How we pay attention in naturalistic visual search settings

Nora Turoman[1,2,3], Ruxandra I. Tivadar[1,4,5], Chrysa Retsa[1,7], Micah M. Murray[1,4,6,7], Pawel J. Matusz[1,2,6 *]

[1] CIBM Center for Biomedical Imaging, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[2] Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland

[3] Working Memory, Cognition and Development lab, Department of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

[4] Department of Ophthalmology, Fondation Asile des Aveugles, Lausanne, Switzerland

[5] Bern University

[6] Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN, USA

[7] The LINE (Laboratory for Investigative Neurophysiology), Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

* Corresponding author:

Pawel Matusz

Information Systems Institute

University of Applied Sciences Western Switzerland (HES-SO Valais)

Rue Technopole 3

3960 Sierre, Switzerland

Tel: +41 27 606 9060

Highlights :

- We tested how goals control attention in naturalistic, multisensory context-rich settings
- Arbitrary target-colour distractors captured attention more than semantically congruent ones
- Nonlinear responses supported interactions between goals, salience and context
- Gain and network-based mechanisms altered stimulus-elicited responses from early on
- Goal-based attention used both predictions and meaning for distractor inhibition

Abstract

Research on attentional control has largely focused on single senses and the importance of one's behavioural goals in controlling attention. However, everyday situations are multisensory and contain regularities, both likely influencing attention. We investigated how visual attentional capture is simultaneously impacted by top-down goals, multisensory nature of stimuli, *and* contextual factors of stimulus' semantic relationship and predictability. Participants performed a multisensory version of the Folk et al., (1992) spatial cueing paradigm, searching for a target of a predefined colour (e.g. a red bar) within an array preceded by a distractor. We manipulated: 1) stimulus' goal-relevance via distractor's colour (matching vs. mismatching the target), 2) stimulus' multisensory nature (colour distractors appearing alone vs. with tones), 3) relationship between the distractor sound and colour (arbitrary vs. semantically congruent) and 4) predictability of the distractor onset. Reaction-time spatial cueing served as a behavioural measure of attentional selection. We also recorded 129-channel event-related potentials (ERPs), analysing the distractor-elicited N2pc component both canonically and using a multivariate electrical neuroimaging (EN) framework. Behaviourally, arbitrary target-matching distractors captured attention more strongly than semantically congruent ones, with no evidence for context modulating multisensory enhancements of capture. Notably, EN analyses revealed context-based influences on attention to both visual and multisensory distractors, in how strongly they activated the brain and type of activated brain networks. In both cases, these context-driven brain response modulations occurred long before the N2pc time-window, with network-based modulations at ~30ms, followed by strength-based modulations at ~100ms post-distractor. This points to meaning being a second source, next to predictions, of contextual information facilitating goal-directed behaviour. More broadly, in everyday situations, attentional is controlled by an interplay between one's goals, stimuli's perceptual salience *and* stimuli's meaning and predictability. Our study calls for a revision of attentional control theories to account for the role of contextual and multisensory control.

Keywords: attentional control, multisensory, real-world, semantic congruence, temporal predictability, context

Introduction

Goal-directed behaviour depends on the ability to allocate processing resources towards the stimuli important to current behavioural goals ("attentional control"). On the one hand, our current knowledge about attentional control may be limited to the rigorous, yet artificial, conditions in which it is traditionally studied. On the other hand, findings from studies assessing attentional control with naturalistic stimuli (audiostories, films) may be limited by confounds from other processes present in such settings. Here, we systematically tested how traditionally studied goal- and salience-based attentional control interact with more naturalistic, context-based mechanisms.

In the real world, the location of goal-relevant information is rarely known in advance. Since the pioneering visual search paradigm (Treisman & Gelade, 1980), we know that in multi-stimulus settings target attributes can be utilised to control attention. Here, research provided conflicting results as to whether primacy in controlling attentional selection lies in task-relevance of objects' attributes (Folk et al., 1992) or their bottom-up salience (e.g. Theeuwes, 1991). Folk et al., (1992) used a version of the spatial cueing paradigm and revealed that attentional capture is elicited only by distractors that matched the target colour. Consequently, they proposed the 'task-set contingent attentional capture' (or TAC) hypothesis, i.e., salient objects will capture attention only if they share features with the target and so are potentially task-relevant. However, subsequently mechanisms beyond goal-relevance were shown to serve as additional sources of attentional control, e.g., spatiotemporal and semantic information within the stimulus and the environment where it appears (e.g., Chun & Jiang 1998; Peelen & Kastner, 2014; Summerfield et al., 2006; van Moorselaar & Slagter 2019; Press et al., 2020), but also multisensory processes (Matusz & Eimer, 2011, 2013; Matusz et al., 2015a; Lunn et al., 2019; Soto-Faraco et al., 2019).

Some multisensory processes occur at early latencies (<100ms post-stimulus), generated within primary cortices (e.g., Talsma & Woldroff, 2005; Raij et al., 2010; Cappe et al., 2010; reviewed in de Meo et al., 2015; Murray et al., 2016). This enables them to influence attentional selection in a bottom-up fashion, potentially independently of the observer's goals. This idea was supported by Matusz and Eimer (2011) who used a multisensory adaptation of Folk et al.'s (1992) task. The authors replicated the TAC effect, and also showed that visual distractors captured attention more strongly when

accompanied by a sound, regardless of their goal-relevance. This demonstrated the importance of bottom-up multisensory enhancement (MSE) for attentional selection of visual objects. However, interactions between such goals, multisensory influences on attentional control, and the stimuli's temporal and semantic context[1] remain unknown.

Top-down contextual factors in attentional control

The temporal structure of the environment is routinely used by the brain to build predictions. Attentional control utilises such predictions to improve the selection of target stimuli (e.g., Correa et al., 2005; Coull et al., 2000; Green & McDonald, 2010; Miniussi et al., 1999; Naccache et al., 2002; Rohenkohl et al., 2014) and inhibition of task-irrelevant stimuli (here, location- and feature-based predictions have been more researched than temporal predictions; e.g. reviewed in Noonan et al., 2018; van Moorselaar & Slagter 2020a). In naturalistic, multisensory settings, temporal predictions are predominantly known to improve language comprehension (e.g. Luo & Poeppel, 2007; ten Oever & Sack, 2015), yet their role as a source of attentional control is less known (albeit see, Zion Golumbic et al., 2012, for their role in the "cocktail party" effect). Semantic relationships are another basic principle of organising information in real-world contexts. Compared to semantically incongruent or meaningless (arbitrary) multisensory stimuli, semantically congruent stimuli are more easily identified and remembered (e.g. Laurienti et al. 2006; Matusz et al., 2015a; Tovar et al., 2020; reviewed in ten Oever et al., 2016; Matusz et al., 2020) but also, notably, more strongly attended (Matusz et al., 2015b, 2019a, 2019b; reviewed in Soto-Faraco et al., 2019; Matusz et al., 2019c). For example, Iordanenscu et al., (2009) demonstrated that search for naturalistic objects is faster when accompanied by irrelevant albeit congruent sounds.

What is unclear from existing research is the degree to which goal-based attentional control interacts with salience-driven (multisensory) mechanisms *and* such contextual factors. Researchers have been clarifying such interactions, but typically in a pair-wise fashion, between e.g. attention and semantic memory, or attention and predictions (reviewed in Summerfield & Egner 2009; Nobre & Gazzaley 2012; Press et al., 2020).

[1] Context has been previously defined as the "immediate situation in which the brain operates… shaped by external circumstances, such as properties of sensory events, and internal factors, such as behavioural goal, motor plan, and past experiences" (van Atteveldt et al., 2014).

However, in everyday situations these processes do not interact in an orthogonalised but rather in a multi-dimensional fashion, with multiple sources of control interacting simultaneously (ten Oever et al., 2016; Nastase et al., 2020). Additionally, in the real world, these processes operate on both unisensory and multisensory stimuli, where the latter are more perceptually salient than the former (e.g. Santangelo & Spence 2007; Matusz & Eimer 2011). Thus, one way to create more complete and "naturalistic" theories of attentional control is by investigating how one's goals interact with *multiple* contextual factors in controlling attentional selection – and doing so in *multi-sensory* settings.

The present study

To shed light on how attentional control operates in naturalistic search settings, we investigated interactions between visual top-down goals, bottom-up multisensory salience and distractor's predictability and semantic congruence when all are manipulated simultaneously. We likewise set out to identify brain mechanisms supporting such complex interactions. To address these questions in a rigorous *and* state-of-the-art fashion, we employed a 'naturalistic laboratory' approach that builds on several methodological advances (Matusz et al., 2019c). First, we used a paradigm isolating specific cognitive process, i.e., the Matusz and Eimer's (2011) multisensory adaptation of the Folk et al.'s (1992) task, where we additionally manipulated distractors' temporal predictability and relationship between their auditory and visual features. In the Folk et al.'s task, attentional control is measured via well-understood spatial cueing effects, where larger effects (e.g. for target-colour and AV distractors) reflect stronger attentional capture. Notably, distractor-related responses have the added value as they isolate attentional from later, motor response-related, process. Second, we measured a well-researched brain correlate of attentional object selection, the N2pc event-related potential (ERP) component. The N2pc is a negative-going voltage deflection ~200ms post-stimulus onset at posterior electrode sites contralateral to stimulus location (Luck & Hillyard, 1994a, 1994b; Eimer, 1996; Girelli & Luck, 1997). Studies canonically analysing N2pc have provided strong evidence for task-set contingence of attentional capture (e.g., Kiss et al., 2008; Eimer et al., 2009). Importantly, N2pc is also sensitive to meaning (e.g. Wu et al., 2015) and predictions (e.g. Burra & Kerzel, 2013), whereas its sensitivity to multisensory enhancement is limited (van der Burg et al., 2011, but see below). This joint evidence makes the N2pc a valuable 'starting point' for

investigating interactions between visual goals and more naturalistic sources of control. Third, analysing the traditional EEG markers of attention with advanced frameworks like electrical neuroimaging (EN) (e.g. Lehmann & Skrandies 1980; Murray et al., 2008; Tivadar & Murray 2019) might offer an especially robust, accurate and informative approach.

Briefly, an EN framework encompasses multivariate, reference-independent analyses of the global features of the electric scalp field. Its main added value is that it readily distinguishes *in surface EEG* the neurophysiological mechanisms driving differences in ERPs across experimental conditions: 1) "gain control" mechanisms, modulating the strength of activation within a non-distinguishable brain network, and 2) topographic, network-based mechanisms, modulating the recruited brain sources (scalp EEG topography differences forcibly flow from changes in the underlying sources; Murray et al., 2008). In this, EN complements interpretational limitations of canonical N2pc analyses. Most notably, a difference in mean N2pc voltages can arise from both strength- and network-based mechanisms (albeit it is assumed to signify gain control); it can also emerge from different brain source configurations (for a full discussion, see Matusz et al., 2019b).

We recently used this approach to better understand brain and cognitive mechanisms of attentional control. We revealed that distinct brain networks are active ~N2pc time-window during visual goal-based *and* multisensory bottom-up attention control (across the lifespan; Turoman et al., 2020a, 2020b). However, these reflect spatially-selective, lateralised brain mechanisms, partly captured by the N2pc (via the contra- and ipsilateral comparison). There is little existing evidence to strongly predict how interactions between goals, stimulus salience and context can occur in the brain. Schroeger et al., (2015) proposed that unpredictable events attract attention more strongly (to serve as a signal to reconfigure the predictive model about the world), visible in larger behavioural responses and ERP amplitudes. Both predictions and semantic memory could be utilised to reduce attention to known (i.e., less informative) stimuli. Indeed, goal-based control gauges knowledge to facilitate visual, auditory and multisensory processing (e.g. Summerfield et al., 2008; Iordanescu et al., 2008; Matusz et al., 2016; Retsa et al., 2018, 2020). However, several questions remain. Does knowledge affect the same way attention to task-*irrelevant* stimuli? How early do contextual factors influence stimulus processing here, if both processes are known to do so <150ms post-stimulus (Thorpe et al., 1996; Doehrmann &

Naumer, 2008; Summerfield & Egner 2009). Finally, do contextual processes operate through lateralised or non-lateralised brain mechanisms? Below we specify our hypotheses.

We expected to replicate the TAC effect: In behaviour, visible as large behavioural capture for target- colour matching distractors and no capture for nontarget-colour matching distractors (e.g., Folk et al., 1992; Folk, et al., 2002; Lamy et al., 2004; Lien et al., 2008); in canonical EEG analyses - enhanced N2pc amplitudes for target-colour than nontarget-colour distractors (Eimer et al., 2009). TAC should modulated by both contextual factors: predictability of the distractor onset and the multisensory relationship between distractor features (semantic congruence vs. arbitrary pairing; Wu et al., 2015; Burra & Kerzel, 2013). However, as discussed above, we had no strong predictions how the context factors would modulate TAC (or if they interact while doing so), as these effects have never been tested systematically together, on audio-visual and task-irrelevant stimuli. For MSE, we expected to replicate it behaviourally (Matusz & Eimer 2011), but without strong predictions about concomitant N2pc modulations (c.f. van der Burg et al., 2011). We expected MSE to be modulated by contextual factors, especially multisensory relationship, based on the extensive literature on the role of semantic congruence in multisensory cognition (Doehrmann & Naumer, 2008; ten Oever et al., 2016). Again, we had no strong predictions as to directionality of these modulations or interaction of their influences.

