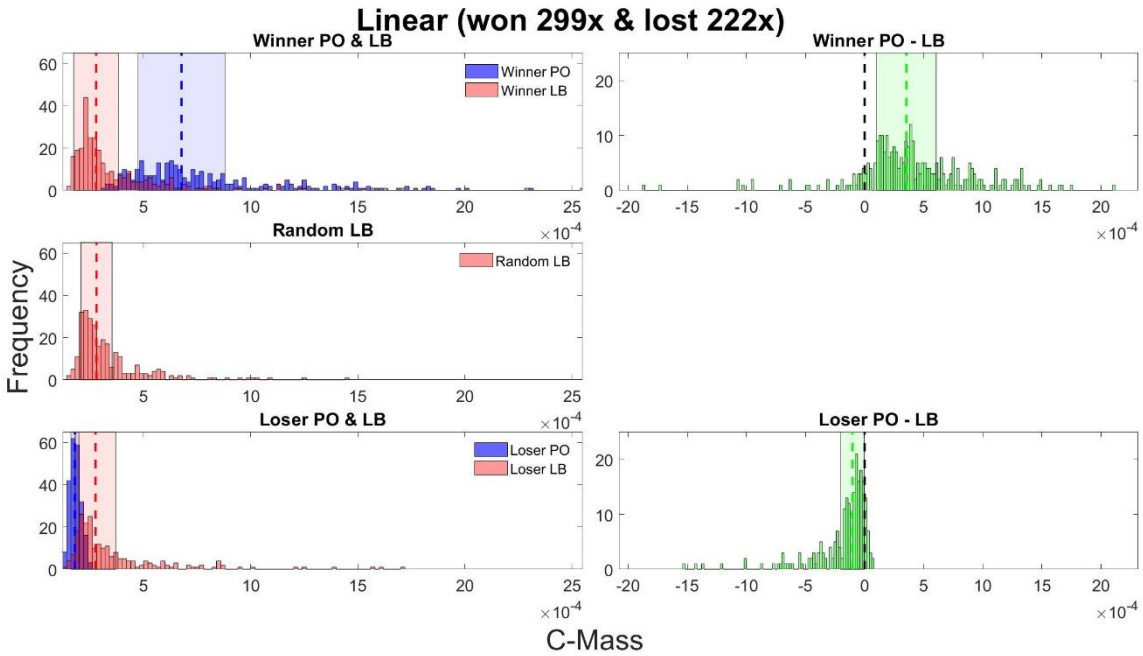


Figure 108. C-Mass distributions for the linear SVM classifier implementing our initial C-Mass measure. Plotting conventions are identical to those of Figure 94.

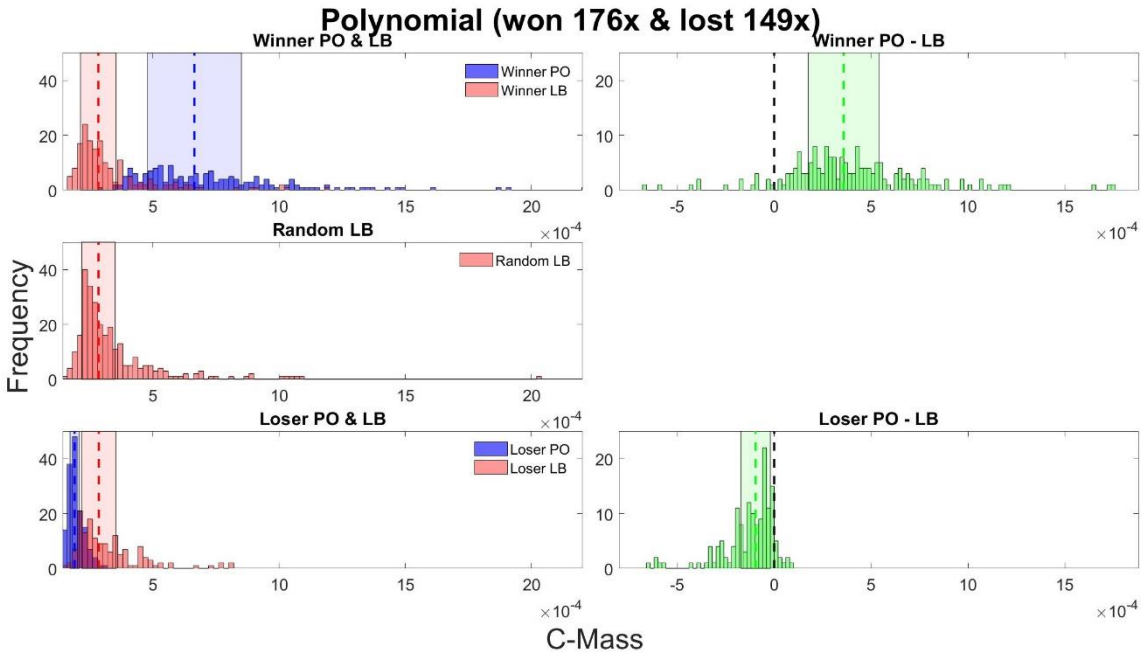


Figure 109. C-Mass distributions for the polynomial (order of 2) classifier implementing our initial C-Mass measure. Plotting conventions are identical to those of Figure 94.

