# Got the gist? The effects of visually evoked expectations and cross-modal stimulation on the rapid processing of real-world scenes

Dominic Owen McLean

100145507

A thesis submitted in partial fulfilment of the requirements of the University of East Anglia for the degree of Doctor of Philosophy

Research undertaken in the School of Psychology, University of East Anglia

September 2021

## Abstract

Scene meaning is processed rapidly, with 'gist' extracted even when presentation duration spans a few dozen milliseconds. This has led some to suggest a primacy of bottom-up visual information. However, gist research has typically relied on showing successions of unrelated scene images, contrary to our everyday experience of a multisensory world unfolding around us in a predictable manner. To address this lack of ecological validity, Study 1 investigated whether top-down information – in the form of observers' predictions of an upcoming scene – facilitates gist processing. Participants (*N*=336) experienced a series of images, organised to represent an approach to a destination (e.g., walking down a sidewalk), followed by a final target scene either congruous or incongruous with the expected destination (e.g., a store interior or a bedroom). A series of behavioural experiments revealed that (i) appropriate expectations facilitated gist processing, (ii) inappropriate expectations interfered with gist processing, (iii) the effect of congruency was driven by provision of contextual information rather than the thematic coherence of approach images, and (iv) expectation-based facilitation was most apparent when destination duration was most curtailed. We then investigated the neural correlates of predictability on scene processing using ERP (*N*=26). Congruency-related differences were found in a putative scene-selective ERP component, related to integrating visual properties (P2), and in later components related to contextual integration including semantic and syntactic coherence (N400 and P600, respectively). Study 2 (*N*=206) then investigated the influence of simultaneous auditory information on gist processing, across two eye-tracking experiments. Search performance as a function of target sound congruency was measured using a flash-preview moving window paradigm. This revealed that a cross-modal effect did exist. Taken together, these results suggest that in real-world situations, both prior expectations and simultaneous cross-modal information influence the earliest stages of scene processing, affecting the integration of visual properties and meaning.

***Keywords:*** scene processing, gist, top-down information, event-related potentials, audio-visual processing, eye tracking

## Contents

**List of Figures**

## List of Tables

# Got the gist? Visually evoked expectations and cross-modal stimulation facilitate rapid processing of real-world scenes

## Preface

Despite their complexity and seemingly infinite variability, the human visual system can process the scenes we encounter with remarkable ease and efficiency. The astonishing speed with which detail and meaning can be extracted from a visual scene has now been appreciated for half a century, founded on the pioneering work of Mary Potter (e.g., 1975, 1976). Subsequent work has suggested that presentation durations of only ~13 ms are sufficient for forming an initial scene percept, potentially even including conceptual understanding (Oliva, 2013; Potter et al., 2014, but see Maguire & Howe, 2016). The mechanisms underlying the extraction of this 'gist' – the initial representation of a scene obtained from the briefest of glances – have become some of the most heavily researched concepts within the scene processing literature.

Much focus has subsequently been placed on unpicking the separate contributions of these mechanisms, with the endeavour to uncover which visual aspects of a scene are most diagnostic in terms of its categorisation. Such work has been incredibly fruitful, and we have gained much understanding as to how this initial gist of a scene is derived. For example, the global analysis of low-level features – such as statistics of local contrast (Scholte et al., 2009) and spectral features (Oliva & Torralba, 2001) – can reveal the spatial properties of a scene. Such global properties are processed rapidly, and so are an efficient method of bypassing more computationally demanding processes, and thus play a crucial role in facilitating gist extraction (Greene & Oliva, 2009; Groen et al., 2013). Furthermore, other properties of scene features have also been demonstrated as diagnostic, such as the position of contour junctions (Walther & Shen, 2014), colour (Oliva & Schyns, 2000; Goffaux et al., 2005), and object information (Davenport & Potter, 2004; Gagne & MacEvoy, 2014; Joubert et al., 2007).

However, this investigation of the speed and efficiency with which gist processing operates has brought with it debate. On the one hand, a scene can be categorised at a basic level in timeframes so rapid as to make it seem unlikely that top-down information would have the opportunity to contribute to processing in any substantial way. This has led to some taking a position which proposes that our initial understanding of a scene is from a 'forward sweep' of information through the processing stream (Fei-Fei et al., 2007; Potter et al., 2014; Thorpe et al., 1996), with little – if any – feedback from higher-order areas. Indeed, the speed with which scenes can be understood is so rapid that it has been proposed as an automatic mechanism able to operate outside of attention (Biederman, 1972; F. Li et al., 2002; Potter, 1975). On the other hand, the ability of an observer to interpret the gist of a scene is necessarily dependent on matching this information to stored representations of typically occurring patterns built through experience (Greene et al., 2014). Importantly, demonstrations of interference to gist extraction when scenes are atypical (Greene et al., 2015; Glanemann et al., 2016) suggest that this process cannot be exclusively stimulus driven. In other words, it appears that even the initial representation of gist is influenced by matching a stored template to current input (Greene et al., 2015) in a top-down manner.

Therefore, while studying the processing of isolated properties has elucidated many important sub-components of scene understanding, fundamental questions as to the role of top-down information in the extraction of gist – and the potential mechanisms underlying this process – remain unanswered. The work presented here aims to address this, through investigation as to whether gist can be influenced by factors outside immediate visual stimulation. Study 1, through behavioural and ERP measures, investigates whether an observer's expectations as to an upcoming scene affect the categorisation process. We find this to be the case, and suggest this provides evidence for top-down influence on both feature extraction and representation matching mechanisms. In Study 2, eye tracking methods are utilised to explore the effects of synchronous audio-visual information on gist processing. We find that gist can indeed be influenced by the congruency of this cross-modal stimulation. Based on these findings a case is made which suggests that gist processing cannot be fully described by a framework founded solely on the forward sweep of information through visual pathways.

In addition, scene research has for many years relied on the measurement of participant performance during the rapid serial visual presentation of unrelated scene images. This methodology has been highly informative in terms of processing speed thresholds, but is clearly far removed from the multisensory, predictable world we experience. As such, in appreciation of recent research demonstrating the divergence of findings from the lab to the outside world (Foulsham & Kingstone, 2017; Foulsham et al., 2011), a principal concern of our work has been to increase the ecological validity of its design. Therefore, while we have remained within the constraints of the laboratory walls, with the additional control that this affords, we have attempted to better approximate processing in daily life, where scenes are rarely (if ever) experienced as an isolated entity rather than an element within a flow of movement, or experienced exclusively through a single modality.

Taken together, therefore, the work contained herein builds upon a robust foundation of previous scene processing literature, but provides strong evidence to suggest that our current understanding of gist processing requires updating, and that future work must be fully appreciative of those factors outside immediate visual stimulation that influence how we experience scenes in the real world.

## STUDY 1

Apart from at waking, every environment we encounter is part of a progression of scenes unfolding around us as we move through our surroundings. As such, any single scene is not confronted in isolation but is instead simply the most recently perceived environment within the continuous experiential flow of our passage through the world. However, scene perception research has largely ignored such an asseveration, focusing more on the mechanisms responsible for processing segregated, individual scene images. In a traditional experiment, a participant may be faced with an image of a mountain, followed by a church, a kitchen, and so forth, a scenario clearly divergent from the progressive and structured environments one inhabits within the course of daily life.

We have, without question, learnt a great deal from investigation of the processing of isolated scene images, and such paradigms have been highly effective in identifying the mechanisms and visual features that facilitate processing of the initial meaning, or conceptual ‘gist’ (see Oliva, 2005), of a scene. Perhaps the most fundamental of findings is that this form of gist – the ability to derive the semantic information contained within a perceptual landscape – can be extracted even under conditions where viewing times span less than a tenth of a second (e.g., Potter, 1975). Such limited durations have led many to infer the primacy of bottom-up visual factors in rapid scene perception (Itti et al., 1998; Potter et al., 2014; Rumelhart, 1970), with the conviction that top-down information can have only a limited role under such brief time frames. This is a fair assessment if one contends that initial scene processing takes place in a classic hierarchical fashion. In such a scenario – of progressive activation through a linear pathway of anatomical areas divergent in terms of functional specificity – it is unlikely top-down feedback would be received prior to such rapid scene categorisation taking place. Therefore, while such models do not deny the role of feedback or re-entrant connectivity as processing continues through time, they propose feature-extraction mechanisms as sufficient for distinguishing conceptual information and meaning within complex natural scenes, with a single ‘forward sweep’ of neural activity through the ventral stream (Potter et al., 2014).