We also investigated if visual goals (TAC), multisensory salience (MSE) and contextual process interactions are supported by lateralised (N2pc-like) or nonlateralised mechanisms. We first analysed if such interactions are captured by canonical N2pc analyses or EN analyses of the lateralised distractor-elicited ERPs ~180-300ms post-stimulus (N2pc-like time-window). These analyses would reveal presence of strength- and network-based *spatially-selective* brain mechanisms contributing to attentional control. However, analyses of the N2pc assume not only lateralised activity, but also symmetry; in brain anatomy but also in scalp electrodes, detecting homologous brain activity over both hemispheres. This may prevent them from detecting other, less-strongly lateralised brain mechanisms of attentional control. We have previously found nonlateralised mechanisms to play a role in attentional control in multisensory settings (Matusz et al., 2019b). Also, semantic information and temporal expectations (and feature-based attention) are known to modulate nonlateralised ERPs (Saenz et al., 2003; Dell'Acqua et al., 2010; Dassanayake et al., 2016). Thus, we tested if strength- and network-based nonlateralised brain mechanisms

reflect interactions between goals, salience and context, analysing the whole post-stimulus time-period activity.

Materials and Methods

Participants

Thirty-nine adult volunteers participated in the study (5 left-handed, 14 male, M_{age}: 27.5years, *SD:* 4years, range: 22–38years). We conducted post-hoc power analyses for the two effects that have been previously behaviourally studied with the present paradigm, namely TAC and MSE. Based on the effect sizes in the original Matusz and Eimer (2011, Exp.2), the analyses revealed sufficient statistical power for both behavioural effects with the collected sample. For ERP analyses, we could calculate power analyses only for the TAC effect. Based on a purely visual study (Eimer et al., 2009) we revealed there to be sufficient statistical power to detect TAC in the N2pc in the current study (all power calculations are available in the SOM's). Participants had normal or corrected-to-normal vision and normal hearing and reported no prior or current neurological or psychiatric disorders. Participants provided informed consent before the start of the experiment. All research procedures were approved by the Cantonal Commission for the Ethics of Human Research (CER-VD; no. 2018-00241).

Task properties and procedures

Each participant took part in one testing session consisting of four experimental tasks (henceforth referred to as 'experiments', Figure 1A), where the first two experiments were followed by a training task (henceforth referred to as 'training', Figure 1C). The first two tasks involved non-semantically related colour-pitch combinations as in the original study, while the last two involved colour-pitch combinations that were semantically congruent (the factor of Multisensory Relationship in Figure 1B). Such a semantic relationship between colours and sounds was created using a training task that was based on the association task from a study by Sui, He and Humphreys (2012). Further, Experiments 1 and 3 involved distractor onsets variable in duration, while Experiments 2 and 4 involved distractors that had a constant onset (the factor of Distractor Onset in Figure 1B). As a pilot study revealed

sufficient proficiency at the experimental task (over 50% accuracy) after a few trials, participants did not practice the task before its administration.

Experimental and training tasks were conducted in a dimly lit, sound-attenuated room, with participants seated at a distance of 90 cm from a 23" LCD monitor with a resolution of 1080 × 1024 (60-Hz refresh rate, HP EliteDisplay E232). All visual elements were approximately equiluminant (~20cd/m^2), as determined by a luxmeter (model SC160, CESVA Instruments) placed at a position adjacent to participants' eyes, measuring the luminance of the screen filled with each respective element's colour. The averages of three measurement values per colour were averaged across colours and transformed from lux to cd/m^2 in order to facilitate comparison with the results of Matusz & Eimer (2011). The testing session lasted no longer than 3h in total, including an initial explanation and obtaining consent, EEG setup, experiments and training, and breaks.

Experiments. Across the experimental tasks, participants searched for a colour predefined target (e.g., a red bar) in a search array, and assessed the target's orientation (vertical vs. horizontal). The search array was always preceded by an array containing distractors. Distractors were visual and audiovisual stimuli that could match the target colour (red set of dots) or not match the target colour (blue set of dots), as in the original Matusz and Eimer (2011) study (Exp.2).

Like in the former study, each experimental trial consisted of the following sequence of arrays: base array (In Experiments 1 and 3: randomly varied between 100, 250, and 450ms; in Experiments 2 and 4: 450ms), followed by distractor array (50ms), followed by a fixation point (150ms), and finally a target array (50ms, see Figure 1A). However, compared to the original Matusz and Eimer (2011) paradigm, the number of elements was reduced from 6 to 4 and targets were reshaped to look like diamonds rather than rectangles, as Experiment 1 served as an adult control for a different, developmental study (reported in Turoman et al., 2020a, 2020b). Thus, the base array contained four differently coloured sets of closely aligned dots, each dot subtending 0.1° × 0.1° of visual angle. Elements were spread equidistally along the circumference of an imaginary circle against a black background, at an angular distance of 2.1° from a central fixation point. Each set element could be one of four possible colours (according to the RGB scale): green (0/179/0), pink (168/51/166), gold (150/134/10), silver (136/136/132). In the distractor array, one of the base array elements changed colour to either a target-matching colour, or a target-

nonmatching colour that was not present in any of the elements before. The remaining three distractor array elements did not change their colour. The distractors and the subsequent target "diamonds" could have either a blue (RGB values: 31/118/220) or red (RGB values: 224/71/52) colour. The target array contained four bars (rectangles) where one was always the colour-defined target. Target colour was counterbalanced across participants. Target orientation (horizontal or vertical) was randomly determined on each trial., The two distractor colours were randomly selected with equal probability before each trial, and the colour change was not spatially predictive of the subsequent target location (same distractor– target location on 25% of trials). On half of all trials, distractor onset coincided with the onset of a pure sine-wave tone, presented from two loudspeakers on the left and right sides of the monitor. Sound intensity was 80 dB SPL (as in Matusz & Eimer, 2011), as measured using an audiometer placed at a position adjacent to participants' ears.

Manipulations of distractors' target colour-matching and the presence/absence of sound resulted in 4 general distractor conditions: TCCV (target colour-cue, Visual), NCCV (nontarget colour-cue, Visual), TCCAV (target colour-cue, AudioVisual), NCCAV (nontarget colour-cue, AudioVisual). These conditions translated into 3 factors that were comparable to Matusz and Eimer's original design: Distractor Colour (target colour-distractor- TCC vs. nontarget colour-distractor- NCC), Distractor Modality (Visual - V vs. AudioVisual - AV) and Cue-Target Location (Same vs. Different). Further manipulations included the introduction of two contextual factors: Distractor Onset and Multisensory Relationship, which were manipulated across the four experimental tasks. For the Distractor Onset factor, in Experiments 2 and 4, base array duration (and therefore distractor onset) was kept constant at 450ms, as in the original Matusz and Eimer (2011) paradigm. Meanwhile in Experiments 1 and 3, base array duration (and therefore distractor onset) was varied between 100, 250 and 450ms. This way, the strength of attentional capture by temporally predictable distractors could be compared with the capture elicited by unpredictable distractors, and if and how these visual and audiovisual distractors differ on that dimension. For the Multisensory Relationship factor, the sound frequency was set to 2000Hz in Experiments 1 and 2, as in the Matusz and Eimer (2011) paradigm, and alternated between 300Hz (low-pitch; chosen based on Matusz & Eimer, 2013) and 4000Hz (high-pitch; chosen for its comparable perceived loudness in relation to the above two sound frequencies per the revised ISO 226:2003 equal-loudness-level contours standard; Spierer et al., 2013) in

Experiments 3 and 4. Then, a Training was presented after Experiment 2, in order to induce in participants a semantic-level association between a specific distractor colour and a specific pitch (Figure 1C). This way, the strength of attentional capture by colour-pitch combinations characterised only by their simultaneous presentation could be compared with the capture elicited by colour-pitch combinations characterised by semantic congruence. Thus, the present study design initially included 5 different factors. These were Distractor Colour, Distractor Modality, and Cue-Target Location, and two new factors: Distractor Onset (DO; Predictable vs. Unpredictable) and Multisensory Relationship (MR; Arbitrary vs. Congruent). However, we simplified our behavioural analyses by using subtracted cueing effects (cue-target location different vs. same) as a foundation for the analyses involving interactions of the other 4 factors (the distractor-evoked ERPs did not capture Cue-Target Location factor).

The full experimental session consisted of 8 blocks of 64 trials each, for each of the 4 experiments, resulting in 2,048 trials in total (512 trials per experiment). Participants were told to respond as quickly and accurately as possible to the targets' orientation by pressing one of two horizontally aligned round buttons (Lib Switch, Liberator Ltd.) that were fixed onto a tray bag on the participants' lap. If participants did not respond within 5000ms of the target presentation, the next trial was initiated, otherwise the next trial was initiated immediately after a button press. Feedback on accuracy was given after each block, followed by a 'progress (treasure) map' which informed participants of the number of blocks remaining until the end, and during which participants could take a break. Breaks were also taken between each experimental task.

Training. The Training procedure consisted of an Association phase and a Testing phase. In the Association phase, participants were shown alternating colour word–pitch pairs. Specifically, each pair consisted of a word, denoting a distractor colour, that was presented on the centre of the screen at the same time as a spatially diffuse pure tone that was either high (4000Hz) or low (300Hz) in pitch. Both the colour word and sound were presented for 2 seconds, after which a central fixation cross was presented for 150ms, followed by the next colour word-pitch pair.

[FIGURE 1 HERE]

Colour words were paired with sounds according to two possible pairing options. In one pairing option, the high-pitch tone was associated with the word 'red' and the low-pitch tone with the word 'blue', and in another pairing option, the high-pitch tone was associated with the word 'blue' and the low-pitch tone with the word 'red' (Figure 1C, Association phase). Pairing options were counterbalanced across participants. Therefore, if the first pairing option was selected, a presentation of the word 'red' with a high-pitch tone would be followed by a presentation of the word 'blue' with a low-pitch tone, which would again be followed by the former pair, etc. There were ten presentations per pair, resulting in a total of 20 trials. Colour words were chosen instead of actual colours to ensure that associations were based on semantic meaning rather than a linking of basic stimulus features (for examples of such taught crossmodal correspondences see e.g., Ernst, 2007). Colour words were shown in participants' native language (speakers: 19 French, 8 Italian, 5 German, 4 Spanish, 3 English). Participants were instructed to observe and try to memorise the pairings as best as they could, as they would be subsequently tested on how well they learnt the pairings.

The strength of colour-pitch associations was assessed in the Testing phase. Here, participants were shown colour word-pitch pairings (as in the training) as well as colour-pitch pairings (a string of x's in either red or blue paired with a sound, Figure 1C, Testing phase). Based on the pairing option that participants were 'taught' in the Association phase, pairings could be either matched or mismatched. For example, if 'red' was paired with a high-pitch tone in the Association phase, in the Testing phase, the word 'red' (or red x's) paired with a high-pitch tone would match, while the word 'red' (or red x's) paired with a low-pitch tone would be mismatched. Participants had to indicate whether a given pair was matched or mismatched by pressing an appropriate button on the same response setup as in the experiments. In a similar paradigm used by Sui, et al., (2012), people were able to reliably associate low-level visual features (colours and geometric shapes) with abstract social concepts such as themselves, their friend, and a stranger. Following their design, in the Testing phase each pairing was shown for 250ms, of which 50ms was the sound (instead of the stimulus duration of 100ms that Sui et al., used, to fit our stimulus parameters), followed by an 800ms blank screen where choices were to be made, and feedback on

performance after each answer was given. Before each trial, a fixation cross was shown for 500ms. Each participant performed three blocks of 80 trials, with 60 trials per possible combination (colour word – sound matching, colour word – sound nonmatching, colour – sound matching, colour – sound nonmatching). A final summary of correct, incorrect, and missed trials was shown at the end of testing phase. Participants whose correct responses were at or below 50% had to repeat the testing.