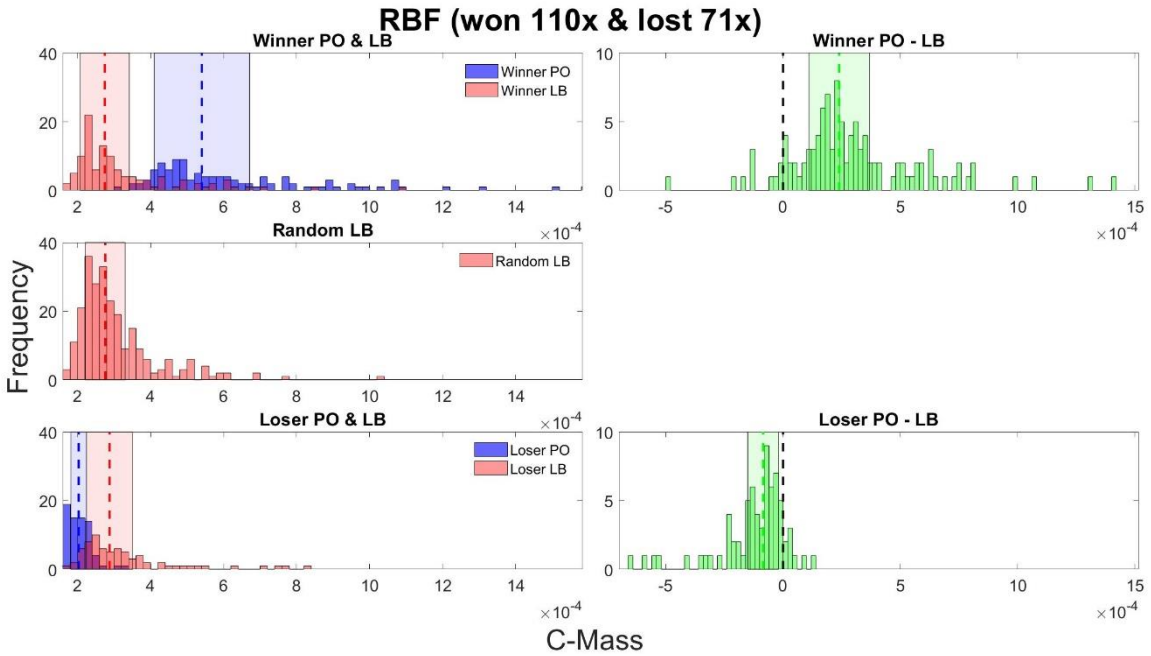
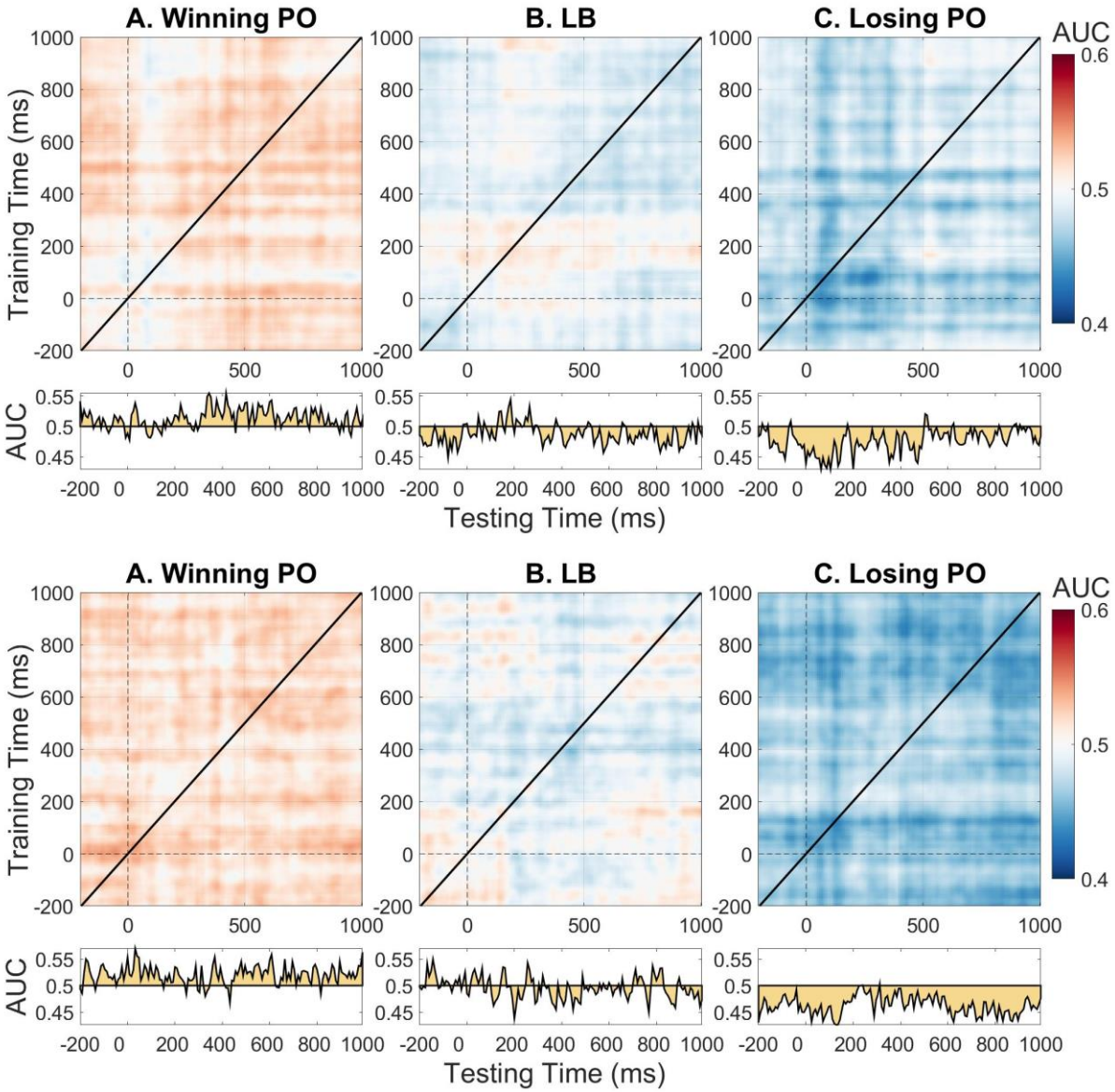
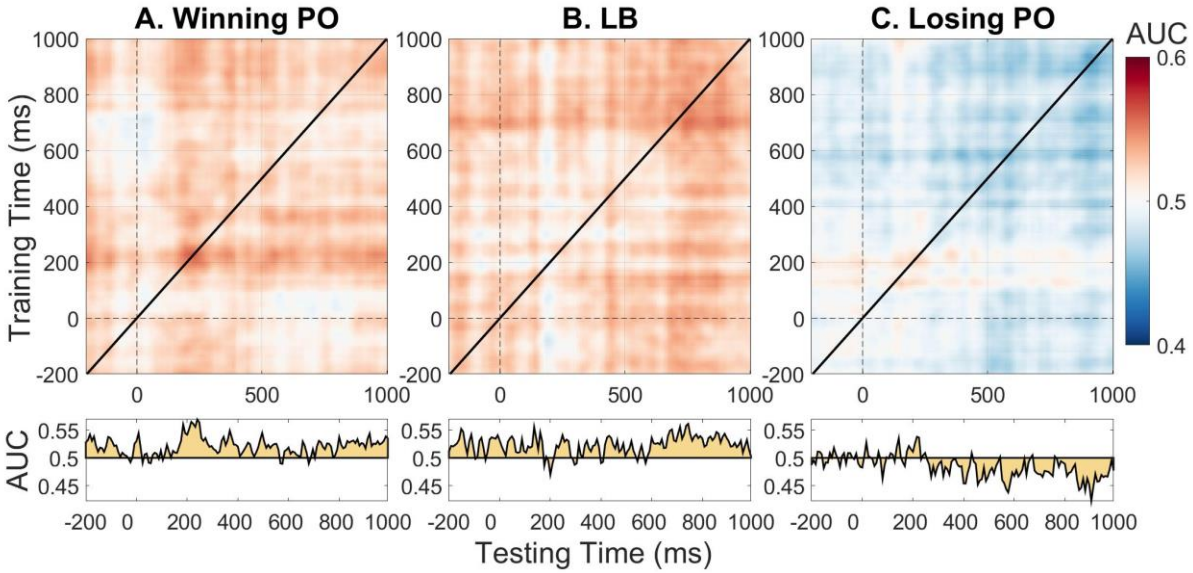


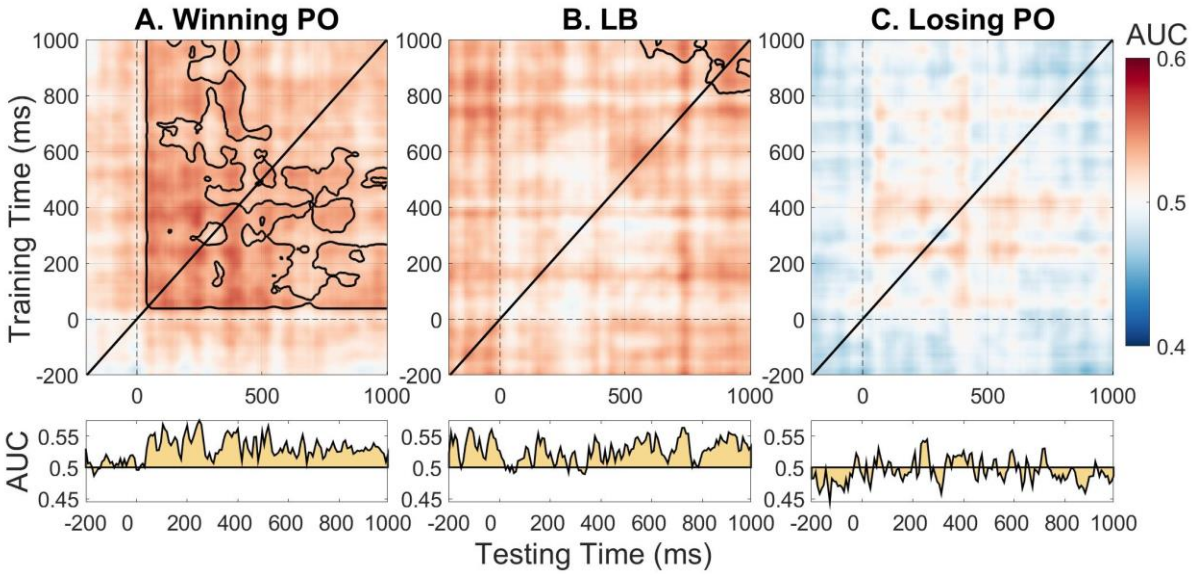
Figure 110. C-Mass distributions for the radial-basis function (RBF) SVM classifier implementing our initial C-Mass measure. Plotting conventions are identical to those of Figure 94.

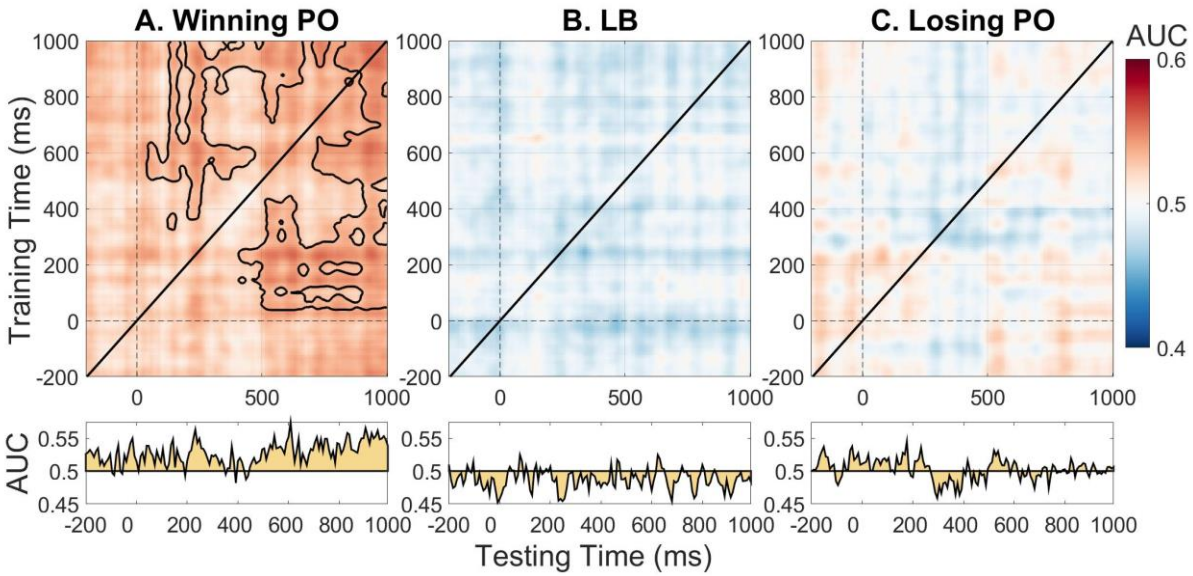
Appendix B – Randomly selected set of 9 Temporal Generalization Map Triplets