However, the traditional view of the serial processing of visual input has been questioned for some time (Engel et al., 2001; Ullman, 1995), and the latest recurrent models can better explain human visual recognition when compared to feedforward neural networks (e.g., Spoerer et al., 2020). Likewise, the past decade has seen great advances in our understanding of the broad extent of reciprocal connections within the neural architecture (e.g., Groen et al., 2017; Kravitz et al., 2013). For instance, research methodologies spanning MEG, EEG and TMS have all provided evidence for rapid local recurrent processes within early visual cortex (Boehler et al., 2008; Camprodon et al., 2010; de Graaf et al., 2014; Foxe & Simpson, 2002), with the proposal that these processes might start only a few tens of milliseconds after the arrival of the visual input (de Graaf et al., 2012). Furthermore, multiple feedforward-feedback loops have been hypothesised as taking place within the first 100 ms of stimulus onset (Bullier, 2001; Juan & Walsh, 2003), a proposition strengthened by the finding of activation in intermediate visual areas prior to the completed contribution of early visual cortex (Koivisto et al., 2011).

Similarly, from the object processing literature, evidence reveals that top-down processes are initiated prior to completion of target recognition, with the suggestion that early activation of higher-order brain regions facilitates the systematic analysis of bottom-up information (Bar et al., 2006). In other words, low spatial frequency information is passed rapidly to higher areas and is then used to form predictions as to the identity of the object being viewed. Consequently, this allows for the pre-activation of a limited set of object representations which are subsequently matched against the continuing flow of bottom-up information (Bar et al., 2006). It seems reasonable to infer that some equivalence may exist within the manner of operation for scene processing, whereby an initial ‘sketch’ (Marr, 1982; Rensink, 2000) of the environment may allow for the pre-activation of scene representations in higher-order areas. Indeed, parallel co-activity within higher regions has been observed even while perceptual coding of visual scenes is actively proceeding (Catherwood et al., 2014). While concurrent activation cannot be taken as direct evidence for interaction between regions, it provides the opportunity for such interactions to a far greater extent than models which assume somewhat step-by-step activation, whereby higher-order processing occurs only as perception subsides.

The above research shows that the selection and processing of those elements within even the precursory stages of the feed-forward wave of activity may be open to facilitation. Moreover, if top-down information can rapidly influence bottom-up processing in scenarios such as these – where no indication as to what will be displayed is provided prior to stimulus onset – then it seems appropriate to suggest that top-down influence might be even more rapid when pre-target cues allow for a subsequent visual image to be predicted. Such a claim is reinforced when considering the growing weight of evidence demonstrating that activity within the visual cortex, including early striate cortex, can be affected by expectations alone (e.g., Aitken et al., 2020; Grill-Spector & Malach, 2004; Kok et al., 2012), that the shape-selectivity of neurons in area V1 is altered depending on what geometric shape is expected (McManus et al., 2011), and that *a priori* expectations generated by scenic context can lead to increased activation in higher-order areas during subsequent visual processing (Caplette et al., 2020). Accordingly, here we present an investigation as to whether an observer’s expectations of an upcoming scene category have a direct effect on the initial stages of processing, i.e., the extraction of conceptual gist. In so doing, we attempt to better replicate how scenes are processed outside the laboratory, namely as predictable settings preceded by contextually relevant visual information, and hence proffer that models based on a progression of activation across successive regions cannot provide an exhaustive account of functionality.

In concordance, considerable evidence signals that expectations can influence subsequent processing of the environment, such as the inadvertent bypassing of crucial but unexpected visual information (Mahon, 1981), ‘looked-but-failed-to-see’ traffic accidents (Langham et al., 2002), and increased task-related errors in situations inconsistent with expectations (e.g., Endsley & Garland, 2000). Relatedly, the influence of context-based expectations on cognitive processing has been widely investigated through the experimental manipulation of object-scene relationships. Such research has repeatedly shown that target objects are found more quickly (Biederman et al., 1973; Võ & Henderson, 2011), and with higher accuracy (Antes et al., 1981; Davenport & Potter, 2004; Underwood, 2005), when within 'appropriate' scenes (i.e., where the scene category and target object are semantically congruous). In addition, such context effects have been found not only during the simultaneous presentation of a scene and target object, but also when a scene image is presented prior to (Demiral et al., 2012; Ganis & Kutas, 2003; Võ & Wolfe, 2013), and independent of (Palmer, 1975), object presentation. Due to the speed with which objects can be detected and identified (e.g., Crouzet & Serre, 2011; Kirchner & Thorpe, 2006; Thorpe et al., 1996), these studies demonstrate that semantic information can rapidly influence visual processing, and also that increased processing ability related to congruency is evident even when natural scene images are used as a precursory means of inducing expectations. If scenes can provide semantic information capable of altering subsequent object processing, it would seem intuitive that such influence similarly extends to subsequent scene processing.

Indeed, experimental evidence has demonstrated that a scene can be primed by a preceding scene-image, termed the 'scene priming' effect, although this has largely concerned priming at the perceptual – rather than conceptual – level. Increased performance regarding spatial layout judgements have been elicited when target scenes are primed using an identical scene image (Sanocki, 2013) or with images of the target scene from different viewpoints (Sanocki & Epstein, 1997, although see Epstein et al., 2005), while image detection ability is improved if primed across scenes more closely matched in terms of spectral information (Caddigan et al., 2017). Furthermore, when primes and targets are adjacent segments of the same complete landscape – thus intrinsically different while being similar in general composition – biases to cortical responses, alongside improved feature detection performance, have been shown (Blondin & Lepage, 2005). However, the mechanisms behind such effects are open to debate, as much of this work is proposed to reflect the maintenance of scene layout information in memory (Oliva & Torralba, 2001) or simply the priming of low-level visual features (Brady et al., 2017; Shafer-Skelton & Brady, 2019; although see Sanocki, 2013 for a potential top-down explanation). The focus of the current study, on the other hand, is investigation of the effect of expectations on scene processing at the semantic level. It is, therefore, equivalent to conceptual (Tulving & Schacter, 1990) or semantic priming (Meyer & Schvaneveldt, 1971), and so more similar to research showing performance increases when a scene's category membership is presented in text prior to presentation of the target image (Reinitz et al., 1989).

Correspondingly, recent research has further suggested a potential influence of top-down factors over the limited duration of gist processing (Greene et al., 2015). Here, briefly presented atypical scenes – such as a boulder in the centre of a living room, or a pillow-fight in a town square – were found to be more difficult to both process and understand compared to frequently encountered scene types (e.g., a car in a driveway). This indicates that an observer's prior semantic knowledge can influence the rapid processing of complex natural scenes, even over highly curtailed presentation durations. However, the design of that study still involved the presentation of single, unrelated images on each trial, and so cannot apprise us of the interaction between immediately preceding information and predictability. So, while such research highlights the cost of violating the expectations held in long-term memory, it speaks less to the violation of expectations built upon the 'on-line' flow of information as it is received.

This gap in understanding needs addressing due to how we experience the world around us, where the daily sequential emergence of scenes takes place in a predictable fashion. This predictability is not only apparent for locations with which we are familiar, such as knowing what scene will greet us when turning the corner of a street travelled daily, but is also related to our expectations when in previously unencountered locations. When walking down an unfamiliar street, in an unfamiliar town, experience with similar environmental surroundings allows one to form predictions as to what awaits past the next corner. The sight of houses at the end of the street may fit within the expected sequential flow of situational contexts built over a lifetime of similar experiences, thereby allowing for efficient cognitive processing (Bartlett, 1995). The sight of a volcano, on the other hand, would most likely violate any such schema (e.g., Mandler & Ritchey, 1977), resulting in the allocation of greater cognitive resources in order to process such unexpected information (Barlow, 1961; Haque et al., 2020).