EEG acquisition and preprocessing

Continuous EEG data sampled at 1000Hz was recorded using a 129-channel HydroCel Geodesic Sensor Net connected to a NetStation amplifier (Net Amps 400; Electrical Geodesics Inc., Eugene, OR, USA). Electrode impedances were kept below 50kΩ, and electrodes were referenced online to Cz. First, offline filtering involved a 0.1Hz high-pass and 40Hz low-pass as well as 50Hz notch (all filters were second-order Butterworth filters with –12dB/octave roll-off, computed linearly with forward and backward passes to eliminate phase-shift). Next, the EEG was segmented into peri-stimulus epochs from 100ms before distractor onset to 500ms after distractor onset. An automatic artefact rejection criterion of ±100µV was used, along with visual inspection. Epochs were then screened for transient noise, eye movements, and muscle artefacts using a semi-automated artefact rejection procedure. Data from artefact contaminated electrodes were interpolated using three-dimensional splines (Perrin et al., 1987). Across all experiment, 11% of epochs were removed on average and 8 electrodes were interpolated per participant (6% of the total electrode montage).

Cleaned epochs were averaged, baseline corrected to the 100ms pre-distractor time interval, and re-referenced to the average reference. To eliminate residual environmental noise in the data, a 50Hz filter was applied[2]. All the above steps were done separately for ERPs from the four distractor conditions, and separately for distractors in the left and right hemifield. We next relabeled ERPs from certain conditions, as is done in traditional lateralised ERP analyses (like those of the N2pc). Namely, we relabelled single-trial data from all conditions where distractors appeared on the *left* so that the electrodes over the

[2] While filtering following epoch creation is normally discouraged (e.g., Widmann et al., 2015), control analyses we have carried out demonstrated that our filtering procedure was necessary and did not harm the data quality within our time-window of interest (0 – ~300ms post-distractor).

left hemiscalp now represented the activity over the right hemiscalp, and electrodes over the right hemiscalp – represented activity over the left hemiscalp, thus creating "mirror distractor-on-the-right" single-trial data. Next, these mirrored data and the veridical "distractor-on-the-right" data from each of the 4 distractor conditions were averaged together, creating a single average ERP for each of the 4 distractor conditions. The contralaterality factor (i.e. contralateral vs. ipsilateral potentials) is normally represented by separate ERPs (one for contralateral activity, and one for ipsilateral activity; logically more pairs for pair-wise N2pc analyses). In our procedure, the lateralised voltage gradients across the whole scalp are preserved within each averaged ERP by simultaneous inclusion of both contralateral and ipsilateral hemiscalp activation. Such a procedure enabled us to fully utilise the capability of the electrical neuroimaging analyses in revealing both lateralised and non-lateralised mechanisms that support the interactions of attentional control with context control. As a result of the relabelling, we obtained 4 different ERPs: TCCV (target colour-cue, Visual), NCCV (nontarget colour-cue, Visual), TCCAV (target colour-cue, AudioVisual), NCCAV (nontarget colour-cue, AudioVisual). Preprocessing and EEG analyses, unless otherwise stated, were conducted using CarTool software (available for free at www.fbmlab.com/cartool-software/; Brunet, Murray, & Michel, 2011).

Data analysis design

Behavioural analyses. Like in Matusz and Eimer (2011), and because mean reaction times (RTs) and accuracy did not differ significantly between the four experiments, the basis of our analyses was RT spatial cueing effects (henceforth "behavioural capture effects"). These were calculated by subtracting the mean RTs for trials where the distractor and target were in the same location from the mean RTs for trials where the distractor and the target location differed, separately for each of the four distractor conditions. RT data were analysed using the repeated-measures analysis of variance (rmANOVA). Error rates (%) were also analysed. As they were not normally distributed, we analysed error rates using the Kruskal–Wallis *H* test and the Durbin test. The former was used to analyse if error rates differed significantly between experiments, while the latter was used to analyse differences between experimental conditions within each experiment separately.

Following Matusz and Eimer (2011), RT data were cleaned by discarding incorrect and missed trials, as well as RTs below 200ms and above 1000ms. Additionally, to enable

more direct comparisons with the developmental study for which current Experiment 1 served as an adult control (Turoman et al., 2020a, 2020b), we have further removed trials with RTs outside 2.5SD of the individual mean RT. As a result, a total of 5% of trials across all experiments were removed. Next, behavioural capture effects were submitted to a four-way 2 × 2 × 2 × 2 rmANOVA with factors: Distractor Colour (TCC vs. NCC), Distractor Modality (V vs. AV), Multisensory Relationship (MR; Arbitrary vs. Congruent), and Distractor Onset (DO; Unpredictable vs. Predictable). Due to the error data not fulfilling criteria for normality, we used Cue-Target location as a factor in the analysis, conducting 3-way Durbin tests for each experiment, with factors Distractor Colour, Distractor Modality, and Cue-Target Location. All analyses, including post-hoc paired *t*-tests, were conducted using SPSS for Macintosh 26.0 (Armonk, NY: IBM Corp). For brevity, we only present the RT results in the Results, and the error rate results can be found in Supplemental Online Materials ("SOMs" henceforth).

ERP analyses. The preprocessing of the ERPs triggered by the visual and audiovisual distractors across the 4 different experimental blocks created ERP averages in which the contralateral versus ipsilateral ERP voltage gradients across the whole scalp were preserved. We first conducted a canonical N2pc analysis, as the N2pc is a well-studied and well-understood correlate of attentional selection in visual settings. However, it is unclear if the N2pc also indexes bottom-up attentional selection modulations by multisensory stimuli, or top-down modulations by contextual factors like multisensory semantic relationships (for visual-only study, see e.g., Wu et al., 2015) or stimulus onset predictability (for visual-only study, see e.g., Burra & Kerzel, 2013). N2pc analyses served also to bridge EN analyses with the existing literature and EEG approaches more commonly used to investigate attentional control. Briefly, EN encompasses a set of multivariate, reference-independent analyses of global features of the electric field measured at the scalp (Biasiucci et al., 2019; Koenig et al., 2014; Michel & Murray, 2012; Murray, Brunet, & Michel, 2008; Lehmann & Skrandies, 1980; Tivadar & Murray, 2019; Tzovara et al., 2012). The key advantages of EN analyses over canonical N2pc analyses and how the former can complement the latter when combined, are described in the Introduction.

Canonical N2pc analysis. To analyse lateralised mechanisms using the traditional N2pc approach, we extracted mean amplitude values from, first, two electrode clusters comprising PO7/8 electrode equivalents (e65/90; most frequent electrode pair used to

analyse the N2pc), and, second, their six immediate surrounding neighbours (e58/e96, e59/e91, e64/e95, e66/e84, e69/e89, e70/e83), over the 180–300ms post-distractor time-window (based on time-windows commonly used in traditional N2pc studies, e.g., Luck & Hillyard, 1994b; Eimer, 1996; including distractor-locked N2pc, Eimer & Kiss 2008; Eimer et al., 2009). Analyses were conducted on the mean amplitude of the N2pc difference waveforms, which were obtained by subtracting the average of amplitudes in the ipsilateral posterior-occipital cluster from the average of amplitudes in the contralateral posterior-occipital cluster. This step helped mitigate the loss of statistical power that could result from the addition of contextual factors into the design. N2pc means were thus submitted to a 4-way 2 × 2 × 2 × 2 rmANOVA with factors Distractor Colour (TCC vs. NCC), Distractor Modality (V vs. AV), MR (Arbitrary vs. Congruent), and DO (Unpredictable vs. Predictable), analogously to the behavioural analysis.

Electrical Neuroimaging of the N2pc component. Our EN analyses separately tested response strength and topography in N2pc-like lateralised ERPs (see e.g. Matusz et al., 2019b for a detailed, tutorial-like description of how EN measures can aid the study of attentional control processes). We assessed if interactions between visual goals, multisensory salience and contextual factors 1) modulated the distractor-elicited lateralised ERPs, and 2) if they do so by altering the strength of responses within statistically indistinguishable brain networks and/or altering the recruited brain networks.

To test for the involvement of strength-based spatially-selective mechanisms, we analysed Global Field Power (GFP) in lateralised ERPs. GFP is the root mean square of potential [μV] across the entire electrode montage (see Lehmann & Skrandies, 1980). To test for the involvement of network-related spatially-selective mechanisms, we analysed stable patterns in ERP topography characterising different experimental conditions using a clustering approach known as the Topographic Atomize and Agglomerate Hierarchical Clustering (TAAHC). This clustering ("segmentation") procedure generates sets of clusters of topographical maps that predict the largest variance within the group-averaged ERP data. Each cluster is labelled with a 'template map' that represents the centroid of its cluster. The optimal number of clusters is one that explains the largest global explained variance (GEV) in the group-averaged ERP data with the smallest number of template maps, and which we identified using the modified Krzanowski–Lai criterion (Murray et al., 2008). In the next step, i.e., the so-called fitting procedure, the single-subject data was 'fitted' back onto the

segmentation results, such that each datapoint of each subject's ERP data over a chosen time-window was labelled by the template map with which it was best spatially correlated. This procedure resulted in a number of timeframes that a given template map was present over a given time-window, which durations (in milliseconds) we then submitted to statistical analyses described below.

In the present study, we conducted strength- and network-based analyses using the same 4-way repeated-measures design as in the behavioural and canonical N2pc analyses, on the lateralised whole-montage ERP data. Since the N2pc is a lateralised ERP, we first conducted an EN analysis of lateralised ERPs in order to uncover the modulations of the N2pc by contextual factors. To obtain *global* EN measures of *lateralised* N2pc effects, we computed a difference ERP by subtracting the voltages over the contralateral and ipsilateral hemiscalp, separately for each of the 4 distractor conditions. This resulted in a 59-channel difference ERP (as the midline electrodes from the 129-electrode montage were not informative). Next, this difference ERP was mirrored onto the other side of the scalp, recreating a "fake" 129 montage (with values on midline electrodes now set to 0). It was on these mirrored "fake" 129-channel lateralised difference ERPs that lateralised strength-based and topography-based EN analyses were performed. Here, GFP was extracted over the canonical 180–300ms N2pc time-window and submitted to a 2 × 2 × 2 × 2 rmANOVA with factors Distractor Colour (TCC vs. NCC), Distractor Modality (V vs. AV), as well as the two new factors, MR (Arbitrary vs. Congruent), and Distractor Onset (DO; Unpredictable vs. Predictable). Meanwhile, for topographic analyses, the "fake" 129-channel data across the 4 experiments were submitted to a segmentation over the entire post-distractor period. Next, the data were fitted back over the 180-300ms period. Finally, the resulting number of timeframes (in ms) was submitted to the same rmANOVA as the GFP data above.

It remains unknown if the tested contextual factors modulate lateralised ERP mechanisms at all. Given evidence that semantic information and temporal expectations can modulate *nonlateralised* ERPs within the first 100-150ms post-stimulus (e.g., Dell'Acqua et al., 2010; Dassanayake et al., 2016), we also investigated the influence of contextual factors on nonlateralised voltage gradients, in an exploratory fashion. It must be noted that ERPs are sensitive to the inherent physical differences in visual and audiovisual conditions. Specifically, on audiovisual trials, the distractor-induced ERPs would be contaminated by brain response modulations induced by sound processing, with these modulations visible in

our data already at 40ms post-distractor. Consequently, any direct comparison of visual-only and audiovisual ERPs would index auditory processing per se and not capture of attention by audiovisual stimuli. Such confounded sound-related activity is eliminated in the canonical N2pc analyses through the contralateral-minus-ipsilateral subtraction. To eliminate this confound in our EN analyses here, we calculated difference ERPs, first between TCCV and NCCV conditions, and then between TCCAV and NCCAV conditions. Such difference ERPs, just as the canonical N2pc difference waveform, subtract out the sound processing confound in visually-induced ERPs. As a result of those difference ERPs, we removed factors Distractor Colour and Distractor Modality, and produced a new factor, Target Difference (two levels: D_{AV} [TCCAV – NCCAV difference] and D_V [TCCV – NCCV difference]), that indexed the enhancement of visual attentional control by sound presence.

All nonlateralised EN analyses involving context factors were conducted on these difference ERPs and included the factor Target Difference. Strength-based analyses, voltage and GFP data were submitted to 3-way rmANOVAs with factors: MR (Arbitrary vs. Congruent), DO (Unpredictable vs. Predictable), and Target Difference (D_{AV} vs. D_V), and analysed using the STEN toolbox 1.0 (available for free at https://zenodo.org/record/1167723#.XS3lsi17E6h),. Follow-up tests involved further ANOVAs and pairwise *t*-tests. To correct for temporal and spatial correlation (see Guthrie & Buchwald, 1991), we applied a temporal criterion of >15 contiguous timeframes, and a spatial criterion of >10% of the 129- channel electrode montage at a given latency for the detection of statistically significant effects at an alpha level of 0.05. As part of topography-based analyses, we segmented the ERP difference data across the post-distractor and pre-target onset period (0 – 300ms from distractor onset) and conducted clustering of the data to obtain template maps. Next, the data were fitted onto the canonical N2pc time-window (180–300ms) as well as also earlier time-periods that were highlighted by the GFP data as representing significant condition differences. The resulting map presence (in ms) over the given time-windows were submitted to 4-way rmANOVAs with factors: MR (Arbitrary vs. Congruent), DO (Unpredictable vs. Predictable), Target Difference (D_{AV} vs. D_V), and Map (different numbers of maps for different time-windows), followed by post-hoc *t*-tests. Maps with durations <15 contiguous timeframes were not included in the analyses. Unless otherwise stated in the Results, map durations were statistically different from 0ms (as confirmed by post-hoc one-sample t-tests), meaning that they were reliably present across

the time-windows of interest. Holm-Bonferroni corrections (Holm, 1979) were used to correct for multiple comparisons between map durations. Comparisons passed the correction unless otherwise stated.