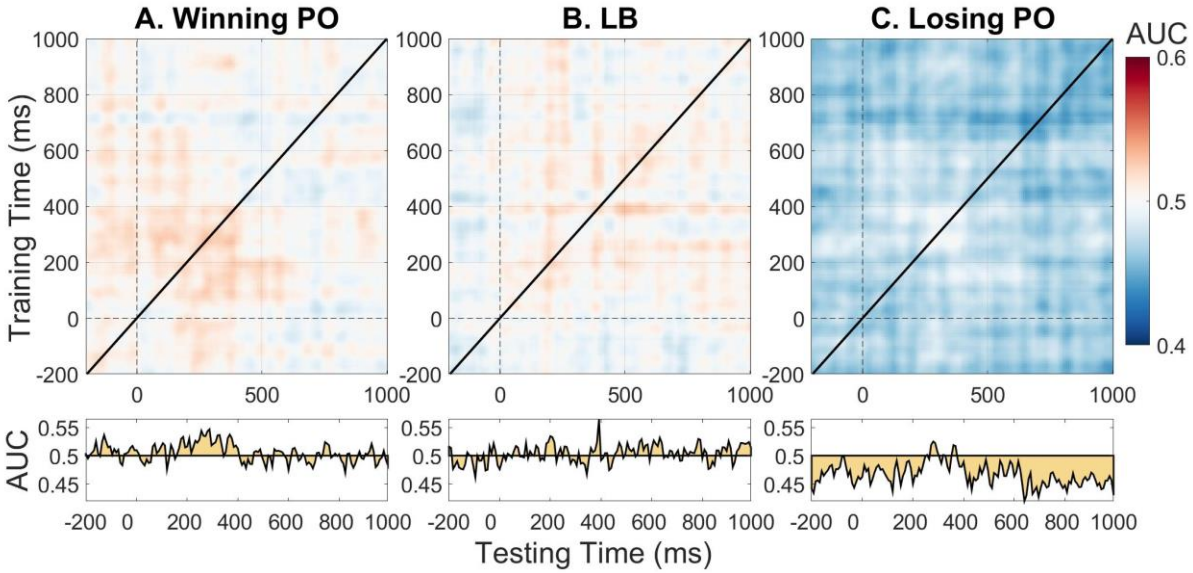
Figure 92 presented in Chapter 8 illustrates just one of the 1000 results that was obtained. In Figure 111, we show 9 additional samples from this distribution of 1000. These were selected at random, with bias neither towards significance of clusters, nor C-Mass. Three of the nine simulations selected at random have significant clusters in the Winning PO.

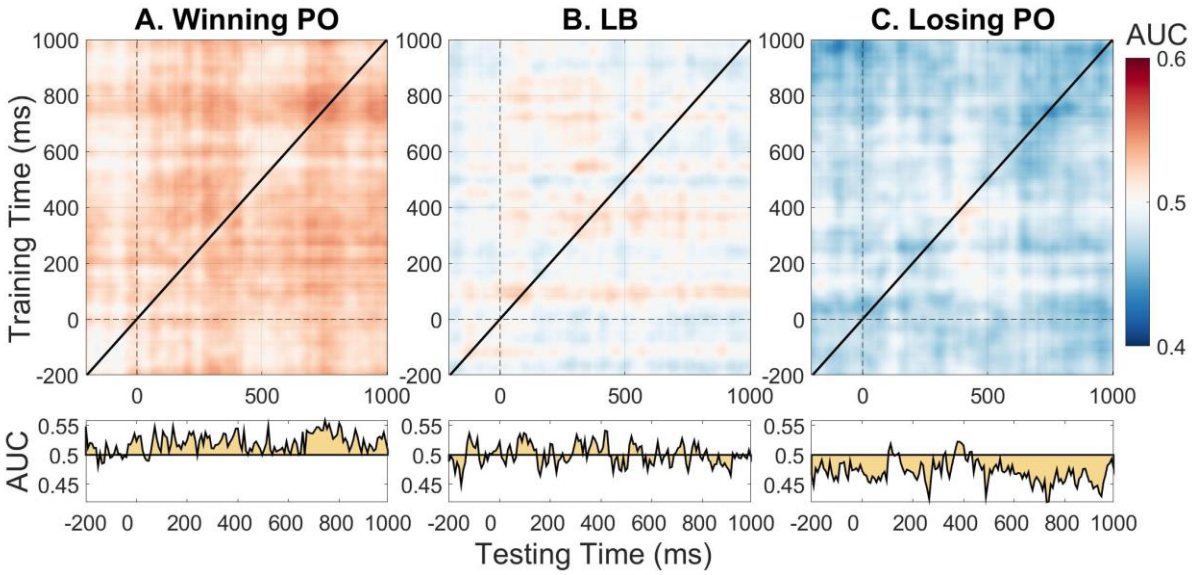



A. Winning PO
B. LB
C. Losing PO
AUC
Training Time (ms)
Testing Time (ms)


A. Winning PO
B. LB
C. Losing PO
AUC
Training Time (ms)
Testing Time (ms)


A. Winning PO
B. LB
C. Losing PO
AUC
Training Time (ms)
Testing Time (ms)


A. Winning PO
B. LB
C. Losing PO
AUC
Training Time (ms)
Testing Time (ms)


A. Winning PO
B. LB
C. Losing PO
AUC
Training Time (ms)
Testing Time (ms)

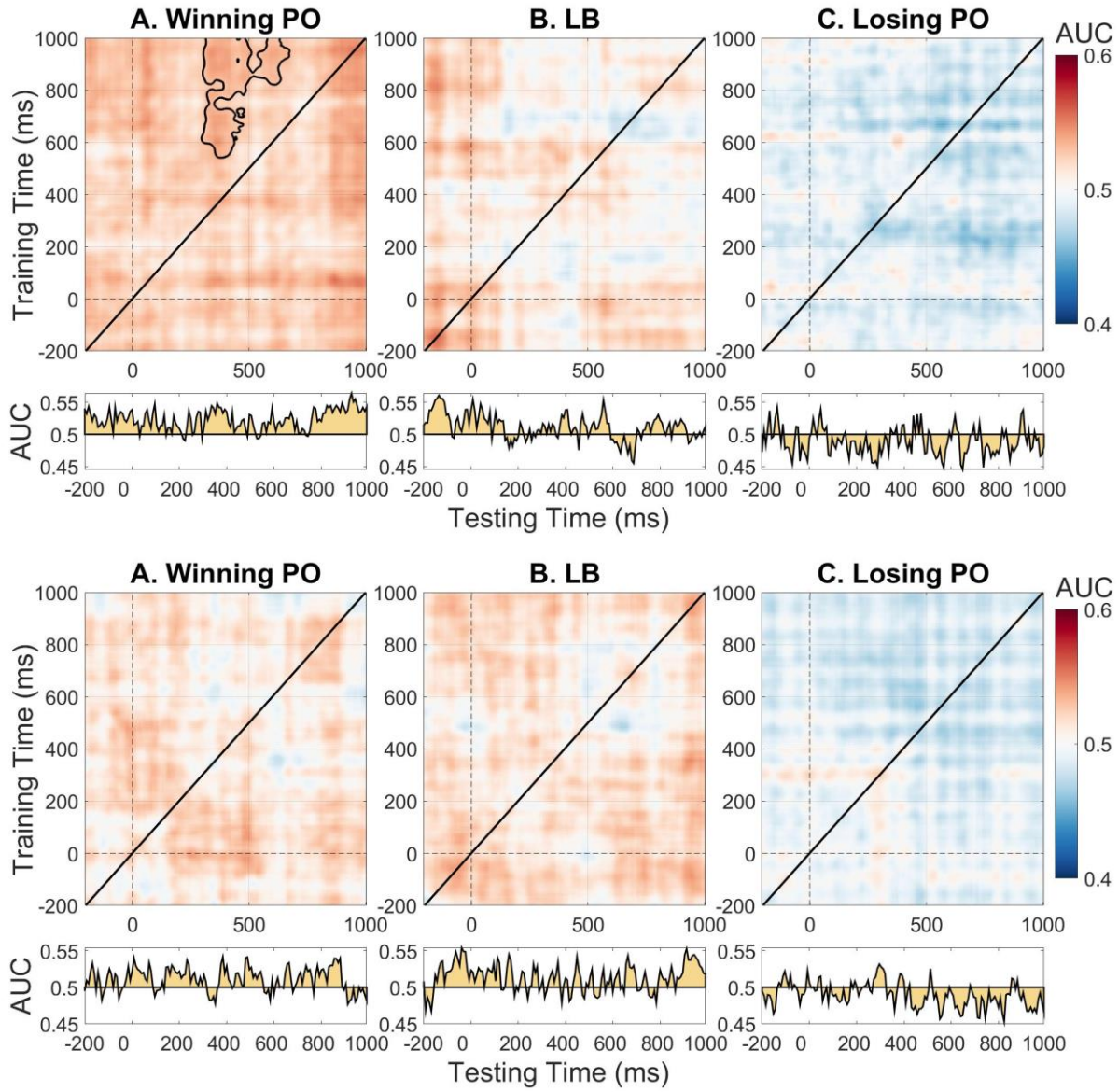


Figure 111. Nine additional temporal generalization map-triplets showing overhyping as demonstrated in Figure 92. These nine candidates were selected randomly and correspond to iteration numbers 98, 127, 279, 547, 633, 815, 906, 914 and 958 (in the order plotted above) of our PO competition. Panels A, B and C show the winning PO, the (winning) LB and the losing PO maps, respectively. Plotting conventions follow those of Figure 92. Temporal generalization maps and permutation tests' cluster-boundaries were again smoothed only to improve visualization and did not affect any analysis whatsoever.