Recent research has started to address this directly, by pointing towards the influence of predictions on gist processing through the use of pre-target narrative sequences (Smith & Loschky, 2019). Here, the spatiotemporal coherence of image sequences depicting different routes (such as from an office to a parking lot) was manipulated. When image sequences were presented in narrative order, as opposed to when randomised, categorisation performance for – and predictability of – target scenes was significantly increased. While this work was concerned with the ordering of pre-target images, rather than their congruency with an upcoming target-scene, it reveals that expectations as to what scene may be encountered next can be informed by what has gone before and, moreover, that these expectations may have a functional role in terms of facilitating scene-gist processing. An explanation for the underlying mechanisms has been offered, whereby narrative sequences help construct a current event model, which then in turn influences the extraction of gist information (Smith & Loschky, 2019). As a consequence, an iterative process is created whereby ‘front end’ information extraction (such as that derived from attentional selection mechanisms) informs ‘back end’ model construction (initially stored in working memory), which in turn influences front end processes, and so forth (Loschky et al., 2019).

So, both directly and indirectly, previous work has indicated that observer expectations can affect scene gist processing. Of equal importance, such a suggestion does not seem unreasonable when considering the typical mechanisms of visual processing more broadly. While we exist within a world of seemingly limitless sources of sensory information the visual system is constrained by limited processing capacity, and so it has long been understood that increased efficiency can be derived through drawing on learned experience to aid our interaction with the environment (Chaumon et al., 2008; Fiser et al., 2016; Gregory, 1997; W. Li et al., 2004; Rock, 1997; Ullman, 1980; although see Gibson, 2014 for an account of 'direct' perception). With this in mind, for the visual system not to use expectations to facilitate scene gist processing would seem to contravene its typical mode of operation.

The emergence of predictive coding models provides a potential framework by which the generation of expectations as to upcoming visual stimulation might, in part, offset inherent signal transmission delays (Hogendoorn & Burkitt, 2018; Nijhawan & Wu, 2009; Rauss et al., 2011). While it is beyond the scope of the current study to make determinations as to the precise mechanisms involved in any top-down influence on gist processing, such models provide a viable solution. For example, any current perceptual environment may lead to predictions of the subsequent environment, resulting in the pre-activation of those internal representations. These expectations may subsequently influence early visual areas by adapting their processing of perceptual features, through adjustment of prediction error thresholds, based on the representations chosen as likely to fit the upcoming landscape (Rauss et al., 2011). As such, the neural signal pattern even at early stages of the processing stream might be a reflection of a perceptual landscape's congruence with predictions, above-and-beyond merely a reflection of the low-level information contained within (Mumford, 1992).

To tease apart the role of on-line expectations within processing, the current study investigated the influence of visual information received immediately prior to target-scene onset. Across all experiments we employed a fundamental change to the traditional methodologies, which either position targets within a rapid serial visual presentation (RSVP) sequence of unrelated images (e.g., Potter, 1975) or present only a single image per trial (e.g., Greene et al., 2015). This was achieved by providing contextual information through presentation of antecedent 'lead-up' images, allowing us to investigate the influence of just-prior experience on the understanding of scenes. These leading images provided a flow of movement through an environment and towards a scene, and so represented an approach to a destination. This is, we suggest, a more naturalistic means by which to generate predictions based on lifelong experience, and as a result is somewhat removed from research investigating the effect of predictions on perception using simplistic pre-target cues (e.g., Summerfield & Koechlin, 2008), or where predictability is manipulated by synthetic means such as the learning of arbitrary contingencies prior to task commencement (e.g., Hindy et al., 2016). A key aim of the current study, therefore, was to provide a more ecologically valid reflection of scene perception. While only an approximation of this can be achieved with a sedentary participant viewing static images on a monitor, careful construction of image-series was considered sufficient in affording an impression of progress through a landscape.

Then, by manipulating whether the target scene was congruous with these leading images, i.e., the 'approach-destination' congruency, we hoped to demonstrate whether there is indeed an influence of predictability on scene categorisation ability. In addition, across the separate behavioural experiments we manipulated the presentation duration of destination images, the spatiotemporal coherence of approach-image sequences, and the provision of pre-destination scenic context in order to more fully investigate the mechanisms underlying the effect of expectations on gist processing. Finally, we turned to electroencephalography to map changes in brain activity relating to the manipulation of approach-destination congruency, with the aim of identifying the forms of cognitive processing most readily affected by the violation of expectations.

## Experiment 1a

The ability to categorise scenes even under the briefest presentation durations has led many to argue that such rapidity of processing must take place largely outside the involvement of top-down influence (Itti et al., 1998; Potter et al., 2014; Rumelhart, 1970). On the other hand, more recent research has found that semantic information can influence scene processing within shorter timeframes than previously thought (Greene et al., 2015; Võ & Wolfe, 2013). However, this research has largely focused on the semantic congruity of objects within a scene, rather than congruity between scenes. We aimed to address this gap by presenting series of ‘approach’ images prior to ‘destination’ target scenes, while manipulating the congruency between the destination and its forerunners, in order to investigate whether semantic predictability of an upcoming scene influences processing.

Furthermore, in Experiment 1a we manipulated target presentation duration to investigate whether the influence of contextual information remained consistent across the different stages of scene processing. Specifically, models assuming primacy of bottom-up factors during gist processing would not expect differences in categorisation performance as a function of congruency at target durations below 100 ms. Under such models (e.g., Potter et al., 2014), the category of the lead-up scenes would be expected to have minimal influence during the gist processing of the subsequent target image. Conversely, if performance differences were found at such brief durations this would lend support to the proposition for top-down influences on gist processing.

We hypothesised that destination scenes preceded by congruous approach images would be more accurately categorised, compared to those with incongruous approaches. Additionally, we predicted that this benefit would be most apparent at briefer presentation durations. This was due to our expectation that, at shorter durations, the ability to extract visual information would be most curtailed whereas, at longer durations, enough visual information would be extracted and processed from destination scenes as to bring categorisation accuracy for all targets towards ceiling performance. Hence, we expected to see the biggest congruency-related differential in performance at the briefest target durations, as this would be the point of maximal benefit from providing participants with a congruous scenic context prior to destination onset.

**Design**

All experiments were programmed and presented using PsychoPy (www.psychopy.org) version 1.85.3, unless otherwise stated (Peirce & MacAskill, 2018). University laboratories were used for all testing across both studies, except for Experiment 3 which took place online. An experimental trial began with participants viewing a sequence of five leading images, organised to represent an approach to a location. These approach images were followed by a target scene, representing a destination, which required a categorisation judgement from six available choices. All series depicted travel on foot, in order to convey a sense of walking through an environment (see Appendix A for additional details relating to the construction of series).

Each participant experienced 120 trials, 75% of which had leading images congruous with the target scene. This ratio was chosen to ensure participants remained attentive to the leading images. Target scenes could be from one of 30 separate categories, split equally between interior and exterior sceneries (see Appendix B for a list of categories used). Indoor and outdoor scenes vary from one another on fundamental characteristics such as level of expansiveness and roughness of textures (e.g., Oliva & Torralba, 2001), and there are suggestions that categorisation performance might differ across these two superordinate categories (Fei-Fei et al., 2007). Therefore, we chose to include both types of environment to provide a more complete picture of gist processing within typically encountered locations. All categories were considered familiar (e.g., ‘bathroom’, ‘beach’, etc.). Further to this, we manipulated target duration as a between-subjects variable, in order to investigate potential changes over the time-course of gist processing. Targets could be presented for 33, 50, 100 or 250 ms (2, 3, 6 or 15 frames on a 60Hz monitor). See Figure 1 for a schematic of the experimental protocol.

**Participants**

An *a priori* power analysis (*G*Power*; Faul et al., 2009) suggested an estimated sample size of ~30 participants per Target Duration condition was required for medium sized effects. As Experiment 1a incorporated four Target Duration conditions, for this initial study we recruited 129 undergraduate psychology students through the University of East Anglia's research pool, who received course credits for participating ($M_{\text{age}}$ = 20.09, $SD_{\text{age}}$ = 3.68; 103 Females, 26 Males; 113 Right-handed, 15 Left-handed, 1 Ambidextrous). All experiments were approved by the ethics committee at the University of East Anglia's School of Psychology (approval code: 2017-0201-000743), and all participants provided written informed consent prior to taking part in the study.