Results

Behavioural analyses

Interaction of TAC and MSE with contextual factors

To shed light on attentional control in naturalistic settings, we first tested whether top-down visual control indexed by TAC interacted with contextual factors in behavioural measures. Our 2 × 2 × 2 × 2 rmANOVA revealed several main effects and interactions, both expected and unexpected (full description of the results in SOMs). We confirmed presence of TAC, via a main effect of Distractor Colour, $F_{(1, 38)} = 340.4$, $p < 0.001$, $\eta_p^2 = 0.9$, with TCC distractors (42ms), but not NCCs (-1ms), eliciting reliable behavioural capture effects.

[FIGURE 2 HERE]

Shedding first light on our question, the strength of TAC was dependent on the multisensory relationship within distractors, demonstrated by a 2-way Distractor Colour × MR interaction, $F_{(1, 38)} = 4.5$, $p = 0.041$, $\eta_p^2 = 0.1$ (Figure 2). This effect was driven by behavioural capture effects elicited by TCC distractors being reliably larger when arbitrary (45ms) than congruent (40ms), $t_{(38)} = 1.9$, $p = 0.027$. NCC distractors showed no evidence of MR modulation (Arbitrary vs. Congruent, $t_{(38)} = 1$, $p = 0.43$). Contrastingly, TAC showed no evidence of modulation by predictability of the distractor onset (no 2-way Distractor Colour × DO interaction, $F_{(1, 38)} = 2$, $p = 0.16$).Thus, visual feature-based control interacted with contextual factor of distractor semantic congruence but not its temporal predictability.

Next, we wanted to shed light on potential interactions of multisensory enhancements with contextual factors. Expectedly, there was behavioural MSE (a significant main effect of Distractor Modality, $F_{(1, 38)}=13.5$, $p=0.001$, $\eta_p^2=0.3$), where visually-elicited behavioural capture effects (18ms) were enhanced on AV trials (23ms). Unlike TAC, this MSE effect showed no evidence of interaction with either contextual factor (Distractor Modality

x MR interaction, $F<1$; Distractor Modality x DO interaction: n.s. trend, $F_{(1, 38)}=3.6$, $p=0.07$, $\eta_p^2= 0.1$). Thus, behaviourally, multisensory enhancement of attentional capture was not modulated by distractor's semantic relationship or its predictability. We have also observed unexpected effects but as these were outside of the focus of the current paper, which aims to elucidate the interactions between visual (goal-based) and multisensory (salience-driven) attentional control and contextual mechanisms, we describe them in SOMs.

ERP analyses

Lateralised (N2pc-like) brain mechanisms

We next set out to investigate the type of brain mechanisms that underlie interactions between more traditional attentional control (TAC, MSE) and contextual control over attentional selection. Our analyses on the lateralised responses, spanning both canonical and EN framework, revealed little evidence for a role of spatially-selective mechanisms in supporting those interactions. Both canonical N2pc and EN analyses confirmed presence of TAC (see Fig.3 for N2pc waveforms across 4 condition), but TAC did not interact with either contextual factors. Lateralised ERPs showed no evidence also for sensitivity to MSE (again in neither canonical, nor EN analyses) or for interactions between MSE and the contextual factors. In fact, even main effects of MR and DO[3] in lateralised responses were absent. (See SOMs for full description of the results of lateralised ERP analyses).

[FIGURE 3 HERE]

Nonlateralised brain mechanisms

A major part of our analyses focused on understanding the role of nonlateralised ERP mechanisms in the interactions between visual goals (TAC), multisensory salience (MSE) and contextual control. To remind the reader, to prevent these nonlateralised ERPs from being confounded by the presence of sound on AV trials, we based our analyses here on the

[3] Any ERP results related to DO are unlikely to be confounded by shifted baseline due to potential dominance of one ISI type (100ms, 250ms, 450ms) over others, as no such dominance was identified in a subsample of data.

difference ERPs indexing visual attentional control under sound absence vs. presence. That is, we calculated ERPs of the difference between TCCV and NCCV conditions, and between TCCAV and NCCAV conditions (D_V and D_{AV} levels, respectively, of the factor Target Difference). We focus the description of these results on the effects of interest (see SOMs for full description of results).

The 2 × 2 × 2 (MR × DO × Target Difference) rmANOVA on electrode-wise voltage analyses revealed a main effect of Target Difference at 53–99ms and 141–179ms. Thus, the three factors interacted at both early and later (N2pc-like) latencies encompassing perceptual and attentional selection processing stages. Across both time-windows, amplitudes were larger D_{AV} (TCCAV – NCCAV difference) than for D_V (TCCV – NCCV difference). This effect was further modulated by the multisensory relationship within the distractors, with a 2-way Target Difference × MR interaction, at the following time-windows: 65–103ms, 143–171ms, and 194–221ms (all *p*'s < 0.05). This effect was driven by semantically congruent distractors showing larger amplitudes for D_{AV} than D_V within all 3 time-windows (65–97ms, 143–171ms, and 194–221ms; all *p*'s < 0.05). No similar differences were found for arbitrary distractors, and there were no other interactions that passed the temporal and spatial criteria for multiple comparisons of >15 contiguous timeframes and >10% of the 129- channel electrode montage.

Interaction of TAC with contextual factors. We next used EN analyses to investigate the contribution of the strength- and topography-based nonlateralised mechanisms to the interactions between TAC and contextual factors. *Strength-based brain mechanisms.* A 2 × 2 × 2 Target Difference × MR × DO rmANOVA on the GFP, mirroring the electrode-wise analysis on ERP voltages, also showed the main effect of Target Difference spanning a large part of the first 300ms post-distractor both before and in N2pc-like time-windows (19–213ms, 221–255ms, and 275–290ms), where, just like the voltages, also GFP was larger for D_{AV} than D_V (all *p*'s < 0.05).

In GFP, Target Difference interacted both with MR (23–255ms) and separately with DO (88–127ms; see SOMs for full description), but most notable was the 3-way Target Difference × MR × DO interaction, spanning 102–124ms and 234–249ms. We then followed up this interaction with series of post-hocs to gauge the modulations of TAC (and MSE, see below) by the two contextual factors. In GFP, MR and DO interacted independently of

Target Difference in the second time-window, which results we describe in the SOMs. To gauge differences in the strength of TAC in GFP across the 4 contexts, we focused the comparisons on only visually-elicited target differences (to minimise any potential confounding influences from sound processing) across the respective levels of the 2 contextual factors. Weakest GFPs were observed for arbitrary predictable distractors (

A). They were weaker than GFPs elicited arbitrary unpredictable distractors (102–124ms and 234–249ms), and predictable congruent distractors (only in the later window, 234–249ms).

Topography-based brain mechanisms. We focused the segmentation of the TAC-related topographic activity on the whole 0–300ms post-distractor time-window (before the target onset), which revealed 10 clusters that explained 82% of the global explained variance within the visual-only ERPs. This time-window was largely composed of distinct template maps across the 4 context conditions, except a time-window of 29–126ms post-distractor where segmentation revealed template maps shared across conditions. A 2 × 2 × 5 rmANOVA on the map presence over the 29–126ms post-distractor time-window revealed a 3-way MR × DO ×Map interaction, $F_{(3.2,122)} = 5.3$, $p = 0.002$, $\eta_p^2 = 0.1$.

Follow-up tests in the 29–126ms time-window focused on maps differentiating between the 4 conditions (results of follow-ups as a function of MR and DO are visible in

B in leftward panel and rightward panel, respectively). These results showed that context altered the processing of distractors from early on. It also did so by engaging different networks for the majority of different combinations of predictability and semantic relationship within the distractor stimuli: arbitrary predictable (Map A1), semantically congruent predictable (Map A2) and semantically congruent unpredictable (Map A5).

Arbitrary predictable distractors, which elicited the weakest GFP, largely recruited Map A1. This map predominated responses to them (37ms) vs. to semantically congruent predictable distractors (21ms), $t_{(38)} = 2.7$, $p < 0.01$ (Fig.4B rightward panel). In contrast, another map they activated (Map A2), predominated responses to arbitrary unpredictable (35ms) distractors, compared to them (18ms), $t_{(38)} = 2.96$, $p = .01$ (Fig.4B leftward panel) .

[FIGURE 4 HERE]

[FIGURE 5 HERE]

Semantically congruent predictable distractors mainly recruited Map A2. Map A2 predominated responses to them (25ms) vs. to semantically congruent unpredictable distractors (14ms), $t_{(38)} = 2$, $p = 0.04$. It likewise predominated responses to them compared to arbitrary unpredictable distractors (14ms), $t_{(38)} = 3.7$, $p = 0.001$.

Semantically congruent unpredictable distractors principally recruited Map A5. Map A5 predominated responses to them (34ms) vs. to semantically congruent predictable (19ms) distractors, $t_{(38)} = 2.7$, $p = 0.04$. It likewise predominated responses to it compared to arbitrary unpredictable (12ms) distractors, $t_{(38)} = 3.7$, $p < 0.001$.

Interaction of MSE with contextual factors. We then analysed the strength- and topography-based nonlateralised mechanisms contributing to the interactions between MSE and contextual factors.

Strength-based brain mechanisms. To gauge the AV-induced MS enhancements between D_{AV} and D_V across the 4 contexts, we explored the abovementioned 2 × 2 × 2 GFP interaction using a series of simple follow-up post-hocs. We first tested if the response strength between D_{AV} and D_V was reliably different within each of the 4 contextual conditions. AV-induced target ERP responses were enhanced (i.e. larger GFP for D_{AV} than D_V distractors) for both predictable and unpredictable semantically congruent distractors, across both earlier and later time-windows. AV enhancements were also found arbitrary predictable distractors, but only in the earlier (102–124ms) time-window; unpredictable distractors showed similar GFPs across D_{AV} and D_V trials. Next, we compare the AV-induced MS enhancements across the 4 contexts, by creating (D_{AV} minus D_V) difference ERPs or each

context. AV-induced enhancements were weaker for predictable arbitrary distractors than predictable semantically congruent distractors (102–124ms and 234–249ms; **Error! Reference source not found.**A).

Topography-based brain mechanisms. We focused the segmentation of the MSE-related topographic activity in the 0–300ms post-distractor time-window, which revealed 7 clusters that explained 78% of the GEV within the AV-V target difference ERPs. We fit the data in three subsequent time-windows where there were clear topographic patterns, 35–110ms, 110– 190ms, and 190– 300ms. To foreshadow the results, in the first and third time-window the MSE-related templates were modulated only by MR, while in the middle time-window – by both contextual factors.

In the first, 35–110ms time-window, the modulation of map presence by MR was evidenced by a 2-way Map × MR interaction, $F_{(2.1,77.9)} = 9.2$, $p < 0.001$, $\eta_p^2 = 0.2$. This effect was driven by one map (map B3) that in this time-window predominated responses to semantically congruent (42ms) vs. arbitrary (25ms) distractors, $t_{(38)} = 4.3$, $p = 0.02$, whereas another map (map B5) predominated responses to arbitrary (33ms) vs. semantically congruent (18ms) distractors, $t_{(38)} = 4$, $p = 0.01$ (**Error! Reference source not found.**B leftward panel).

In the second, 110–190ms time-window, map presence was modulated by both contextual factors, with a 3-way Map × MR × DO interaction, $F_{(2.6,99.9)} = 3.7$, $p = 0.02$, $\eta_p^2 = 0.1$, as it did for TAC. We focused follow-up tests in that time-window again on maps differentiating between the 4 conditions, as we did for the 3-way interaction for TAC (results of follow-ups as a function of MR and DO are visible in **Error! Reference source not found.**B, middle upper and lower panels, respectively). Context processes again interacted to modulate the processing of distractors, although now it did so after the first 100ms. It did so again by engaging different networks for different combinations of predictability and semantic relationship within the distractor stimuli: arbitrary predictable (Map B1), semantically congruent predictable (Map B3), semantically congruent unpredictable (Map B6), and now also arbitrary unpredictable (Map B5) distractors.

Arbitrary predictable distractors, which again elicited the weakest GFP, mainly recruited map B1. This map predominated responses to them (35ms; **Error! Reference source not found.**B middle upper panel) vs. to arbitrary unpredictable distractors (18ms,

$t_{(38)}$ = 2.8, p = 0.01), and vs. semantically congruent predictable distractors (17ms, $t_{(38)}$ = 2.8, p = 0.01; **Error! Reference source not found.**B middle lower panel).