Appendix C – PO & LB C-Mass distributions for each classifier

We present equivalents of Figure 94 for each respective classifier in Figure 112 - Figure 115 to further illustrate that overhyping was evident across all classification algorithms of our analyses (LDA and SVMs implementing a linear, polynomial (order of 2) and radial-basis function (RBF) kernel). For these plots we extracted all C-Mass values of those iterations in which the respective algorithm was determined a winner or loser of the PO competition. We similarly extracted the C-Mass values in which the respective classifier was randomly chosen for our *Random LB* panel (middle left panel) of Figure 94 to be able to plot the equivalent panels in the classifier-specific figures below. For example, the top left *Winner PO & LB* panel of Figure 112 shows all 350 iterations' C-Mass values in which the LDA classifier yielded maximum C-Mass. We provide information on how often a given classifier won and lost the PO competition in the figure-titles. Again, across all classifiers, the median difference in C-Mass between winning PO and LB was positive and significantly different from zero after performing the permutation test introduced in Chapter 8 (linear SVM: median = 0.0146, n = 329; polynomial SVM: median = 0.0125, n = 189; RBF SVM: median = 0.0127, n = 132; LDA: median = 0.0184, n = 350).

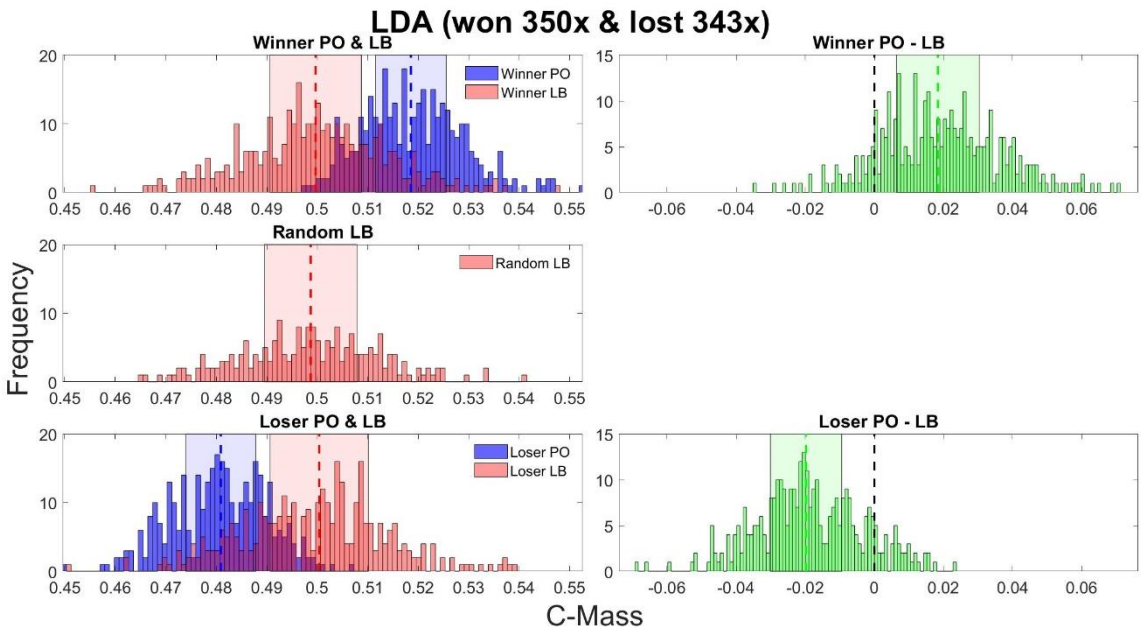


*Figure 112.*C-Mass distributions for the LDA classifier. Plotting conventions are identical to those of Figure 94.

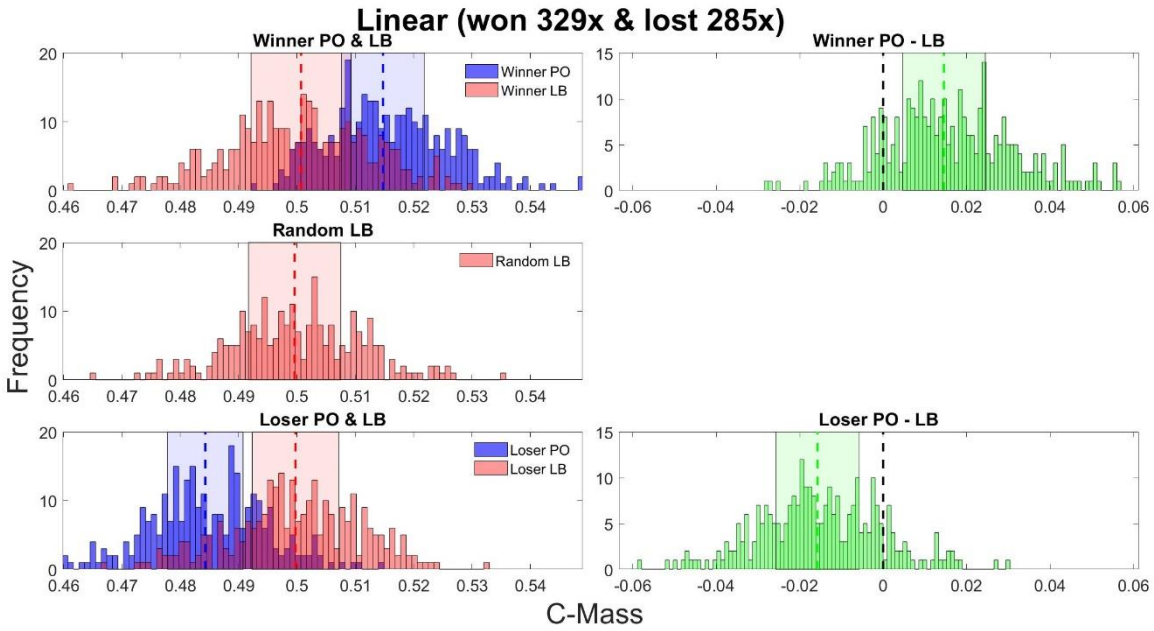


Figure 113. C-Mass distributions for the linear SVM classifier. Plotting conventions are identical to those of Figure 94.

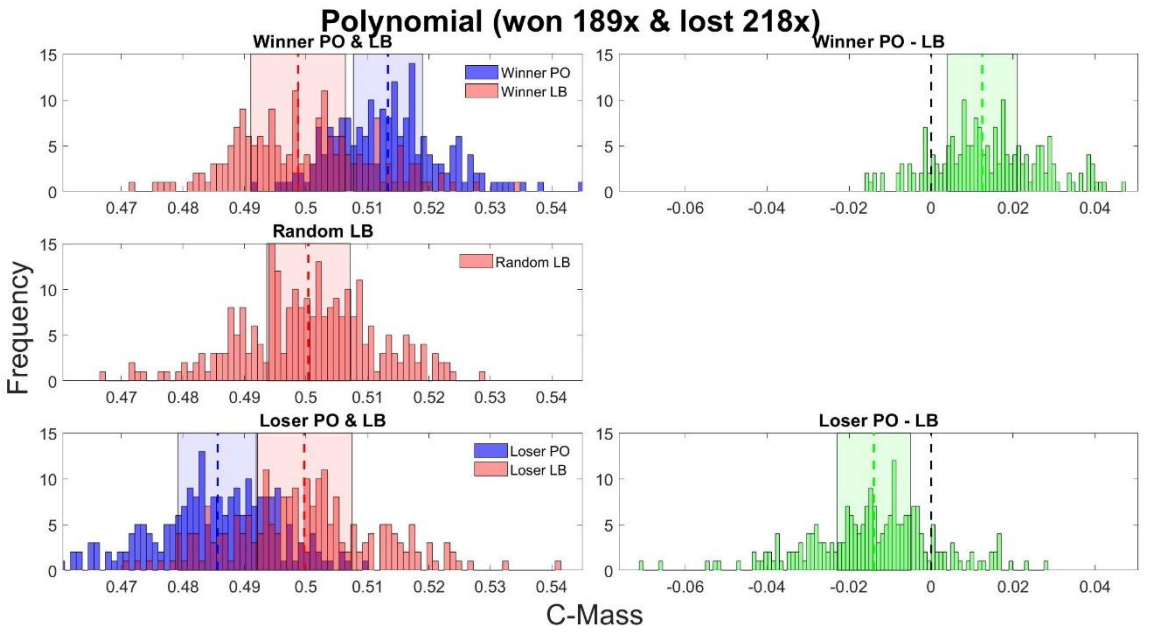


Figure 114. C-Mass distributions for the polynomial (order of 2) SVM classifier. Plotting conventions are identical to those of Figure 94.