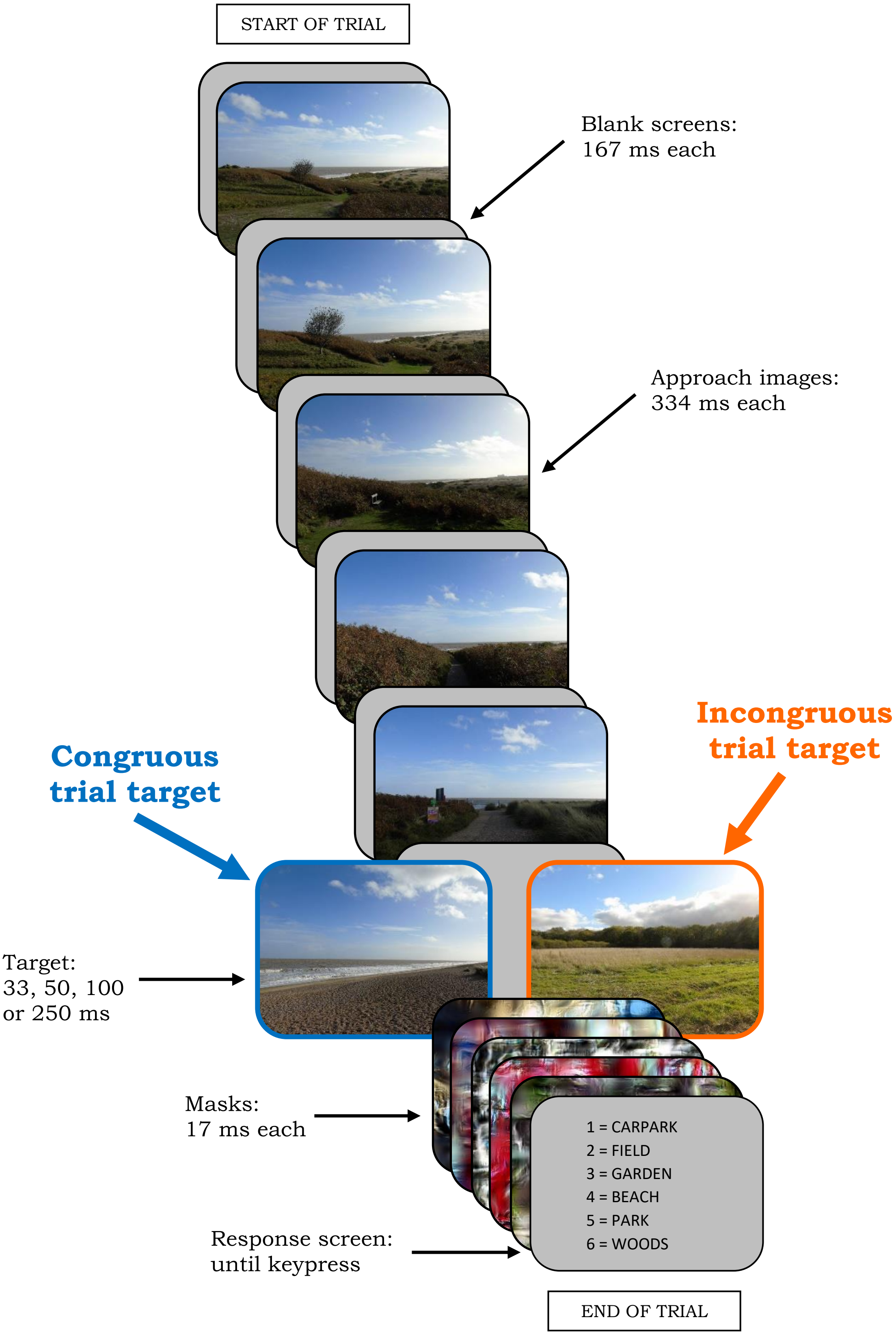
**Stimuli**

The collection of images was comprised of photographs taken by the researchers alongside high-definition images of sceneries and video-stills freely available on the internet. A total of 756 images were used as stimuli, of which 720 appeared in the experimental trials. No images were repeated. Each trial consisted of five spatiotemporally coherent approach images, followed by a target scene, resulting in 120 individual series. There were four series for each of the 30 scene categories. Approach images were sequential, first-person viewpoints heading towards a specific destination, with the aim of imbuing in participants a sense of progression through an environment.

One series from each of the scene categories was selected at random to become an Incongruous trial. The target scenes of each of these 30 series were then randomly reallocated amongst each other. This redistribution was conducted in adherence to two principles. Firstly, a target could not replace another target of the same scene category as, although it would be a different exemplar than what might be expected, it would still be semantically related to the approach images. Secondly, a target could only replace another target of the same superordinate category (in terms of interior / exterior distinction). This division was maintained due to suggestions that discriminating between superordinate categories is not analogous to discriminating between basic-level categories. While there is still debate as to the exact order with which these different levels are processed (see, for example, Banno & Saiki, 2015; Fei-Fei et al., 2007; Kadar & Ben-Shahar, 2012; Loschky & Larson, 2010), it was considered necessary to follow this principle to avoid potential changes in processing strategy from trial to trial. Each target image was followed by a set of five

**Figure 1**

*Schematic of the Protocol for Experiments 1a and 1b*



masks, presented rapidly in sequence. A different set of masks was used after each target. To achieve this, 600 masks were generated from the approach images by using Portilla and Simoncelli's (2000) texture synthesis algorithm in Matlab, in line with previous research showing this to be an effective method for placing temporal constraints on bottom-up processing of scene images (Evans et al., 2011; Greene et al., 2015).

Performance was judged through participants selecting the category that best described each destination scene from a list of six options. The available category options on each response screen were allocated randomly and were also randomised in terms of item position. All options were of the same superordinate category (indoor / outdoor) as the target. This was to ensure participants could not reject certain options based simply on superordinate-level membership. For Incongruous trials, the category the approach images would be expected to lead to was also included. In other words, an Incongruous trial displaying an approach to a 'Park' followed by a 'High Street' destination would have a response screen that included 'Park', 'High Street' and four other exterior scene categories. All images and masks were displayed with an image resolution of 800 x 600. All images were presented in colour, on a monitor with a refresh rate of 60Hz.

**Procedure**

Prior to starting the experiment, a series of instruction screens were displayed explaining the task and prompting participants to imagine travelling through the environments that were presented. This was followed by six practice trials, with the opportunity to ask any questions of the researcher on their completion. The same set of practice trials, in the same order, was experienced by each participant.

The 120 trials were presented in a different randomised order for each participant. Each trial included five sequential approach images, separated by blank screens, followed by a destination image. A series of five masks began at target-offset, prior to a 6AFC response screen. Once a response had been given, by pressing the number on the keyboard corresponding to the chosen category, the next trial began. At three equally spaced points within the task an ‘optional break’ screen was displayed, where participants could choose to pause if they wished and recommence once any key was pressed.