Semantically congruent predictable distractors mostly recruited Map B3. Map B3 predominated responses to them (25ms) vs. to semantically congruent unpredictable distractors (12ms, $t_{(38)}$ = 2.2, p = 0.05), and vs. predictable arbitrary distractors (8ms, $t_{(38)}$ = 2.2, p = 0.005).

Semantically congruent unpredictable distractors principally recruited Map B6. That map predominated responses to them (37ms) vs. to semantically congruent predictable distractors (21ms, $t_{(38)}$ = 2.5, p = 0.02), and vs. arbitrary unpredictable distractors (24ms, $t_{(38)}$ = 2.3, p = 0.04).

Finally, now also *arbitrary unpredictable* distractors largely recruited one map, Map B5. Map B5 predominated responses to them (33ms) vs. to arbitrary predictable distractors (17ms, $t_{(38)}$ = 2.6, p = 0.04), and vs. semantically congruent unpredictable distractors (13ms, $t_{(38)}$ = 3.4, p = 0.002).

In the third, 190–300ms time-window, the 2-way Map × MR interaction was reliable at $F_{(3.2, 121.6)}$ = 3.7, p = 0.01, η_p^2= 0.1. Notably, the same map as before (map B3) predominated responses to semantically congruent (50ms) vs. arbitrary distractors (33ms), $t_{(38)}$ = 3.6, p = 0.08, and another map (map B1) predominated responses to arbitrary (25ms) distractors vs. semantically congruent (14ms) distractors, $t_{(38)}$ = 2.3, p = 0.02 (**Error! Reference source not found.**B rightward panel).

Discussion

Attentional control is known to be necessary to cope with the multitude of stimulation in everyday situations. However, in such situations observer's goals and stimulus' salience routinely interact with contextual processes, but such multi-pronged interactions between control processes have never been studied. Below, we discuss our findings on how visual and multisensory attentional control, respectively, interact with distractor temporal predictability and semantic relationship. We then discuss the spatiotemporal dynamics in nonlateralised brain mechanisms underlying these interactions. Finally, we discuss how our results enrich the understanding of attentional control in real-world settings.

Interaction of task-set contingent attentional capture with contextual control

Visual control interacted most robustly with stimulus' semantic relationship. Behaviourally, *target-matching* visual distractors captured attention more strongly when they were arbitrary than semantically congruent. This was accompanied by a sequence of modulations of nonlateralised brain responses, spanning both the attentional selection, N2pc-like stage and much earlier, perceptual stages. Arbitrary distractors, but only predictable ones, first recruited one particular brain network (Map A1), to a larger extent than predictable semantically congruent distractors, and did so early on (29–126ms post-distractor). Arbitrary predictable distractors elicited also suppressed responses, in the later part of this early time-window (102–124ms; where they elicited the weakest responses). In the later, N2pc-like (234–249ms) time-window, responses to arbitrary predictable distractors were again weaker, now compared to semantically congruent predictable distractors.

This potential cascade of network- and strength-based modulations of nonlateralised brain responses might epitomise a potential brain mechanism for interactions between visual top-down control and multiple sources of contextual control, as they are consistent with existing literature. The discovered early (~30-100ms) network-based modulations for predictable target-matching (compared to unpredictable) distractors is consistent with predictions attenuating the earliest visual perceptual stages (C1 component, ~50–100ms post-stimulus; Dassanayake et al., 2016). The subsequent, mid-latency response suppressions (102–124ms, where we found also network-based modulations) for predictable distractors are in line with N1 attenuations for self-generated sounds (Baess et al., 2011; Klaffehn et al., 2019), and the latencies where the brain might promote the processing of unexpected events (Press et al., 2020). Notably, these latencies are also in line with the onset (~115ms post-stimulus) of the goal-based suppression of salient visual distractors (here: presented simultaneously with targets), i.e., distractor positivity (Pd; Sawaki & Luck 2010). Finally, the response suppressions we found at later, N2pc-like, attentional selection stages (234–249ms), are also consistent with some extant (albeit scarce) literature. Van Moorselaar and Slagter (2019) showed that when such salient visual distractors appear in predictable locations, they elicit the N2pc but no longer a (subsequent, post-target) Pd, suggesting that once the brain learns the distractor's location, it can suppress it without the need for active inhibition. More recently, van Moorselaar et al.,

(2020b) showed that the representation of the predictable distractor feature could be decoded already from pre-stimulus activity. While our paradigm was not optimised for revealing such effects, pre-stimulus mechanisms could indeed explain our early-onset (~30ms) context-elicited neural effects. The robust response suppressions for predictable stimuli are also consistent with recent proposals for interactions between predictions and auditory attention processes. Schroeger et al., (2015) suggested that larger attention is deployed to more "salient" stimuli, i.e., those for which a prediction is missing, so that the predictive model can be reconfigured to encompass such predictions in the future. This reconfiguration, in turn, requires top-down goal-based attentional control. Our results extend this model to the visual domain. Our findings involving the response modulation 'cascade' and behavioural benefits may also support the Schroeger et al.'s tenet that different, but connected, predictive models exist at different levels of the cortical hierarchy.

These existing findings jointly strengthen our interpretations that goal-based top-down control utilises contextual information to alter visual processing from its very early stages. Our findings also extend the extant ideas in several ways. First, they show that in context-rich settings (i.e., involving multiple sources of contextual control), goal-based control will use both stimulus-related predictions and stimulus meaning to facilitate task-relevant processing. Second, context information modulates not only early, pre-stimulus and late, attentional stages, but also early responses elicited by a stimulus. Third, our findings also suggest candidate mechanisms for supporting interactions between goal-based control and multiple sources of contextual information. Namely, context will modulate the early stimulus processing by recruiting distinct brain networks for stimuli representing different contexts, e.g. the brain networks recruited by predictable distractors differed for arbitrarily linked and semantically congruent stimuli (Map A1 and A2, respectively). Also, the distinct network recruitment might lead to the suppressed (potentially more efficient; c.f. repetition suppression, Grill-Spector et al., 2006) brain responses. These early response attenuations will extend also to later stages, associated with attentional selection. Thus, it is the early differential brain network recruitment that might trigger a cascade of spatiotemporal brain dynamics leading effectively to the stronger behavioural capture, here for predictable (arbitrary) distractors. However, for distractors, these behavioural benefits may be most robust for arbitrary target-matching stimuli (as opposed to semantically congruent), with prediction-based effects are less apparent. We develop this point below.

Interaction of multisensory enhancement of attentional capture with contextual control

Multisensory-induced processes likewise interacted with contextual processes, but these effects were found only in brain responses. To measure effects related to multisensory-elicited modulations and to its interactions with contextual information, we analysed AV–V differences within the Target Difference ERPs.

The interactions between multisensory modulations and context processes were also instantiated via an early-onset 'cascade' of strength- and network-based nonlateralised brain mechanisms. This sequence of response modulations again started early (now 35–110ms post-distractor). A separate segmentation analysis revealed that in the multisensory-modulated responses the brain first distinguished only between semantically congruent and arbitrarily linked distractors. These distractors recruited predominantly different brain networks (Map B3 and B5, respectively). Around the end of these network-based modulations, at 102–124ms, multisensory-elicited brain responses were also modulated in their strength. Arbitrary predictable distractors again triggered weaker responses, now compared to semantically congruent predictable distractors. Multisensory-elicited responses predominantly recruited distinct brain networks for the four distractor types from 110ms and till 190ms post-distractor, thus spanning stages linked to perception and attentional selection. The two brain networks activated earlier (reflected by maps B3 and B5) were now recruited for responses to semantically congruent predictable and arbitrary unpredictable distractors, respectively. In turn, two other brain networks (reflected by B1 and B6), were recruited for responses to arbitrary predictable and semantically congruent unpredictable distractors, respectively. In the subsequent time-window (190–300ms) that mirrors the time-window used in the canonical N2pc analyses, the multisensory-related responses again recruited different brain networks, now distinguishing (as in the first window of topographic modulations) between distractors based on their multisensory relationship. There, Map B3 (in the middle time-window recruited by congruent predictable distractors) again was predominantly recruited by semantically congruent over arbitrary distractors, and now Map B1 (in the middle time-window recruited by arbitrary predictable distractors) - for arbitrary distractors over congruent ones. In the middle of this time-window (234–249ms), responses were again modulated in their strength, with predictable

arbitrary distractors again eliciting weaker responses compared to semantically congruent predictable distractors.

To summarise, distractor's semantic relationship played a dominant (but not absolute) role in interactions between multisensory-elicited and contextual processes. The AV–V difference ERPs were modulated exclusively by multisensory relationships both in the earliest, perceptual (35–110ms) time-window and latest, N2pc-like (190–300ms) time-window linked to attentional selection. At both stages, distinct brain networks were recruited predominantly by semantically congruent and arbitrary distractors. These results suggest that stimulus processing is affected by whether it holds a meaning to the observer, from early perceptual stages to the stages of its attentional selection. Notably, the same brain network (Map B3) supported multisensory processing of semantically congruent distractors across both time-windows, while different networks were recruited by arbitrarily linked distractors.

Thus, a single network is potentially recruited for processing of meaningful multisensory stimuli. Behavioural results suggest that this brain network is involved in suppressing attentional capture by semantically congruent (over arbitrarily linked) distractors in service of top-down goal-driven attentional control. This idea is supported by the interactions between distractors' multisensory-driven modulations, their multisensory relationship and temporal predictability in the second, 110–190ms time-window. During that time interval, the brain network reflected by the same, "semantic" (Map B3) template map was still active, albeit now it was recruited for responses to semantically congruent (rather than arbitrary) *predictable* distractors. Based on the existing evidence that predictions are used as means for goal-based behaviour (Schroeger et al., 2015; van Moorselaar et al., 2020a; Matusz et al., 2016), one could argue that the brain network reflected by Map B3 serves to integrate contextual information across both predictions and objects' meaning. Alternatively, this brain network could be sensitive predominantly to the latter (as it remained recruited by semantically congruent distractors throughout the distractor-elicited response). The activity of this network might have contributed to the overall stronger brain responses (indicated by GFP results) to semantically congruent multisensory stimuli, which in turn contributed to the null behavioural multisensory enhancements of behavioural indices of attentional capture. While these are the first results of this kind, they open an exciting possibility that surface-level EEG/ERP studies can

reveal the network- and strength-related brain mechanisms (potentially a single network for "gain control" up-modulation) by which goal-based processes control (i.e., suppress) multisensorily-driven enhancements of attentional capture.

How we pay attention in naturalistic settings

It is now relatively well-established that the brain facilitates goal-directed processing (from perception to attentional selection) via processes based on observer's goals (e.g. Folk et al., 1992; Desimone & Duncan 1995), predictions about the outside world (Summerfield & Egner 2009; Schroeger et al., 2015; Press et al., 2020), and long-term memory contents (Summerfield et al., 2006; Peelen & Kastner 2014). Also, multisensory processes are increasingly recognised as an important source of bottom-up, attentional control (e.g. Spence & Santangelo 2007; Matusz & Eimer 2011; Matusz et al., 2019a; Fleming et al., 2020). By studying these processes largely in isolation, researchers clarified how they support goal-directed behaviour. However, in the real world, observer's goals interact with multisensory processes and multiple types of contextual information. Our study sheds first light on this "naturalistic attentional control".

Understanding of attentional control in the real world has been advanced by research on feature-related mechanisms (Theeuwes 1991; Folk et al., 1992; Desimone & Duncan 1995; Luck et al., 2020), which support attentional control where target location information is missing. Here, we aimed to increase the ecological validity of this research by investigating how visual feature-based attention (TAC) transpires in context-rich, multisensory settings (see SOMs for a discussion of our replication of TAC). Our findings of reduced capture for semantically congruent than artificially linked target-colour matching distractors is novel and important, as they suggest stimuli's meaning is also utilised to suppress attention (to distractors). Until now, known benefits of meaning were limited to target selection (Thorpe et al., 1996; Iordanescu et al., 2008; Matusz et al., 2019a). Folk et al (1992) famously demonstrated that attentional capture by distractors is sensitive to the observer's goals; we reveal that distractor's meaning may serve as a second source of goal-based attentional control. This provides a richer explanation for how we stay focused on task in everyday situations, despite many objects matching attributes of our current behavioural goals.

To summarise, in the real world, attention should be captured more strongly by stimuli that are unpredictable (Schroeder et al., 2015), but also by those unknown or without a clear meaning. On the other hand, stimuli with high strong spatial and/or temporal alignment across senses (and so stronger bottom-up salience) may be more resistant to such goal-based attentional control (suppression), as we have shown here (MSE effect; also Santangelo & Spence 2007; Matusz & Eimer 2011; van der Burg et al., 2011; Turoman et al., 2020a; Fleming et al., 2020). As multisensory distractors captured attention more strongly even in current, context-rich settings, this confirms the importance of multisensory salience as a source of *potential* bottom-up attentional control in naturalistic environments (see SOMs for a short discussion of this replication).