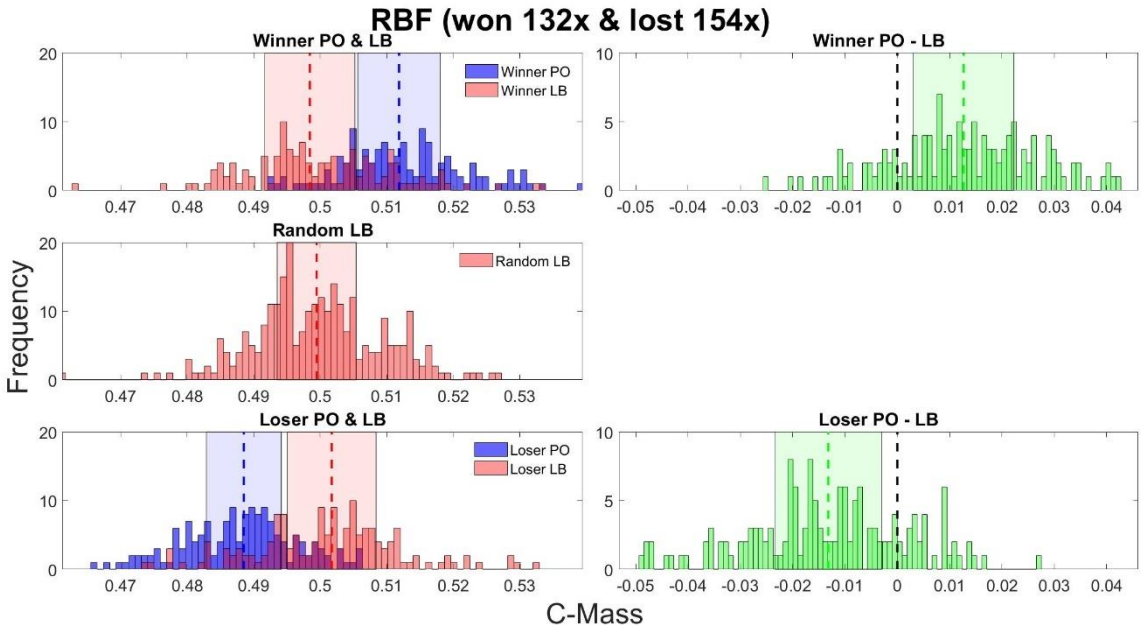


Figure 115. C-Mass distributions for the radial-basis function (RBF) SVM classifier. Plotting conventions are identical to those of Figure 22.

Appendix D – 2x2 Jackknife Bootstrap Interaction

In addition to the approach presented in Chapter 6, we conducted the 2x2 interaction analysis of N2pc temporal jitters (TJs), adopting the factors experiment (1 vs. 2) and response condition (correct vs. intrusion), with a bootstrap approach. This approach only differed in how N2pc TJ was measured. In our main approach, we measured N2pc TJ after computing dynamic time warping (DTW) between a jackknife n-1 Grand Average N2pc and each jackknife sample's subject-level N2pc. The variance of this latency distribution was then defined as the excluded subject's N2pc TJ value.

In this bootstrap approach, we again jackknifed first. We then took a bootstrap sample of N2pcs of a given n-1 jackknife sample, keeping the same number of subjects as the jackknife sample contained (e.g., n-1 being 23-1, so 22, for Experiment 1). We computed the average of the bootstrap subject-level N2pcs next, which generated a bootstrap n-1 GA N2pc. Subsequently, the DTW between the jackknife n-1 GA N2pc and the bootstrap n-1 GA N2pc was computed and the DTW distance value was stored in a distribution. This procedure was repeated 100000 and the variance of these 100000 DTW distance values was defined as our measure of N2pc TJ. Similar to our main analysis, this whole procedure was repeated *for each subject in each condition* (i.e., in each bin of our 2x2 interaction), again yielding a N2pc TJ value for each bin of the 2x2 on which the repeated measures ANOVA was performed as explained in Chapter 6. Note that we corrected F-values according to Ulrich and Miller (2001) in this analysis, too, as it again involved jackknifing.

The results are presented in Figure 116, which show the equivalents of the plots presented in Chapter 6's Figure 54 for the current bootstrap approach. Both N2pc time-windows were again found to not contain statistically significant interaction effects (onset time-window (150 – 200 ms): $p = .4281$ & full N2pc (150 – 400 ms): $p = .2003$).

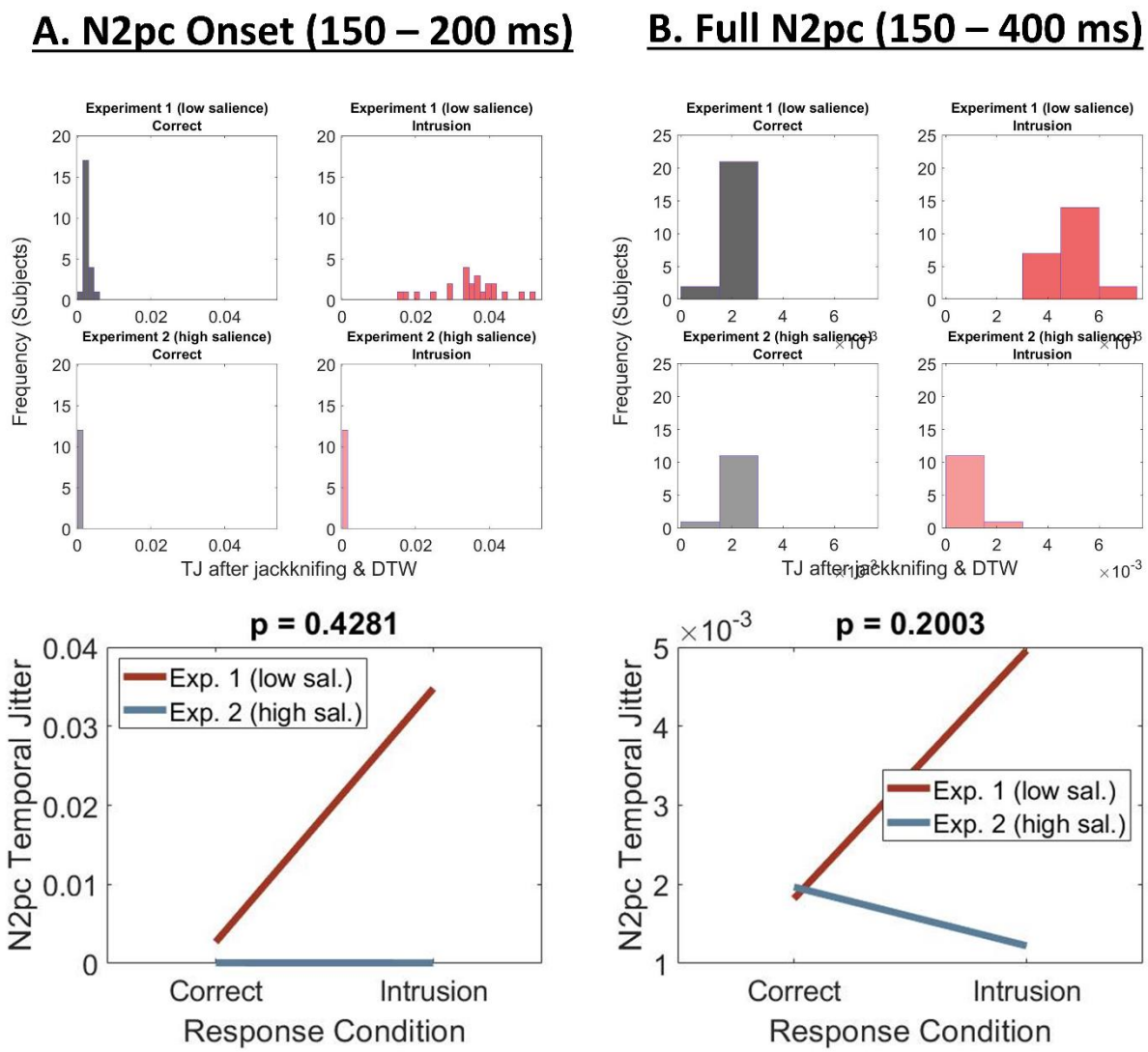


Figure 116. 2x2 Jackknife Bootstrap Interaction. Panels A and B show the results of performing the jackknife bootstrap interaction analysis for the time-windows of N2pc onset (150 – 200 ms) and the full N2pc (150 – 400 ms), respectively. The top eight panels display the subject-level distributions of N2pc TJ after jackknifing, bootstrapping & DTW on which the repeated measures ANOVAs were conducted. The bottom two panels illustrate the interaction plots, depicting the average N2pc TJ values of each bin of the 2x2 interaction. Note that N2pc TJ value-ranges (as in, e.g., the y-axis of the interaction plots) is lower for the full N2pc (B) than for the N2pc's onset (A).

References

Akyürek, E. G., Eshuis, S. A. H., Nieuwenstein, M. R., Saija, J. D., Başkent, D., & Hommel, B. (2012). Temporal target integration underlies performance at Lag 1 in the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1448–1464. https://doi.org/10.1037/a0027610

Akyürek, E. G., Kappelmann, N., Volkert, M., & van Rijn, H. (2017). What You See Is What You Remember: Visual Chunking by Temporal Integration Enhances Working Memory. *Journal of Cognitive Neuroscience*, *29*(12), 2025–2036. https://doi.org/10.1162/jocn_a_01175

Akyürek, E. G., Leszczyński, M., & Schubö, A. (2010). The temporal locus of the interaction between working memory consolidation and the attentional blink. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.2010.01033.x

Akyürek, E. G., & Wolff, M. J. (2016). Extended temporal integration in rapid serial visual presentation: Attentional control at Lag 1 and beyond. *Acta Psychologica*, *168*, 50–64. https://doi.org/10.1016/j.actpsy.2016.04.009

Allefeld, C., Görgen, K., & Haynes, J. D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage*, *141*, 378–392. https://doi.org/10.1016/j.neuroimage.2016.07.040

Anderson, M. J., & Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, *43*(1), 75–88.