**Results and Discussion**

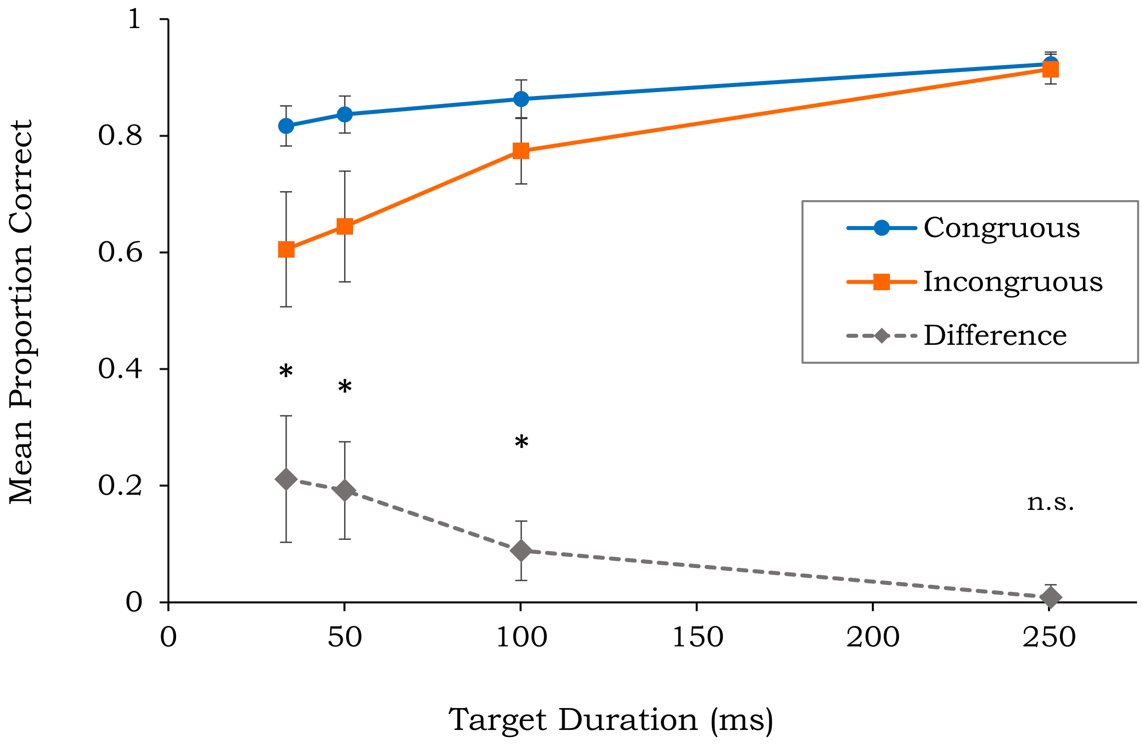
One participant was removed from the analysis due to a zero score for the Incongruous condition, suggestive of a misunderstanding of the task. A further five participants were removed due to a score in either congruency condition being outside 3 standard deviations of the mean for the respective target duration. Analysis was conducted on the remaining 123 participants ($M_{\text{age}}$ = 20.11, $SD_{\text{age}}$ = 3.76; 97 Females, 26 Males; 107 Right-handed, 15 Left-handed, 1 Ambidextrous). The proportion of correct answers on both congruency conditions was calculated for each participant, and the mean scores across participants were plotted (see Figure 2). As can be seen, when the opportunity to extract visual information was limited due to brief target durations (33 and 50 ms conditions), performance on Incongruous trials was discernibly below that on Congruous trials. As the target duration increased, however, this disparity across congruency conditions narrowed (100 ms), and subsequently disappeared (250 ms).

For each participant, accuracy scores for Incongruous trials were subtracted from scores for Congruous trials, to assess performance differences across congruency conditions. The values were first used to test the data for normality, which was found to be positively skewed at the 33, 50 and 100 ms target durations, and also to be leptokurtic at 100 ms. The non-normality for these three conditions was confirmed through Shapiro-Wilk tests (all *p*s < .005), and for this reason non-parametric alternatives were chosen for the analysis.

A Kruskal-Wallis test showed the differences in congruency-related scene categorisation accuracy to be significantly affected by the presentation duration of target sceneries, $H(3) = 17.46$, $p = .001$. Further to this, pairwise comparisons with Bonferroni-adjusted p values showed the disparity in performance relating to congruency differed significantly across target durations. The disparity in performance at a target duration of 33 ms was significantly different to that at 250 ms (*p* = .004, *r* = 0.44), and likewise at 50 ms compared to 250 ms (*p* = .001, *r* = 0.50). The difference between the 100 ms and 250 ms conditions only approached significance (*p* = .087, *r* = 0.32), and no other significant differences were found across durations.

**Figure 2**

*Scene Categorisation Accuracy for Congruous and Incongruous Trials as a Function of Target Duration*



*Note.* Error bars represent 95% CIs. * denotes *p* < .05. "n.s." denotes non-significance.

Follow-up Wilcoxon tests were then employed to investigate potential congruency-related differences in categorisation accuracy at each target duration. Difference scores (Congruous accuracy minus Incongruous accuracy) were compared to zero at each duration using a one-sample Wilcoxon test. Bonferroni adjusted p values are reported. At 33 ms, participants were significantly more accurate at categorising Congruous trials (*Mdn* = 0.83) than Incongruous trials (*Mdn* = 0.73), *Z* = -3.74, *p* < .001, *r* = -0.47, representing a medium-to-large effect. The same pattern was true at 50 ms, with greater accuracy for Congruous trials (*Mdn* = 0.86) than Incongruous trials (*Mdn* = 0.77), *Z* = -4.39, *p* < .001, *r* = -0.54, representing a large effect. This was again apparent at 100 ms, with greater accuracy for Congruous trials (*Mdn* = 0.89) than Incongruous trials (*Mdn* =0.80), *Z* = -3.40, *p* = .003, *r* = -0.43, representing a medium effect. However, no congruency-related differences were found at 250 ms (*p* = 1), with similar accuracy scores for both Congruous (*Mdn* = 0.94) and Incongruous trials (*Mdn* = 0.93).

As predicted, in Experiment 1a we found a significant benefit to categorisation performance when a target scene was preceded by semantically congruous approach images, revealing that participants' expectations were influencing scene processing. Furthermore, the greatest differential in performance across congruency conditions was seen at the briefest target durations (33 and 50 ms), indicative of gist extraction being modulated by top-down information. These findings sit in agreement with previous reports of expectations influencing subsequent processing of the environment (e.g., Langham et al., 2002), as well as research showing that object processing can be facilitated if situated within semantically compatible sceneries (e.g., Underwood, 2005). While facilitation of processing across scene-images has previously been observed in relation to the priming of visual features (Brady et al., 2017), we suggest that the benefit of congruency seen in Experiment 1a was due to the provision of semantically relevant context, and so more akin to the semantic priming of a scene when preceded by a relevant written word (Reinitz et al., 1989), or to work finding a disruption to gist processing when an observer views improbable sceneries (Greene et al., 2015). Therefore, these results revealed that top-down information – in the form of expectations generated prior to target scene appearance – was able to influence gist processing, a proposition at odds with models assuming minimal higher-order modulation of gist processing (e.g., Itti et al., 1998; Rumelhart, 1970).

## Experiment 1b

The results from Experiment 1a demonstrated an advantage in categorisation performance for Congruous trials, apparent at target durations where the opportunity to process visual information was most limited. However, it was important to confirm that these findings were due to the congruency manipulation as opposed to unintended residual effects based on the experimental design. Specifically, 75% of trials were congruous in Experiment 1a, and so higher performance on these trials was feasibly based on their increased frequency compared to Incongruous trials. We addressed this possibility in Experiment 1b, by switching the relative presentation frequencies of the congruency conditions.

The reduction in the number of Congruous trials in Experiment 1b also served a further purpose: in a task where most trials are incongruous it is not beneficial for participants to take account of the contexts provided by the approach images, as these are more often than not unrelated to the destination. If a pattern of results similar to those from Experiment 1a emerged, therefore, in terms of higher performance for Congruous compared to Incongruous trials, this would suggest that predictions as to an upcoming scene category were being generated automatically.

### Design

We again employed a 3:1 split across trial congruency, but now with 75% of trials having destinations incongruous to the approach images. The decision was also taken to limit Experiment 1b to three target-duration conditions. This was due to the preceding iteration showing very similar levels of performance, in terms of both congruency conditions, across the 33 and 50 ms target durations. There was also a noticeable amount of variation in performance across participants at 33 ms, with some failing to achieve scores above chance level. As a result, it was decided that the 50 ms condition provided the most reliable reflection of general performance under circumstances of limited availability of visual stimulation.

Although the selection of Incongruous trials in Experiment 1a had been achieved by random assignment, it was prudent to ensure this had not led to any bias through unintentional systematic differences across the two congruency conditions. As such, Experiment 1b introduced a Latin Square design. Four separate versions of the protocol were programmed, each with a different set of 30 Congruous trials (one from each scene category). This meant that, over the course of the experiment as a whole, all series were presented in both congruous and incongruous fashion, with the specific makeup of conditions determined by which version a participant sat. Versions were cycled through for each new participant, separated by target-duration condition.