The investigation of brain mechanisms underlying known EEG/ERP correlates (N2pc, for TAC) via advanced multivariate analyses has enabled us to provide a comprehensive, novel account of attentional control in a multi-sensory, context-rich setting. Our results jointly support the predominance of goal-based control in naturalistic settings. Multisensory semantic congruence reduced behavioural attentional capture by target-matching colour distractors compared to arbitrarily linked distractors. Context modulated nonlateralised brain responses to target-related (TAC) distractors via a sequence of strength- and network-based mechanisms from early (~30ms post-distractor) through later perceptual to later, attentional selection stages. While these results are first of this kind and need replication, they suggest that context-based goal-directed modulations of distractor processing onset at early stages (potentially involving pre-stimulus processes; e.g. van Moorselaar & Slagter, 2020) to control behavioural attentional selection. Responses to predictable arbitrary (target-matching) distractors revealed by our EN analyses might have driven the larger behavioural capture for arbitrary than semantically congruent distractors. The former engaged distinct brain networks and triggered the weakest and potentially most efficient (Grill-Spector et al., 2006) responses. One reason for absence of our distractor-elicited effects in behavioural measures is their small magnitude: while the TAC effect is ~50 ms, both MSE and semantically-driven suppression were small, ~5ms. This may also be the reason why context-driven effects were absent in behavioural measures of multisensory enhancement of attentional capture, despite involving a complex, early-onsetting cascade of strength- and network-based modulations, like visual goal-based control. Alternatively, or additionally, our results point to a potential brain mechanism by which the stimulus'

semantic relationship is utilised for goal-directed behaviour. Namely, our EN analyses of surface-level EEG identified a brain network that is recruited by semantically congruent stimuli at early, perceptual stages, and remains actively recruited for them also attentional selection stages captured by the N2pc time-window. While remaining cautious with the interpretation of our results, this network might have contributed to the consistently enhanced AV-induced responses for semantically congruent, than arbitrary, multisensory distractors. These enhanced brain responses and suppressed behavioural attention are consistent with a "gain control" mechanism, in the context of distractor processing (e.g. Sawaki & Luck 2010; Luck et al., 2020). Our results reveal that such "gain control", at least in some cases, operates by relaying processing of certain stimuli to specific brain networks. We have purported the existence of such a mechanism in a different study on top-down multisensory attention (e.g. Matusz et al., 2019c). We emphasise that such in-depth insights into the nature of brain mechanisms of attentional control are readily gauged by multivariate, global (EN) analyses of surface-level data (within our methodological approach; Matusz et al., 2019c).

N2pc as an index of attentional control

We have previously discussed the limitations of canonical N2pc analyses in capturing neurocognitive mechanisms by which visual top-down goals and multisensory bottom-up salience simultaneously control attention selection (Matusz et al., 2019b). The mean N2pc amplitude modulations are commonly interpreted as "gain control", but they can be driven by both strength- (i.e., "gain") and network-based mechanisms. Canonical N2pc analyses cannot distinguish between those two brain mechanisms. Contrastingly, Matusz et al., (2019b) have shown evidence for both brain mechanisms underlying N2pc-like responses. These and other results of ours (Turoman et al., 2020a, 2020b) provided evidence from surface-level data to infer about distinct brain sources contributing to the N2pc's, a finding that has been previously shown only in *source*-level data (Hopf et al., 2000). These findings point to a certain limitation of the N2pc (canonically analysed), which is an EEG *correlate* of attentional selection, but where other analytical approaches are necessary to reveal brain mechanisms of attentional selection.

Here, we have shown that the lateralised, spatially-selective brain mechanisms, approximated by the N2pc and revealed by EN analyses are limited in how they contribute

to attentional control in some settings. Rich, multisensory, and context-laden influences over goal-based top-down attention are, in our current paradigm, not captured by such lateralised mechanisms. In contrast, nonlateralised (or at least *relatively less* lateralised, see Figures 4 and 5) brain networks seem to support such interactions for visual and multisensory distractors - from early on to the stage of attentional selection. We nevertheless want to reiterate that paradigms that can gauge N2pc offer an important starting point for studying attentional control in less traditional multisensory and/or context-rich settings. There, multivariate analyses and an EN framework might be useful in readily revealing new mechanistic insights into attentional control.

Broader implications

Our findings are important to consider when aiming to study attentional control, and information processing more generally, in naturalistic settings (e.g. while viewing movies, listening to audiostories) and veridical real-world environments (e.g. the classroom or the museum). Additionally, conceptualisations of ecological validity (Peelen et al., 2014; Shamay-Tsoory & Mendelsohn 2019; Vanderwal et al., 2019; Eickhoff et al., 2020; Cantlon 2020) should go beyond traditionally invoked components (e.g. observer's goals, context, socialness) to encompass contribution of multisensory processes. For example, naturalistic studies should compare unisensory and multisensory stimulus/material formats, to measure/estimate the contribution of multisensory-driven bottom-up salience to the processes of interest. More generally, our results highlight that hypotheses about how neurocognitive functions operate in everyday situations can be built already in the laboratory, if one manipulates systematically, together and across the senses, goals, salience and context (van Atteveldt et al., 2014, 2018; Matusz et al., 2019c). Such a cyclical approach (Matusz et al., 2019a; see also Naumann et al., 2020 for a new tool to measure ecological validity of an experiment) involving testing of hypotheses across laboratory and veridical real-world settings is most promising for successfully bridging the two, typically separately pursued types of research, creating more complete theories of naturalistic attentional control.

References

Baess, P., Horváth, J., Jacobsen, T., & Schröger, E. (2011). Selective suppression of self-initiated sounds in an auditory stream: An ERP study. *Psychophysiology*, *48*(9), 1276-1283.

Brunet, D., Murray, M. M., & Michel, C. M. (2011). Spatiotemporal analysis of multichannel EEG: CARTOOL. *Computational intelligence and neuroscience*, *2011*.

Burra, N., & Kerzel, D. (2013). Attentional capture during visual search is attenuated by target predictability: Evidence from the N2pc, Pd, and topographic segmentation. *Psychophysiology*, *50*(5), 422-430.

Cantlon, J. F. (2020). The balance of rigor and reality in developmental neuroscience. *NeuroImage*, *216*, 116464.

Cappe, C., Thut, G., Romei, V., & Murray, M. M. (2010). Auditory–visual multisensory interactions in humans: timing, topography, directionality, and sources. *Journal of Neuroscience*, 30(38), 12572-12580.

Chen, Y. C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389-404.

Chun, M. M., & Jiang, Y. (1998). Contextual cueingcueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, *36*(1), 28-71.

Correa, Á., Lupiáñez, J., & Tudela, P. (2005). Attentional preparation based on temporal expectancy modulates processing at the perceptual level. *Psychonomic bulletin & review*, 12(2), 328-334.

Coull, J. T., Frith, C. D., Büchel, C., & Nobre, A. C. (2000). Orienting attention in time: behavioural and neuroanatomical distinction between exogenous and endogenous shifts. *Neuropsychologia*, *38*(6), 808-819.

Dassanayake, T. L., Michie, P. T., & Fulham, R. (2016). Effect of temporal predictability on exogenous attentional modulation of feedforward processing in the striate cortex. *International Journal of Psychophysiology*, 105, 9-16.

De Meo, R., Murray, M. M., Clarke, S., & Matusz, P. J. (2015). Top-down control and early multisensory processes: chicken vs. egg. *Frontiers in integrative neuroscience*, *9*(17), 1-6.

Dell'Acqua, R., Sessa, P., Peressotti, F., Mulatti, C., Navarrete, E., & Grainger, J. (2010). ERP evidence for ultra-fast semantic processing in the picture–word interference paradigm. *Frontiers in psychology*, 1, 177.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193-222.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research, 1242,* 136–50. https://doi.org/10.1016/J.BRAINRES.2008.03.071

Eickhoff, S. B., Milham, M., & Vanderwal, T. (2020). Towards clinical applications of movie fMRI. Neuroimage, 116860.

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. *Electroencephalography and Clinical Neurophysiology, 99(3),* 225–234.

Eimer, M., & Kiss, M. (2008). Involuntary attentional capture is determined by task set: Evidence from event-related brain potentials. *Journal of cognitive neuroscience*, *20*(8), 1423-1433.

Eimer, M., Kiss, M., Press, C., & Sauter, D. (2009). The roles of feature-specific task set and bottom-up salience in attentional capture: An ERP study. *Journal of Experimental Psychology: Human Perception and Performance, 35(5),* 1316–1328.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*(5), 7-7.

Fleming, J. T., Noyce, A. L., & Shinn-Cunningham, B. G. (2020). Audio-visual spatial alignment improves integration in the presence of a competing audio-visual stimulus. *Neuropsychologia*, *146*, 107530.

Folk, C. L., Leber, A. B., & Egeth, H. E. (2002). Made you blink! Contingent attentional capture produces a spatial blink. *Perception & psychophysics*, *64*(5), 741-753.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18(4),* 1030–1044.

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in cognitive sciences*, 16(2), 129-135.

Girelli, M., & Luck, S. J. (1997). Are the same attentional mechanisms used to detect visual search targets defined by color, orientation, and motion? *Journal of Cognitive Neuroscience*, *9*(2), 238-253.

Green, J. J., & McDonald, J. J. (2010). The role of temporal predictability in the anticipatory biasing of sensory cortex during visuospatial shifts of attention. *Psychophysiology*, 47(6), 1057-1065.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14-23.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology, 28(2),* 240-244.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70.

Hopf, J.-M., Luck, S. J., Girelli, M., Mangun, G. R., Scheich, H., & Heinze, H.-J. (2000). Neural sources of focused attention in visual search. *Cerebral Cortex, 10*, 1233–1241.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*(3), 548-554.

Kiss, M., Jolicœur, P., Dell'Acqua, R., & Eimer, M. (2008). Attentional capture by visual singletons is mediated by top-down task set: New evidence from the N2pc component. *Psychophysiology*, *45*(6), 1013-1024.

Klaffehn, A. L., Baess, P., Kunde, W., & Pfister, R. (2019). Sensory attenuation prevails when controlling for temporal predictability of self-and externally generated tones. *Neuropsychologia*, *132*, 107145.

Koenig, T., Stein, M., Grieder, M., & Kottlow, M. (2014). A tutorial on data-driven methods for statistically assessing ERP topographies. *Brain topography*, *27*(1), 72-83.

Lamy, D., Leber, A., & Egeth, H. E. (2004). Effects of task relevance and stimulus-driven salience in feature-search mode. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(6), 1019.

Laurienti, P. J., Burdette, J. H., Maldjian, J. A., & Wallace, M. T. (2006). Enhanced multisensory integration in older adults. *Neurobiology of aging*, *27*(8), 1155-1163.

Lehmann, D., Skrandies, W. (1980). Reference-free identification of components of checkerboard evoked multichannel potential fields. *Electroencephalography in Clinical Neurology, 48*, 609–621.

Lien, M. C., Ruthruff, E., Goodin, Z., & Remington, R. W. (2008). Contingent attentional capture by top-down control settings: converging evidence from event-related potentials. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 509.

Luck, S. J., Gaspelin, N., Folk, C. L., Remington, R. W., & Theeuwes, J. (2020). Progress toward resolving the attentional capture debate. *Visual Cognition*, 1-21. DOI: 10.1080/13506285.2020.1848949

Luck, S. J., & Hillyard, S. A. (1994a). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology, 31,* 291–308.

Luck, S. J., & Hillyard, S. A. (1994b). Spatial filtering during visual search: Evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception and Performance, 20(5),* 1000–1014.

Lunn, J., Sjoblom, A., Ward, J., Soto-Faraco, S., & Forster, S. (2019). Multisensory enhancement of attention depends on whether you are already paying attention. *Cognition*, *187*, 38-49.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001-1010.

Matusz, P. J., & Eimer, M. (2013). Top-down control of audiovisual search by bimodal search templates. *Psychophysiology*, *50*(10), 996-1009.

Matusz, P. J., & Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychonomic bulletin & review*, *18*(5), 904.

Matusz, P. J., Wallace, M. T., & Murray, M. M. (2020). Multisensory contributions to object recognition and memory across the life span. In *Multisensory Perception* (pp. 135-154). Academic Press.

Matusz, P. J., Merkley, R., Faure, M., & Scerif, G. (2019a). Expert attention: Attentional allocation depends on the differential development of multisensory number representations. *Cognition*, *186*, 171-177.

Matusz, P. J., Turoman, N., Tivadar, R. I., Retsa, C., & Murray, M. M. (2019b). Brain and cognitive mechanisms of top–down attentional control in a multisensory world: Benefits of electrical neuroimaging. *Journal of cognitive neuroscience*, *31*(3), 412-430.