Arnell, K. M., & Jenkins, R. (2004). Revisiting within-modality and cross-modality attentional blinks: Effects of target-distractor similarity. *Perception and Psychophysics*, *66*(7), 1147–1161. https://doi.org/10.3758/BF03196842

Arnell, K. M., & Jolicœur, P. (1999). The attentional blink across stimulus modalities: Evidence for central processing limitations. *Journal of Experimental Psychology:*

Human Perception and Performance, *25*(3), 630–648. https://doi.org/10.1037/0096-1523.25.3.630

Arnell, K. M., & Larson, J. M. (2002). Cross-modality attentional blinks without preparatory task-set switching. *Psychonomic Bulletin and Review*, *9*(3), 497–506. https://doi.org/10.3758/BF03196305

Baars, B. J. (1988). *A cognitive theory of consciousness.* Cambridge University Press.

Baars, B. J. (1997). *In the Theater of Consciousness*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195102659.001.1

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences*, *6*(1), 47–52. https://doi.org/10.1016/S1364-6613(00)01819-2

Bao, C., Fountas, Z., Olugbade, T., & Bianchi-Berthouze, N. (2020). Multimodal Data Fusion Based on the Global Workspace Theory. *Proceedings of the 2020 International Conference on Multimodal Interaction*, 414–422. https://doi.org/10.1145/3382507.3418849

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Botella, J. (1992). Target-specified and target-categorized conditions in RSVP tasks as reflected by detection time. *Bulletin of the Psychonomic Society*, *30*(3), 197–200. https://doi.org/10.3758/BF03330440

Botella, J., Barriopedro, M. I., & Suero, M. (2001). A model of the formation of illusory conjunctions in the time domain. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(6), 1452–1467. https://doi.org/10.1037/0096-1523.27.6.1452

Botella, J., & Eriksen, C. W. (1992). Filtering versus parallel processing in RSVP tasks. *Perception & Psychophysics*. https://doi.org/10.3758/BF03211627

Botella, J., Garcia, M. L., & Barriopedro, M. (1992). Intrusion patterns in rapid serial visual presentation tasks with two response dimensions. *Perception & Psychophysics*, *52*(5), 547–552. https://doi.org/10.3758/BF03206716

Botella, J., Privado, J., de Liaño, B. G. G., & Suero, M. (2011). Illusory conjunctions reflect the time course of the attentional blink. *Attention, Perception, and Psychophysics*, *73*(5), 1361–1373. https://doi.org/10.3758/s13414-011-0112-z

Botella, J., Rodríguez, C., Rubio, M. E., Valle-Inclán, F., & De Liaño, B. G. G. (2008). Event-related potentials and illusory conjunctions in the time domain. *Psicologica*, *29*(2), 153–169.

Bowman, H., Brooks, J. L., Hajilou, O., Zoumpoulaki, A., & Litvak, V. (2020). Breaking the circularity in circular analyses: Simulations and formal treatment of the flattened average approach. In *PLoS Computational Biology* (Vol. 16, Issue 11). https://doi.org/10.1371/journal.pcbi.1008286

Bowman, H., & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review*, *114*(1), 38–70. https://doi.org/10.1037/0033-295X.114.1.38

Bowman, H., Wyble, B., Chennu, S., & Craston, P. (2008). A reciprocal relationship between bottom-up trace strength and the attentional blink bottleneck: Relating the LC-NE and ST2 models. *Brain Research*, *1202*, 25–42. https://doi.org/10.1016/j.brainres.2007.06.035

Bridge, H. (2011). Mapping the visual brain: How and why. *Eye*, *25*(3), 291–296. https://doi.org/10.1038/eye.2010.166

Broadbent, D. E., & Broadbent, M. H. (1987). From detection to identification: Response to multiple targets in rapid serial visual presentation. *Perception & Psychophysics*, *42*(2), 105–113. https://doi.org/10.3758/BF03210498

Brooks, J. L., Zoumpoulaki, A., & Bowman, H. (2017). Data-driven region-of-interest selection without inflating Type I error rate. *Psychophysiology*, *54*(1), 100–113. https://doi.org/10.1111/psyp.12682

Buzkova, P. (2016). Interaction Testing: Residuals-Based Permutations and Parametric Bootstrap in Continuous, Count, and Binary Data. *Epidemiologic Methods*, *5*(1). https://doi.org/10.1515/em-2015-0010

Callahan-Flintoft, C., Chen, H., & Wyble, B. (2018). A hierarchical model of visual processing simulates neural mechanisms underlying reflexive attention. *Journal of Experimental Psychology: General*, *147*(9), 1273–1294. https://doi.org/10.1037/xge0000484.supp

Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., & Turret, J. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, *11*(10), 9–9. https://doi.org/10.1167/11.10.9

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1–1. https://doi.org/10.1167/13.10.1

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chartier, S., Cousineau, D., & Charbonneau, D. (2004). A Connectionist Model of the Attentional Blink Effect During a Rapid Serial Visual Presentation Task. *Proceedings of the 6th International Conference on Cognitive Modelling*, 64–69.

Chennu, S. (2009). *The temporal spotlight of attention: Computatioanl and electrophysiological explorations.* [University of Kent]. http://www.cs.kent.ac.uk/pubs/2009/3054

Chennu, S., Bowman, H., & Wyble, B. (2011). Fortunate Conjunctions Revived: Feature Binding with the 2f-ST2 Model. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society, 1992*, 182–196.

Chennu, S., Craston, P., Wyble, B., & Bowman, H. (2009). Attention Increases the Temporal Precision of Conscious Perception: Verifying the Neural-ST2 Model. *PLOS Computational Biology*, *5*(11), e1000576. https://doi.org/10.1371/journal.pcbi.1000576

Chun, M. M. (1997). Temporal binding errors are redistributed by the attentional blink. *Perception and Psychophysics*, *59*(8), 1191–1199. https://doi.org/10.3758/BF03214207

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(1), 109–127. https://doi.org/10.1037/0096-1523.21.1.109

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cost, R. S., & Salzberg, S. L. (2004). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning*, *10*, 57–78.

Craston, P., Wyble, B., Chennu, S., & Bowman, H. (2008). The Attentional Blink Reveals Serial Working Memory Encoding: Evidence from Virtual and Human Event-related Potentials. *Journal of Cognitive Neuroscience*, *21*(3), 550–566. https://doi.org/10.1162/jocn.2009.21036

De Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. *Breakthroughs in Statistics*, *1*, 134x174. https://doi.org/10.1007/978-1-4612-0919-5_10

Dehaene, S., & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. USA*, 6.

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences*, *100*(14), 8520–8525. https://doi.org/10.1073/pnas.1332574100

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Di Lollo, V., Kawahara, J. I., Shahab Ghorashi, S. M., & Enns, J. T. (2005). The attentional blink: Resource depletion or temporary loss of control? *Psychological Research*, *69*(3), 191–200. https://doi.org/10.1007/s00426-004-0173-x

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508.

Duncan, J., & Humphreys, G. W. (1989). Visual Search and Stimulus Similarity. *Psychological Review*. https://doi.org/10.1037/0033-295X.96.3.433

Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attentional dwell time in human vision. *Nature*, *369*(6478), 313–315. https://doi.org/10.1038/369313a0

Dux, P. E., & Marois, R. (2009). The attentional blink: A review of data and theory. *Atten Percept Psychophys*, *71*(8), 1683–1700. https://doi.org/10.3758/APP.71.8.1683

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. *Electroencephalography and Clinical Neurophysiology*. https://doi.org/10.1016/0013-4694(96)95711-9

Fahrenfort, J. J., van Driel, J., van Gaal, S., & Olivers, C. N. L. (2018). From ERPs to MVPA Using the Amsterdam Decoding and Modeling Toolbox (ADAM). In *Frontiers in Neuroscience* (Vol. 12). https://www.frontiersin.org/article/10.3389/fnins.2018.00368

Flandin, G., & Friston, K. J. (2015). Topological Inference. In *Brain Mapping: An Encyclopedic Reference* (Vol. 1, pp. 495–500). https://doi.org/10.1016/B978-0-12-397025-1.00322-5

Ford, J. M., Mohs, R. C., Pfefferbaum, A., & Kopell, B. S. (1980). On the Utility of P3 Latency and RT for Studying Cognitive Processes. In H. H. Kornhubek & L. Deecke (Eds.), *Progress in Brain Research* (Vol. 54, pp. 661–667). Elsevier. https://doi.org/10.1016/S0079-6123(08)61687-8

Fragopanagos, N., Kockelkoren, S., & Taylor, J. G. (2005). A neurodynamic model of the attentional blink. *Cognitive Brain Research*, *24*(3), 568–586. https://doi.org/10.1016/j.cogbrainres.2005.03.010

Franklin, S., & Graesser, A. C. (1999). A Software Agent Model of Consciousness. *Consciousness and Cognition*, *8*, 285–301.