**Participants**

Our second experiment included 90 undergraduate psychology students, recruited through the University of East Anglia's research pool, who received course credits for participating ($M_{\text{age}}$ = 20.89, $SD_{\text{age}}$ = 4.98; 68 Females, 22 Males; 81 Right-handed, 9 Left-handed). This sample size was in line with our previous power calculation, but was a lower number than the total required for Experiment 1a, as Experiment 1b only included three Target Duration conditions rather than the four of the previous iteration.

**Stimuli**

Experiment 1b used the same image set and masks as Experiment 1a. The response screens were redrawn, using the same randomisation procedures as the first experiment.

**Procedure**

The procedure for Experiment 1b mirrored that of 1a, with one alteration. A handful of participants had asked for clarification of certain category words during the previous iteration, most notably 'Quay'. To eliminate this issue, prior to beginning Experiment 1b participants were shown a list of the 30 scene categories and were provided with explanations by the researcher where needed. Participants were assured that the list did not need to be memorised.

**Results and Discussion**

Four participants were removed due to a score in either congruency condition being outside 3 standard deviations of the mean for the respective target duration. Analysis was conducted on the remaining 86 participants ($M_{\text{age}}$ = 20.98, $SD_{\text{age}}$ = 5.08; 65 Females, 21 Males; 77 Right-handed, 9 Left-handed). The proportion of correct answers on both congruency conditions was calculated for each participant, and the mean scores across participants were plotted (see Figure 3). As in Experiment 1a, when the opportunity to extract visual information was limited due to a brief target duration (50 ms), performance on Incongruous trials was some distance below that on Congruous trials. As target duration increased, the disparity across congruency conditions narrowed (100 and 250 ms).

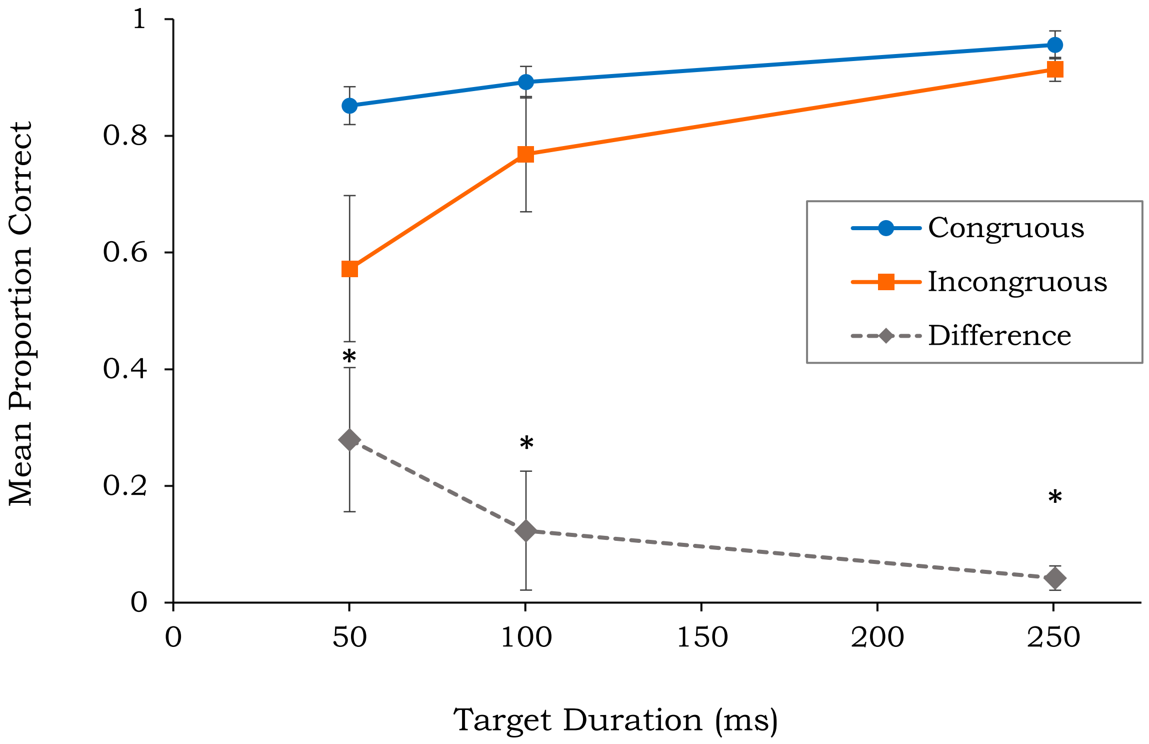
Accuracy scores for Incongruous trials were subtracted from scores for Congruous trials, to assess performance differences across congruency conditions. The values were first used to test the data for normality, which was found to be positively skewed for the 50 and 100 ms target durations, and to be leptokurtic for the 100 ms duration. The non-normality of these two conditions was confirmed through Shapiro-Wilk tests (all *p*s < .001), and for this reason non-parametric alternatives were chosen for the analysis.

A Kruskal-Wallis test showed the differences in congruency-related scene categorisation accuracy to be significantly affected by the presentation duration of target scenes, $H(2) = 13.40$, $p = .001$. Further to this, pairwise comparisons with Bonferroni-adjusted p values showed the disparity in performance differed significantly across target durations. The disparity in performance across congruency conditions at a target duration of 50 ms was significantly different to that at 100 ms ($p = .004$, $r = 0.42$), and at 250 ms ($p = .005$, $r = 0.41$). No significant difference was found between the performance disparity at 100 ms and 250 ms. Follow-up one-sample Wilcoxon tests were then used, comparing congruency-related differences in categorisation accuracy at each target duration to zero (representing no difference). Bonferroni adjusted p values are reported. At 50 ms, participants were significantly more accurate at categorising Congruous trials (*Mdn* = 0.87) than Incongruous trials (*Mdn* = 0.74), $Z = -4.17$, $p < .001$, $r = -0.55$, representing a large effect. The same pattern was true at

100 ms, with greater accuracy for Congruous trials (*Mdn* = 0.90) than Incongruous trials (*Mdn* = 0.87), *Z* = -2.68, *p* = .022, *r* = -0.35, representing a medium effect. This was also apparent at 250 ms, with greater accuracy for Congruous trials (*Mdn* = 0.97) than Incongruous trials (*Mdn* = 0.92), *Z* = -3.27, *p* = .003, *r* = -0.44. This again represents a medium effect.

**Figure 3**

*Scene Categorisation Accuracy for Congruous and Incongruous Trials as a Function of Target Duration*



*Note.* Error bars represent 95% CIs. * denotes *p* < .05.

As predicted, the results from Experiment 1b mirrored those from Experiment 1a. Again, categorisation ability was significantly higher when target scenes were preceded by congruous approach images, and this differential in performance was greatest when target presentation duration was at its most brief (50 ms). Consequently, Experiment 1b confirmed our findings were due to the congruency manipulation, as opposed to simply being based on the presentation frequency of experimental trials. In addition, these results show that context-based predictions were being generated automatically by participants as they viewed the approach images, in line with work demonstrating that pre-target natural scene images lead to the automatic generation of expectations as to the identity of an upcoming target object (Caplette et al., 2020).

Taken together, the findings from across these two experiments revealed that approach images influenced subsequent scene processing, and so suggest a role for top-down information in rapid gist processing. They do not support, therefore, narratives which propose the extraction of scene-gist is exclusively based on feedforward processes.

## Experiment 2

While Experiments 1a-b found an influence of trial congruity, it remained to be determined the specific mechanisms responsible for such an advantage. Divergent explanations as to the mechanisms underlying the findings of the previous experiments are possible. On one hand, the presentation order of approach images may have comparatively little bearing on performance, whereby these images simply serve to provide a semantic context which increases the predictability of the subsequent destination. For instance, observing an approach image which depicts surroundings commonly associated with the countryside may be sufficient for expectations to be formed as to the most likely eventual destination (e.g., a field, woods, etc.). In this scenario there would be no cost to performance if approach images were not arranged in a meaningful sequence, as participants would still be provided with the same contextual information prior to target presentation. On the other hand, there may be an additional benefit, above-and-beyond that based on semantic context, from the spatiotemporally progressive nature of the series. If this were true, then we would expect to see lower performance on trials where there was disruption to the ordering of images within a sequence.