Matusz, P. J., Dikker, S., Huth, A. G., & Perrodin, C. (2019c). Are We Ready for Real-world Neuroscience?. *Journal of Cognitive Neuroscience*, *31*(3), 327.

Matusz, P. J., Retsa, C., & Murray, M. M. (2016). The context-contingent nature of cross-modal activations of the visual cortex. *Neuroimage*, *125*, 996-1004.

Matusz, P. J., Thelen, A., Amrein, S., Geiser, E., Anken, J., & Murray, M. M. (2015a). The role of auditory cortices in the retrieval of single-trial auditory–visual object memories. *European Journal of Neuroscience*, *41*(5), 699-708.

Matusz, P. J., Broadbent, H., Ferrari, J., Forrest, B., Merkley, R., & Scerif, G. (2015b). Multi-modal distraction: Insights from children's limited attention. *Cognition*, *136*, 156-165.

Michel, C. M., & Murray, M. M. (2012). Towards the utilization of EEG as a brain imaging tool. *Neuroimage*, *61*(2), 371-385.

Miniussi C, Wilding EL, Coull JT, Nobre AC. (1999). Orienting atten- tion in the time domain: modulation of potentials. *Brain, 122,* 1507-18.

Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP analyses: a step-by-step tutorial review. *Brain topography*, *20*(4), 249-264.

Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., & Matusz, P. J. (2016). The multisensory function of the human primary visual cortex. *Neuropsychologia*, *83*, 161-169.

Murray, M. M., Michel, C. M., De Peralta, R. G., Ortigue, S., Brunet, D., Andino, S. G., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *Neuroimage*, *21*(1), 125-135.

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science*, 13(5), 416-424.

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage*. In print.

Naumann, S., Byrne, M. L., de la Fuente, L. A., Harrewijn, A., Nugiel, T., Rosen, M. L., ... & Matusz, P. J. (2020). Assessing the degree of ecological validity of your study: Introducing the Ecological Validity Assessment (EVA) Tool. *PsyArXiv*. DOI: 10.31234/osf.io/qb9tz.

Noonan, M. P., Crittenden, B. M., Jensen, O., & Stokes, M. G. (2018). Selective inhibition of distracting input. *Behavioural brain research*, *355*, 36-47.

Peelen, M. V., & Kastner, S. (2014). Attention in the real world: toward understanding its neural basis. *Trends in cognitive sciences*, *18*(5), 242-250.

Perrin, F., Pernier, J., Bertnard, O., Giard, M. H., & Echallier, J. F. (1987). Mapping of scalp potentials by surface spline interpolation. *Electroencephalography and clinical neurophysiology*, *66*(1), 75-81.

Press, C., Kok, P., & Yon, D. (2020). The perceptual prediction paradox. *Trends in Cognitive Sciences*, *24*(1), 13-24.

Raij, T., Ahveninen, J., Lin, F. H., Witzel, T., Jääskeläinen, I. P., Letham, B., ... & Hämäläinen, M. (2010). Onset timing of cross-sensory activations and multisensory interactions in auditory and visual sensory cortices. *European Journal of Neuroscience*, 31(10), 1772-1782.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, *28*(2), 617-625.

Retsa, C., Matusz, P. J., Schnupp, J. W., & Murray, M. M. (2018). What's what in auditory cortices?. *NeuroImage*, *176*, 29-40.

Retsa, C., Matusz, P. J., Schnupp, J. W., & Murray, M. M. (2020). Selective attention to sound features mediates cross-modal activation of visual cortices. *Neuropsychologia*, *144*, 107498.

Rohenkohl, G., Gould, I. C., Pessoa, J., & Nobre, A. C. (2014). Combining spatial and temporal expectations to improve visual perception. *Journal of vision*, 14(4), 8-8.

Sawaki, R., & Luck, S. J. (2010). Capture versus suppression of attention by salient singletons: Electrophysiological evidence for an automatic attend-to-me signal., *Attention, Perception, & Psychophysics*, *72*(6), 1455-1470.

Shamay-Tsoory, S. G., & Mendelsohn, A. (2019). Real-life neuroscience: an ecological approach to brain and behavior research. *Perspectives on Psychological Science*, *14*(5), 841-859.

Soto-Faraco, S., Kvasova, D., Biau, E., Ikumi, N., Ruzzoli, M., Morís-Fernández, L., & Torralba, M. (2019). *Multisensory interactions in the real world*. Cambridge University Press.

Spierer, L., Manuel, A. L., Bueti, D., & Murray, M. M. (2013). Contributions of pitch and bandwidth to sound-induced enhancement of visual cortex excitability in humans. *Cortex*, *49*(10), 2728-2734.

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human perception and performance*, *38*(5), 1105.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in cognitive sciences*, *13*(9), 403-409.

Summerfield, J. J., Lepsien, J., Gitelman, D. R., Mesulam, M. M., & Nobre, A. C. (2006). Orienting attention based on long-term memory experience. *Neuron*, *49*(6), 905-916.

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience,* 17, 1098–1114.

Ten Oever, S., & Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *Proceedings of the National Academy of Sciences*, *112*(52), 15833-15837.

Ten Oever, S., Romei, V., van Atteveldt, N., Soto-Faraco, S., Murray, M. M., & Matusz, P. J. (2016). The COGs (context, object, and goals) in multisensory processing. *Experimental brain research*, 234(5), 1307-1323.

Theeuwes, J. (1991). Cross-dimensional perceptual selectivity. *Perception & Psychophysics, 50(2),* 184–193.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520-522.

Tivadar, R. I., & Murray, M. M. (2019). A primer on electroencephalography and event-related potentials for organizational neuroscience. *Organizational Research Methods*, *22*(1), 69-94.

Tovar, D. A., Murray, M. M., & Wallace, M. T. (2020). Selective enhancement of object representations through multisensory integration. *Journal of Neuroscience*. In press. DOI: https://doi.org/10.1523/JNEUROSCI.2139-19.2020

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12(1),* 97–136.

Turoman, N., Tivadar, R. I., Retsa, C., Maillard, A. M., Scerif, G., and Matusz, P. (2020a). The development of attentional control mechanisms in multisensory environments. *bioRxiv*. doi: https://doi.org/10.1101/2020.06.23.166975

Turoman, N., Tivadar, R. I., Retsa, C., Maillard, A. M., Scerif, G., and Matusz, P. (2020b). Uncovering the mechanisms of real-world attentional control over the course of primary education. *bioRxiv*. doi: https://doi.org/10.1101/2020.10.20.342758

Tzovara, A., Murray, M. M., Michel, C. M., & De Lucia, M. (2012). A tutorial review of electrical neuroimaging from group-average to single-trial event-related potentials. *Developmental neuropsychology*, *37*(6), 518-544.

Van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron*, *81*(6), 1240-1253.

van Atteveldt, N., van Kesteren, M. T. R., Braams, B.; Krabbendam, L. (2018). Neuroimaging of learning and development: improving ecological validity. *Frontline Learning Research,* 6 (3), 186–203. DOI: 10.14786/flr.v6i3.366.

Van der Burg, E., Talsma, D., Olivers, C. N. L., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage, 55,* 1208–1218.

van Moorselaar, D., & Slagter, H. A. (2019). Learning what is irrelevant or relevant: Expectations facilitate distractor inhibition and target facilitation through distinct neural mechanisms. *Journal of Neuroscience*, *39*(35), 6953-6967.

van Moorselaar, D., & Slagter, H. A. (2020a). Inhibition in selective attention. *Annals of the New York Academy of Sciences*, *1464*(1), 204.

van Moorselaar, D., Daneshtalab, N., & Slagter, H. (2020b). Neural mechanisms underlying distractor inhibition on the basis of feature and/or spatial expectations. bioRxiv.

Vanderwal, T., Eilbott, J.; Castellanos, F. X (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience, 36*, 100600.

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data–a practical approach. *Journal of Neuroscience Methods, 250*, 34-46.

Wu, R., Nako, R., Band, J., Pizzuto, J., Ghoreishi, Y., Scerif, G., & Aslin, R. (2015). Rapid attentional selection of non-native stimuli despite perceptual narrowing. *Journal of Cognitive Neuroscience, 27*(11), 2299-2307.

Zion-Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain and language*, *122*(3), 151-161.

Author contributions:

Nora Turoman: Investigation, Formal analysis, Data curation, Software, Visualisation, Writing - original draft, Writing - review & editing.

Ruxandra I. Tivadar: Software, Writing - review & editing.

Chrysa Retsa: Software, Writing - review & editing.

Pawel J. Matusz: Conceptualization, Funding acquisition, Methodology, Resources, Formal analysis, Software, Supervision, Writing - original draft, Writing - review & editing.

Micah M. Murray: Funding acquisition, Methodology, Resources, Formal analysis, Software, Supervision, Writing - review & editing.

Conflict of interest statement:

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements and data sharing statement

This project was supported by the Pierre Mercier Foundation to P.J.M. Financial support was likewise provided by the Swiss National Science Foundation (grants: 320030_149982 and 320030_169206 to M.M.M., PZ00P1_174150 to P.J.M., and the National Centre of Competence in research project "SYNAPSY, The Synaptic Bases of Mental Disease" [project 51AU40_125759]). P.J.M. and M.M.M. are both supported by Fondation Asile des Aveugles. Unfortunately, the current data set was acquired before both researchers and the broader public began to understand the scientific importance of data sharing. As we had not consented participants with an explicit data sharing statement, we cannot upload the current data set to an open data repository. Researchers interested in analysing the current data set are very welcome to contact the Corresponding Author (pawel.matusz@hevs.ch) for data sharing.

Figure legends

Figure 1. A) An example trial of the General experimental task is shown, with four successive arrays. The white circle around the target location (here the target is a blue diamond) and the corresponding distractor location serves to highlight, in this case, a non-matching distractor colour condition, with a concomitant sound, i.e., NCCAV. B) The order of tasks, with the corresponding conditions of Multisensory Relationship (MR) in red, and Distractor Onset (DO) in green, shown separately for each experiment. Predictable and unpredictable blocks before and after the training (1 & 2 and 3 & 4, respectively) were counterbalanced across participants. C) Events that were part of the Training. *Association phase*: an example pairing option (red – high pitch, blue – low pitch) with trial progression is shown. *Testing phase*: the pairing learnt in the Association phase would be tested using a colour word or a string of x's in the respective colour. Participants had to indicate whether the pairing was correct via a button press, after which feedback was given.

Figure 2. The violin plots show the attentional capture effects (spatial cueing in milliseconds) for TCC and NCC conditions, and the distributions of single-participant scores (vertical error areas) according to whether Multisensory Relationship within these distractors was Arbitrary (light green) or Congruent (dark green). The dark grey boxes within each violin plot show the interquartile range from the 1st to the 3rd quartile, and white dots in the middle of these boxes represent the median. Larger behavioural capture elicited by target-colour distractors (TCC) was found for arbitrary than semantically congruent distractors. Expectedly, regardless of Multisensory Relationship, attentional capture was larger for target-colour (TCC) distractors than for non-target colour distractors (NCC).

Figure 3. Overall contra- and ipsilateral ERP waveforms representing a mean amplitude over electrode clusters (plotted on the head model at the bottom of the figure in blue and black), separately for each of the four experimental conditions (Distractor Colour x Distractor Modality), averaged across all four adult experiments. The N2pc time-window of 180–300ms following distractor onset is highlighted in grey, and significant contra-ipsi differences are marked with an asterisk ($p < 0.05$). As expected, only the TCC distractors elicited statistically significant contra-ipsi differences.

Figure 4. Nonlateralised GFP and topography results for the visual only difference ERPs (DV condition of Target Difference), as a proxy for TAC. A) Mean GFP over the post-distractor and pre-target time-period across the 4 experimental tasks (as a function of the levels of MR and DO that they represent), as denoted by the colours on the legend. The time-windows of interest (102–124ms and 234–249ms) are highlighted by grey areas. B) Template maps over the post-distractor time-period as revealed by the segmentation (Maps A1 to A5) are shown in top panels. In lower panels are the results of the fitting procedure over the 29–126ms time-window. The results displayed here are the follow-up tests of the 3-way Map x MR x DO interaction as a function of MR (leftward panel) and of DO (rightward panel). Bars are coloured according to the template maps that they represent. Conditions are represented by full colour or patterns per the legend. Error bars represent standard errors of the mean.

Figure 5. Nonlateralised GFP and topography results for the difference ERPs between the D_{AV} and D_V conditions of Target Difference, as a proxy for MSE. **A)** Mean GFP over the post-distractor and pre-target time-period across the 4 experimental tasks (as a function of the levels of MR and DO that they represent), as denoted by the colours on the legend. The time-windows of interest (102–124ms and 234–249ms) are highlighted by grey bars. **B)** Template maps over the post-distractor time-period as revealed by the segmentation (Maps A1 to A7) are shown on top. Below are the results of the fitting procedure over the three time-windows: 35–110, 110–190, and 190–300 time-window. Here we display the follow-ups of the interactions observed in each time-window: in 35–110 and 190–300 time-windows, the 2-wau Map x MR interaction (leftward and rightward panels, respectively), and in the 110–190 time-window, follow-ups of the 3-way Map x MR x DO interaction as a function of MR and of DO (middle panel). Bars are coloured according to the template maps that they represent. Conditions are represented by full colour or patterns per the legend. Error bars represent standard errors of the mean.