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. In *NeuroImage* (Vol. 61, Issue 4, pp. 1300–1310). Academic Press. https://doi.org/10.1016/j.neuroimage.2012.04.018

Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. J. (1991). Comparing functional (PET) images: The assessment of significant change. *Journal of Cerebral Blood Flow and Metabolism*, *11*(4), 690–699. https://doi.org/10.1038/jcbfm.1991.122

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. https://doi.org/10.1002/hbm.460020402

Gathercole, S. E., & Broadbent, D. E. (1984). Combining attributes in specified and categorized target search: Further evidence for strategic differences. *Memory & Cognition*. https://doi.org/10.3758/BF03198292

Goodbourn, P. T., & Holcombe, A. O. (2015). 'Pseudoextinction': Asymmetries in simultaneous attentional selection. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/a0038734

Gordon-Salant, S. (2005). Hearing loss and aging: New research findings and clinical implications. *Journal of Rehabilitation Research & Development*, *42*.

Handel, S. (2006). *Perceptual coherence: Hearing and seeing*. Oxford University Press.

Handy, T. C. (2005). Basic principles of ERP quantification. In *Event-related potentials: A methods handbook*.

Harris, K., Miller, C., Jose, B., Beech, A., Woodhams, J., & Bowman, H. (2020). Breakthrough percepts of online identity: Detecting recognition of email addresses on the fringe of awareness. *European Journal of Neuroscience*, *n/a*(n/a). https://doi.org/10.1111/ejn.15098

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hogeboom, M., & van Leeuwen, C. (1997). Visual search strategy and perceptual organization covary with individual preference and structural complexity. *Acta Psychologica*, *95*(2), 141–164. https://doi.org/10.1016/S0001-6918(96)00049-2

Hommel, B., & Akyürek, E. G. (2005). Lag-1 sparing in the attentional blink: Benefits and costs of integrating two events into a single episode. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *58*(8), 1415–1433. https://doi.org/10.1080/02724980443000647

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience and Biobehavioral Reviews*, *119*(April), 456–467. https://doi.org/10.1016/j.neubiorev.2020.09.036

Ikkai, A., McCollough, A. W., & Vogel, E. K. (2010). Contralateral Delay Activity Provides a Neural Measure of the Number of Representations in Visual Working Memory. *Journal of Neurophysiology*, *103*(4), 1963–1968. https://doi.org/10.1152/jn.00978.2009

Intraub, H. (1985). Visual Dissociation. An Illusory Conjunction of Pictures and Forms. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/0096-1523.11.4.431

Isaak, M. I., Shapiro, K. L., & Martin, J. (1999). The attentional blink reflects retrieval competition among multiple rapid serial visual presentation items: Tests of an interference model. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1774–1792. https://doi.org/10.1037/0096-1523.25.6.1774

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Jolicœur, P., & Dell'Acqua, R. (1999). Attentional and structural constraints on visual encoding. *Psychological Research*, *62*, 154–164. https://doi.org/10.1007/s004260050048

Jones, W., Pincham, H., Gootjes-Dreesbach, E. L., & Bowman, H. (2020). Fleeting Perceptual Experience and the Possibility of Recalling Without Seeing. *Scientific Reports*, *10*(1), 1–19. https://doi.org/10.1038/s41598-020-64843-2

Kawahara, J. I., Enns, J. T., & Di Lollo, V. (2006). The attentional blink is not a unitary phenomenon. *Psychological Research, 70*(6), 405–413. https://doi.org/10.1007/s00426-005-0007-5

Keele, S. W., & Neill, W. T. (1978). Mechanisms of attention. In E. C. Carterette & M. P. Friedman (Eds.). In *Handbook of Perception Vol. 9* (pp. 3–47). New York: Academic Press.

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.2007.00618.x

Kilner, J. M., & Friston, K. J. (2010). Topological inference for EEG and MEG. *Annals of Applied Statistics, 4*(3), 1272–1290. https://doi.org/10.1214/10-AOAS337

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences, 18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

King, J. R., Faugeras, F., Gramfort, A., Schurger, A., El Karoui, I., Sitt, J. D., Rohaut, B., Wacongne, C., Labyt, E., Bekinschtein, T., Cohen, L., Naccache, L., & Dehaene, S. (2013). Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. *NeuroImage, 83*, 726–738. https://doi.org/10.1016/j.neuroimage.2013.07.013

King, J. R., Gramfort, A., Schurger, A., Naccache, L., & Dehaene, S. (2014). Two distinct dynamic modes subtend the detection of unexpected sounds. *PLoS ONE, 9*(1). https://doi.org/10.1371/journal.pone.0085791

Kiss, M., Van Velzen, J., & Eimer, M. (2008). The N2pc component and its links to attention shifts and spatially selective visual processing. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.2007.00611.x

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*. https://doi.org/10.1017/S0048577201990559

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Lorca-Puls, D. L., Gajardo-Vidal, A., White, J., Seghier, M. L., Leff, A. P., Green, D. W., Crinion, J. T., Ludersdorfer, P., Hope, T. M. H., Bowman, H., & Price, C. J. (2018). The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2018.03.014

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, *4*(2). https://doi.org/10.1088/1741-2560/4/2/R01

Luck, S. J. (2014). An Introduction to the Event-Related Potential Technique, second edition. In *The MIT Press*.

Luck, S. J., & Hillyard, S. (1994). Spatial filtering during visual search: Evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1000–1014. https://doi.org/10.1037/0096-1523.20.5.1000

Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature*, *383*(6601), 616–618. https://doi.org/10.1038/383616a0

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. In *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2016.01.003

Maki, W. S., Frigen, K., & Paulson, K. (1997). Associative priming by targets and distractors during rapid serial visual presentation: Does word meaning survive the attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *23*(4), 1014–1034. https://doi.org/10.1037/0096-1523.23.4.1014

Manly, B. F. (2018). *Randomization, Bootstrap and Monte Carlo Methods in Biology: Texts in Statistical Science*. chapman and hall/CRC.

Marois, R., Yi, D. J., & Chun, M. M. (2004). The Neural Fate of Consciously Perceived and Missed Events in the Attentional Blink. *Neuron*, *41*(3), 465–472. https://doi.org/10.1016/S0896-6273(04)00012-1

Martens, S., Wolters, G., & Van Raamsdonk, M. (2002). Blinks of the mind: Memory effects of attentional processes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(6), 1275–1287. https://doi.org/10.1037//0096-1523.28.6.1275

Martens, S., & Wyble, B. (2010). The attentional blink: Past, present, and future of a blind spot in perceptual awareness. *Neuroscience and Biobehavioral Reviews*, *34*(6), 947–957. https://doi.org/10.1016/j.neubiorev.2009.12.005

Marti, S., & Dehaene, S. (2017). Discrete and continuous mechanisms of temporal selection in rapid visual streams. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-02079-x

Marti, S., King, J. R., & Dehaene, S. (2015). Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. *Neuron*, *88*(6), 1297–1307. https://doi.org/10.1016/j.neuron.2015.10.040

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584 PART B), 2145–2166. https://doi.org/10.1256/003590002320603584

Mayeli, A., Zotev, V., Refai, H., & Bodurka, J. (2016). Real-time EEG artifact correction during fMRI using ICA. *Journal of Neuroscience Methods*, *274*, 27–37. https://doi.org/10.1016/j.jneumeth.2016.09.012

Mcclelland, J. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287–330. https://doi.org/10.1037/0033-295X.86.4.287

Medina, J. M., Wong, W., Díaz, J. A., & Colonius, H. (2015). Advances in modern mental chronometry. *Frontiers in Human Neuroscience*, *9*.

Mennes, M., Wouters, H., Vanrumste, B., Lagae, L., & Stiers, P. (2010). Validation of ICA as a tool to remove eye movement artifacts from EEG/ERP. *Psychophysiology*, *47*(6), 1142–1150. https://doi.org/10.1111/j.1469-8986.2010.01015.x

Michel, C. M., & He, B. (2019). EEG source localization. *Handbook of Clinical Neurology*, *160*, 85–101. https://doi.org/10.1016/B978-0-444-64032-1.00006-0

Nicholls, M. E. R., Churches, O., & Loetscher, T. (2018). Perception of an ambiguous figure is affected by own-age social biases. *Scientific Reports*, *8*(1), 12661. https://doi.org/10.1038/s41598-018-31129-7

Nichols, T., & Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*.

Nieuwenhuis, S., Holmes, B. D., Gilzenrat, M. S., & Cohen, J. D. (2005). The role of the locus coeruleus in mediating the attentional blink: A neurocomputational theory. *Journal of Experimental Psychology: General*, *134*(3), 291–307. https://doi.org/10.1037/0096-3445.134.3.291

Nieuwenhuis, S., van Nieuwpoort, I. C., Veltman, D. J., & Drent, M. L. (2007). Effects of the noradrenergic agonist clonidine on temporal and spatial attention. *Psychopharmacology*, *193*(2), 261–269. https://doi.org/10.1007/s00213-007-0770-7

Nieuwenstein, M. R. (2006). Top-down controlled, delayed selection in the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 973–985. https://doi.org/10.1037/0096-1523.32.4.973

Nieuwenstein, M. R., Potter, M. C., & Theeuwes, J. (2009). Unmasking the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 159.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nunez, P. L., & Srinivasan, R. (2007). Electroencephalogram. *Scholarpedia*, *2*, 1348.