An advantage of approach-image sequentiality, if apparent, could be due to several factors. For example, the importance of narrative coherence for efficient processing has previously been demonstrated (Cohn et al., 2012; Foulsham et al., 2016; Smith & Loschky, 2019). Through disruption to the order and content of comic strips, Cohn and colleagues have investigated the individual contributions of the semantic relationship between images and their overall narrative structure on the processing of image sequences (2012). It was found that both semantic relatedness and narrative structure were advantageous, whereby the processing of a subsequent image was influenced by both the structure and meaning of the series that preceded it. Alternatively, a case could be made that sequentiality allows for the generation of a ‘perceived flow’ of movement through the environment, potentially facilitating processing by allowing for the extraction of more information, such as that derived from the semblance of optical flow (Gibson, 1966) or through aiding the transformation of the viewer-centred 2½D sketch into a three-dimensional representation (Marr, 1982). Finally, the further away in space a leading image is from its eventual destination, the potentially weaker its predictive power. As an observer progresses through a series, each new leading image may further ‘fine-tune’ expectations, which could be a more additive process compared to that occurring from experiencing the same images in random order.

Hence, to investigate whether the sequentiality of series plays a role in gist processing, in Experiment 2 we manipulated the presentation order of approach images while also continuing to manipulate congruency. We predicted a categorisation advantage for sequentially coherent trials, as compared to disordered trials. This was due to the assumption that sequentiality would create a flow of information that more closely mirrored typical functioning in everyday environments, and due to research identifying an important role of narrative sequences for processing (e.g., Cohn et al., 2012; Foulsham et al., 2016; Smith & Loschky, 2019). Additionally, owing to the provision of semantically relevant context, we predicted that performance on Congruous trials, regardless of sequentiality, would still exceed that on Incongruous trials.

## Design

While maintaining the approach-destination congruency manipulation of Experiments 1a-b, Experiment 2 departed from the previous iterations by also manipulating the sequentiality of approach images. Therefore, trials included approach images displayed either in a sequential or randomised order. This led to four within-participant conditions: Congruous-Sequential; Congruous-Disordered; Incongruous-Sequential; and Incongruous-Disordered. Each condition consisted of 30 trials and included one series for each of the scene categories. A Latin Square design was employed, so that each series alternated across all conditions within the four versions of the experiment. For each version, the destination images for those series selected to constitute Incongruous trials were randomly reallocated amongst each other, following the same principles as previous iterations. Similarly, the presentation order of approach images within Disordered trials was randomly selected, but with two important constraints. First, the approach image in the closest geographical location to the destination scene could not be the final pre-target image in a Disordered trial. This parameter was to ensure that congruous targets were not simply being primed by the final approach image in isolation. Secondly, such trials could not contain more than two approach images displayed in their original order. This was to safeguard the non-sequentiality of Disordered trials.

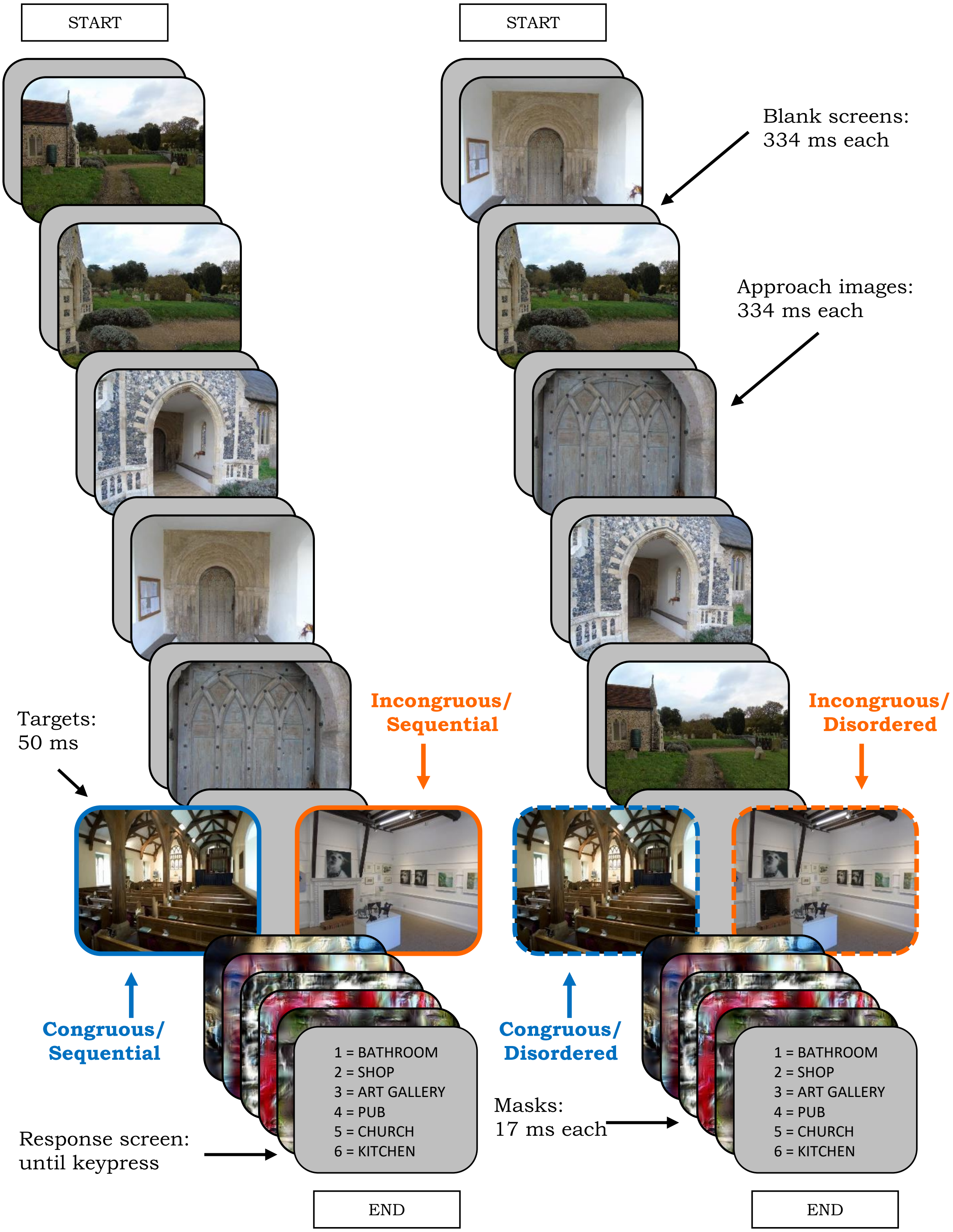
The presentation order of trials was randomised independently for each participant. Target duration was not manipulated in Experiment 2. Targets were presented for 50 ms, due to the findings from the previous experiments. This was based on the demonstration that the effect of congruity was most apparent at brief target durations, diminishing as presentation length increased. See Figure 4 for a schematic of the experimental protocol.

## Participants

Due to the removal of the target duration manipulation, and in line with our initial power calculation, thirty-six participants were originally included in Experiment 2. This comprised students and staff of the university, receiving either course credits or a small payment for taking part.

**Figure 4**

*Schematic of the Protocol for Experiment 2*



However, analysis of this initial data showed a much smaller effect of approach-image sequentiality compared to the size of effect related to trial congruency. To better determine the veracity of this effect, we took the decision to double the sample size while halving the alpha level during the subsequent analysis ($\alpha$ = .025). This technique is considered appropriate for controlling the Type 1 error rate in situations where a sample is increased due to the size of the observed effect (Lakens, 2014). The complete sample, therefore, consisted of 72 participants ($M_{\text{age}}$ = 23.58, $SD_{\text{age}}$ = 11.22; 54 Females, 18 Males; 61 Right-handed, 10 Left-handed, 1 Ambidextrous).

**Stimuli**

Experiment 2 used the same image set and masks as Experiment 1a and 1b. The response screens were again redrawn, using the same randomisation procedures as the preceding experiments.

**Procedure**

The procedure followed the same routine as Experiment 1b, except that the display duration of blank screens was increased from 167 ms to 334 ms (10 to 20 frames on a 60Hz monitor). This was judged to provide a more comfortable viewing experience for participants, which better mimicked the sense of traversing an environment.