Figure 1

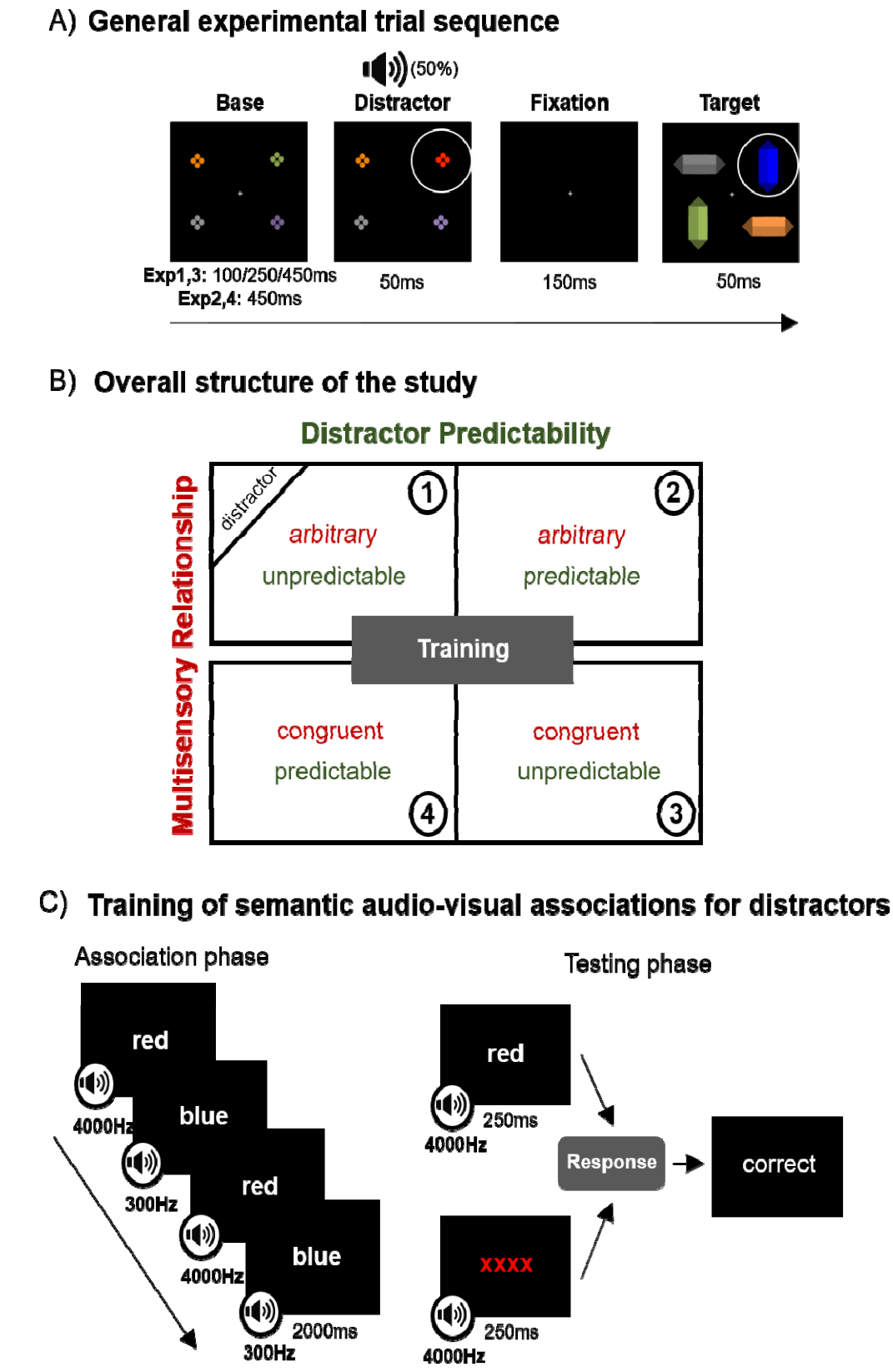

Figure 2.

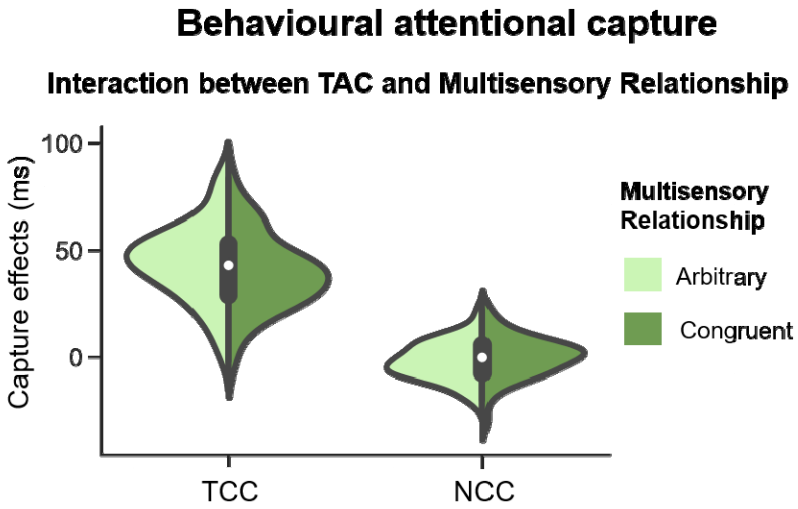

Figure 3.

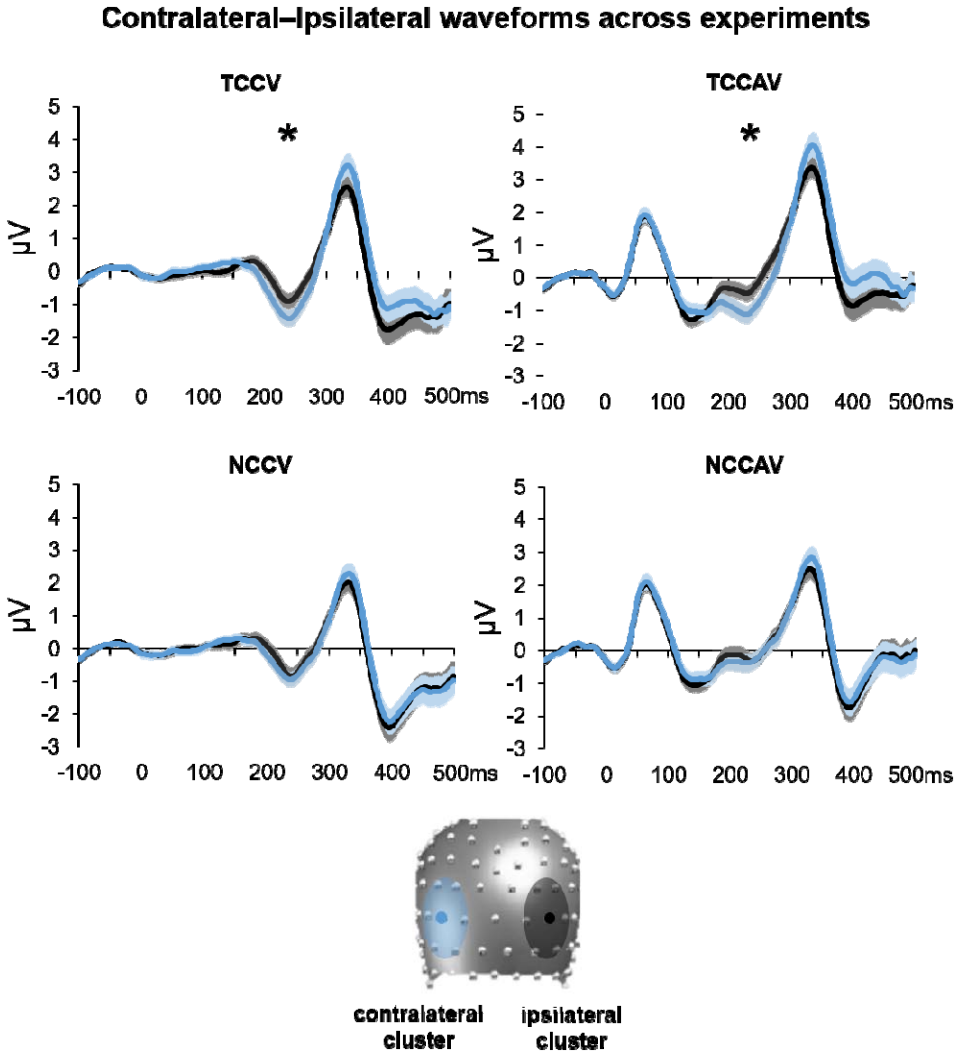

Figure 4.

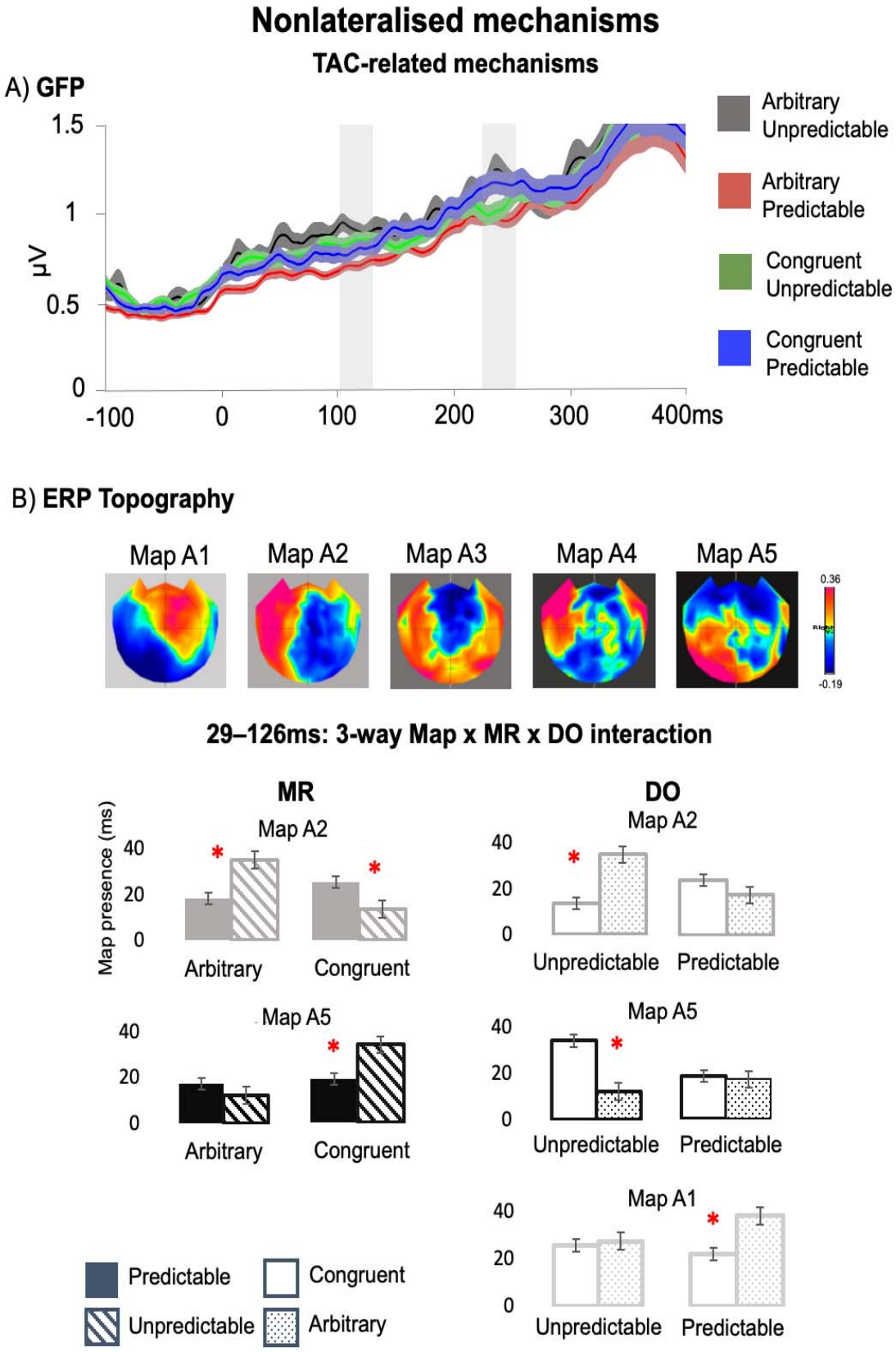

Figure 5.