Olav Kallenberg. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer. https://doi.org/10.1007/0-387-28861-9

Olivers, C. N. L., & Meeter, M. (2008). A Boost and Bounce Theory of Temporal Attention. *Psychological Review*, *115*(4), 836–863. https://doi.org/10.1037/a0013395

Olivers, C. N. L., Van Der Stigchel, S., & Hulleman, J. (2007). Spreading the sparing: Against a limited-capacity account of the attentional blink. *Psychological Research*, *71*(2), 126–139. https://doi.org/10.1007/s00426-005-0029-z

Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85–93. https://doi.org/10.1016/j.jneumeth.2014.08.003

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

Pincham, H. L., Bowman, H., & Szucs, D. (2016). The experiential blink: Mapping the cost of working memory encoding onto conscious perception in the attentional blink. *Cortex*, *81*, 35–49. http://dx.doi.org/10.1016/j.cortex.2016.04.007

Plous, S. (1993). *The psychology of judgment and decision making.* (pp. xvi, 302). Mcgraw-Hill Book Company.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. In *Clinical Neurophysiology* (Vol. 118, Issue 10, pp. 2128–2148). https://doi.org/10.1016/j.clinph.2007.04.019

Posner, M. I. (2005). Timing the Brain: Mental Chronometry as a Tool in Neuroscience. *PLOS Biology*, *3*(2), e51. https://doi.org/10.1371/journal.pbio.0030051

Popple, A. V., & Levi, D. M. (2007). Attentional blinks as errors in temporal binding. *Vision Research*. https://doi.org/10.1016/j.visres.2007.06.022

Potter, M. C. (1993). Very short-term conceptual memory. *Memory & Cognition*, *21*(2), 156–161. https://doi.org/10.3758/BF03202727

Potter, M. C., Dell'Acqua, R., Pesciarelli, F., Job, R., Peressotti, F., & O'Connor, D. H. (2005). Bidirectional semantic priming in the attentional blink. In *Psychonomic Bulletin and Review* (Vol. 12, Issue 3, pp. 460–465). https://doi.org/10.3758/BF03193788

Potter, M. C., & Fox, L. F. (2009). Detecting and remembering simultaneous pictures in a rapid serial visual presentation. In *Journal of Experimental Psychology: Human Perception and Performance* (Vol. 35, Issue 1, pp. 28–38). American Psychological Association. https://doi.org/10.1037/a0013624

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15. https://doi.org/10.1037/h0027470

Potter, M. C., Staub, A., & O'Connor, D. H. (2002). The Time Course of Competition for Attention: Attention Is Initially Labile. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(5), 1149–1162. https://doi.org/10.1037//0096-1523.28.5.1149

Potter, M. C., Wyble, B., Pandav, R., & Olejarczyk, J. (2010). Picture Detection in Rapid Serial Visual Presentation: Features or Identity? *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/a0018730

Prajapati, G. L., & Patle, A. (2010). On Performing Classification Using SVM with Radial Basis and Polynomial Kernel Functions. *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, 512–515. https://doi.org/10.1109/ICETET.2010.134

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary Suppression of Visual Processing in an RSVP Task: An Attentional Blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849–860. https://doi.org/10.1037/0096-1523.18.3.849

Recht, S., Mamassian, P., & de Gardelle, V. (2019). Temporal attention causes systematic biases in visual confidence. *Scientific Reports*. https://doi.org/10.1038/s41598-019-48063-x

Rolke, B., Heil, M., Streb, J., & Hennighausen, E. (2001). Missed prime words within the attentional blink evoke an N400 semantic priming effect. *Psychophysiology*, *38*(2), 165–174. https://doi.org/10.1017/S0048577201991504

Safieddine, D., Kachenoura, A., Albera, L., Birot, G., Karfoul, A., Pasnicu, A., Biraben, A., Wendling, F., Senhadji, L., & Merlet, I. (2012). Removal of muscle artifact from EEG data: Comparison between stochastic (ICA and CCA) and deterministic (EMD and wavelet-based) approaches. *EURASIP Journal on Advances in Signal Processing*, *2012*(1), 127. https://doi.org/10.1186/1687-6180-2012-127

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391–1400. https://doi.org/10.1038/nn1549

Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition*, *15*(2), 433–449. https://doi.org/10.1016/j.concog.2005.11.005

Shapiro, K., Driver, J., Ward, R., & Sorensen, R. E. (1997). Priming from the attentional blink: A Failure to Extract Visual Tokens but Not Visual Types. *Psychological Science*, *8*(2), 95–100. https://doi.org/10.1111/j.1467-9280.1997.tb00689.x

Shapiro, K. L., Raymond, J. E., & Arnell, K. M. (1994). Attention to visual pattern information produces the attentional blink in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 357–371. https://doi.org/10.1037/0096-1523.20.2.357

Shih, S. I. (2008). The attention cascade model and attentional blink. *Cognitive Psychology*, *56*(3), 210–236. https://doi.org/10.1016/j.cogpsych.2007.06.001

Simione, L., Akyürek, E. G., Vastola, V., Raffone, A., & Bowman, H. (2017). Illusions of integration are subjectively impenetrable: Phenomenological experience of Lag 1 percepts during dual-target RSVP. *Consciousness and Cognition*, *51*, 181–192. https://doi.org/10.1016/j.concog.2017.03.004

Simione, L., Raffone, A., Wolters, G., Salmas, P., Nakatani, C., Belardinelli, M. O., & van Leeuwen, C. (2012). ViSA: A Neurodynamic Model for Visuo-Spatial Working Memory, Attentional Blink, and Conscious Access. *Psychological Review*, *119*(4), 745–769. https://doi.org/10.1037/a0029345

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, Ser. B*, *36*, 111–147.

Taatgen, N. A., Juvina, I., Schipper, M., Borst, J. P., & Martens, S. (2009). Too much control can hurt: A threaded cognition model of the attentional blink. *Cognitive Psychology*, *59*(1), 1–29. http://dx.doi.org/10.1016/j.cogpsych.2008.12.002

Treder, M. (2020). MVPA-Light: A Classification and Regression Toolbox for Multi-Dimensional Data. *Frontiers in Neuroscience*, *14*.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*. https://doi.org/10.1016/0010-0285(80)90005-5

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology*, *38*(5), 816–827. https://doi.org/10.1017/S0048577201000610

Van Leeuwen, C., & Lachmann, T. (2004). Negative and positive congruence effects in letters and shapes. *Perception & Psychophysics*, *66*(6), 908–925. https://doi.org/10.3758/BF03194984

Verleger, R., Jaśkowski, P., & Wascher, E. (2005). Evidence for an integrative role of P3b in linking reaction to perception. *Journal of Psychophysiology*. https://doi.org/10.1027/0269-8803.19.3.165

Visser, T. A. W., Merikle, P. M., & Di Lollo, V. (2005). Priming in the attentional blink: Perception without awareness? *Visual Cognition*, *12*(7), 1362–1372. https://doi.org/10.1080/13506280444000733

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(6), 1656–1674. https://doi.org/10.1037/0096-1523.24.6.1656

Von Ehrenfels, C. (1937). On Gestalt-qualities. *Psychological Review*, *44*(6), 521–524. https://doi.org/10.1037/h0056968

Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as Inference: Selection Is Probabilistic; Responses Are All-or-None Samples. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/a0017352

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, *138*(6), 1218–1252. PubMed. https://doi.org/10.1037/a0029334

Wolfe, J. M. (2014). Approaches to Visual Search. In *The Oxford Handbook of Attention.* https://doi.org/10.1093/oxfordhb/9780199675111.013.002

Woodman, G. F., & Luck, S. J. (1999). Electrophysiological measurement of rapid shifts of attention during visual search. *Nature*. https://doi.org/10.1038/23698

Wyble, B., Bowman, H., & Nieuwenstein, M. (2009). The Attentional Blink Provides Episodic Distinctiveness: Sparing at a Cost. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 787–807. https://doi.org/10.1037/a0013902

Wyble, B., Callahan-Flintoft, C., Chen, H., Marinov, T., Sarkar, A., & Bowman, H. (2020). Understanding Visual Attention With RAGNAROC: A Reflexive Attention Gradient Through Neural AttRactOr Competition. *Psychological Review*. https://doi.org/10.1037/rev0000245

Wyble, B., Potter, M. C., Bowman, H., & Nieuwenstein, M. (2011). Attentional episodes in visual perception. *Journal of Experimental Psychology: General*, *140*(3), 488–505. https://doi.org/10.1037/a0023612

Zivony, A., & Eimer, M. (2020a). Distractor Intrusions Are the Result of Delayed Attentional Engagement: A New Temporal Variability Account of Attentional Selectivity in Dynamic Visual Tasks. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000789

Zivony, A., & Eimer, M. (2020b). Perceptual competition between targets and distractors determines working memory access and produces intrusion errors in rapid serial visual presentation (RSVP) tasks. In *Journal of Experimental Psychology: Human Perception and Performance* (Vol. 46, Issue 12, pp. 1490–1510). American Psychological Association. https://doi.org/10.1037/xhp0000871

Zivony, A., & Lamy, D. (2018). Contingent Attentional Engagement: Stimulus- and Goal-Driven Capture Have Qualitatively Different Consequences. *Psychological Science*. https://doi.org/10.1177/0956797618799302

Zoumpoulaki, A., Alsufyani, A., Filetti, M., Brammer, M., & Bowman, H. (2015). Latency as a region contrast: Measuring ERP latency differences with Dynamic Time Warping. *Psychophysiology*, *52*(12), 1559–1576. https://doi.org/10.1111/psyp.12521