**Results and Discussion**

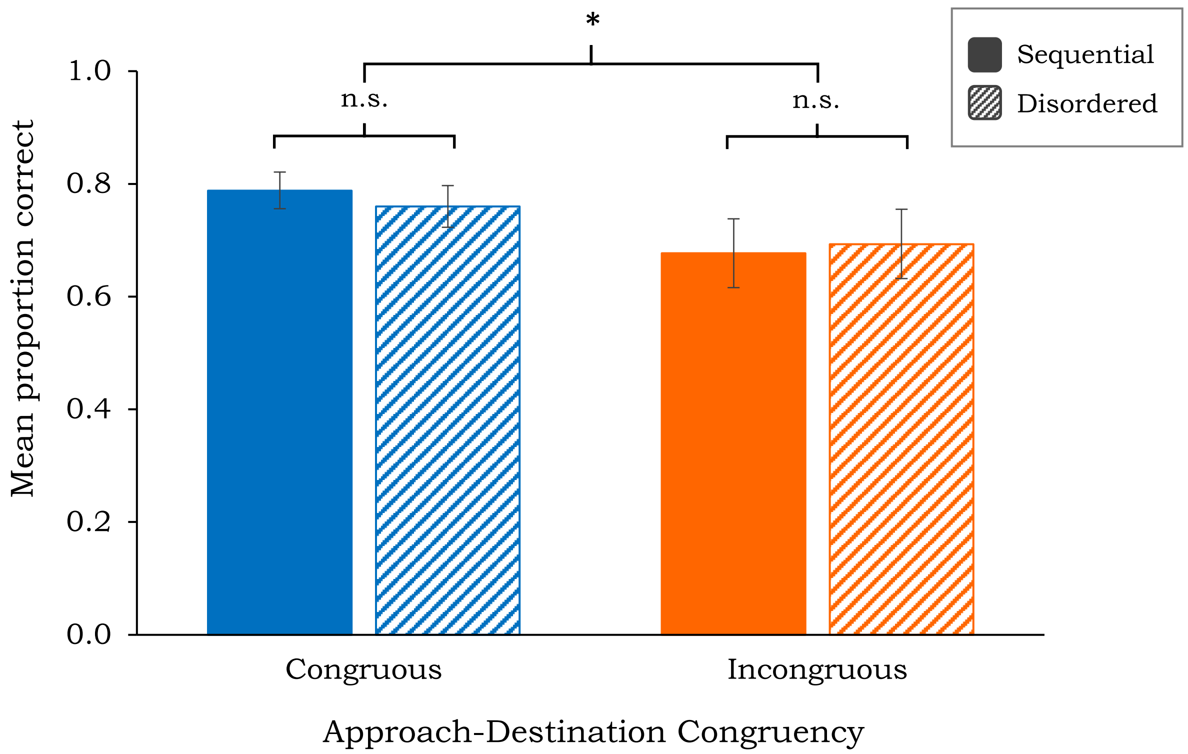
One participant was removed due to a score outside three standard deviations of the mean in one of the experimental conditions. Analysis was conducted on the remaining 71 participants ($M_{\text{age}}$ = 23.62, $SD_{\text{age}}$ = 11.29; 54 Females, 17 Males; 60 Right-handed, 10 Left-handed, 1 Ambidextrous). For each participant, accuracy scores for Congruous-Disordered trials were subtracted from scores for Congruous-Sequential trials, and Incongruous-Disordered scores were subtracted from Incongruous-Sequential scores, to assess performance differences across Congruency and Sequentiality conditions. Additionally, differences across Incongruous trials were subtracted from the differences across Congruous trials. These three sets of values were used to test the data for normality, displaying no issues relating to skewness or kurtosis, as confirmed through Shapiro-Wilk tests (all *p*s > .3). On account of this, a 2 (Congruency: Congruous; Incongruous) x 2 (Sequentiality: Sequential; Disordered) repeated measures ANOVA with planned comparisons was chosen for the analysis.

There was a main effect of Congruency, *F*(1, 70) = 9.74, *p* = .003, $\eta p^2$ = .12, with significantly higher performance on Congruous (*M* = 0.77, *SE* = 0.02) than Incongruous (*M* = 0.69, *SE* = 0.03) trials. This represents a medium effect size. There was no main effect of Sequentiality, *F*(1, 70) = 0.32, *p* = .574, $\eta p^2$ = .01. There was, however, a significant Congruency X Sequentiality interaction, *F*(1, 70) = 7.00, *p* = .010, $\eta p^2$ = .09, also representing a medium effect size. This indicated that the sequentiality of approach images had different effects on categorisation performance depending on whether the approach images were congruous with the target image (see Figure 5). To investigate this interaction further, paired samples t-tests were conducted. On average, when approaches were congruous with destinations, participants performed better if the approach images were presented in sequential (*M* = 0.79, *SE* = 0.02) compared to random order (*M* = 0.76, *SE* = 0.02). This difference, 0.03, 95% CI [0.003, 0.05], was not significant at the 0.025 alpha level, *t*(70) = 2.24, *p* = .028, *r* = .26. This represents a small-to-medium effect. It should be noted, though, that the confidence interval did not bridge zero which can be taken as support for the existence of such an effect. When approaches were incongruous with destinations, participants scored slightly higher if approach images were presented in random order (*M* = 0.69, *SE* = 0.03) compared to sequentially (*M* = 0.68, *SE* = 0.03). However, this difference, -0.02, 95% CI [-0.05, 0.01] did not reflect a significant difference in performance, *t*(70) = -1.16, *p* = .249, *r* = .14.

As with the previous experiments, we again found a benefit to categorisation performance related to trial congruency. However, the predicted effect of approach-image sequentiality did not reach statistical significance. As a consequence, it appears that simply providing observers with an appropriate contextual setting allowed for sufficiently accurate predictions of the upcoming destination to be formed. For instance, approach images displaying movement along a pavement, even if out of sequence, still allowed for expectations to be generated (i.e., previous experience would teach us that pavements lead to, say, high streets much more frequently than to woods).

**Figure 5**

*Scene Categorisation Accuracy for Sequential and Disordered Trials as a Function of Approach-Destination Congruency*



*Note.* Error bars represent 95% CIs. * denotes *p* < .05. "n.s." denotes non-significance.

That there was no significant additional benefit to gist processing when trials depicted a continuous spatiotemporal journey suggests that the generation of a perceived 'flow of movement' (e.g., Gibson, 1966) had little bearing on the accuracy of the expectations constructed by participants. This finding was also surprising due to research demonstrating the importance of narrative coherence within pictorial sequences (Cohn et al., 2012; Foulsham et al., 2016; Smith &

Loschky, 2019). However, that previous work incorporated narratives more complex in nature than our short approaches to proximal destinations. It stands to reason that the disruption to processing caused by the disarrangement of chronological elements would be a function of the complexity of the story being told, whether that be within the panels of a comic strip or the pictorial representation of an extended journey.

## Experiment 3

The findings from Experiment 2 suggested that the influence of approach images on subsequent scene processing was primarily due to participants being provided a semantic context prior to target onset. Up to this point the assumption had been made that this effect was driven by congruous approaches facilitating the subsequent processing of destinations. However, a possible alternative explanation remained, namely that the difference across experimental conditions was the result of incongruous approaches interfering with the processing of destinations.

To answer this question, in Experiment 3 we introduced a third, neutral condition whereby approach images were replaced by images of coloured patterns. As such, provision of semantic context was absent within the trials of this condition, meaning participants were unable to generate expectations as to the identity of the upcoming destination. As this condition maintained the trial format of other conditions, namely the display of series of pre-target images, it was considered a suitable reflection of baseline performance. Therefore, by comparing categorisation performance for this condition to performance on Congruous and Incongruous trials, respectively, we hoped to uncover more fully the role of the congruency manipulation on gist processing. We expected better performance on Congruous trials, compared to No-context and Incongruous trials; due to a lack of direct evidence from previous research, we made no predictions as to whether performance on Incongruous trials would be significantly lower than that of the No-context condition.

### Design

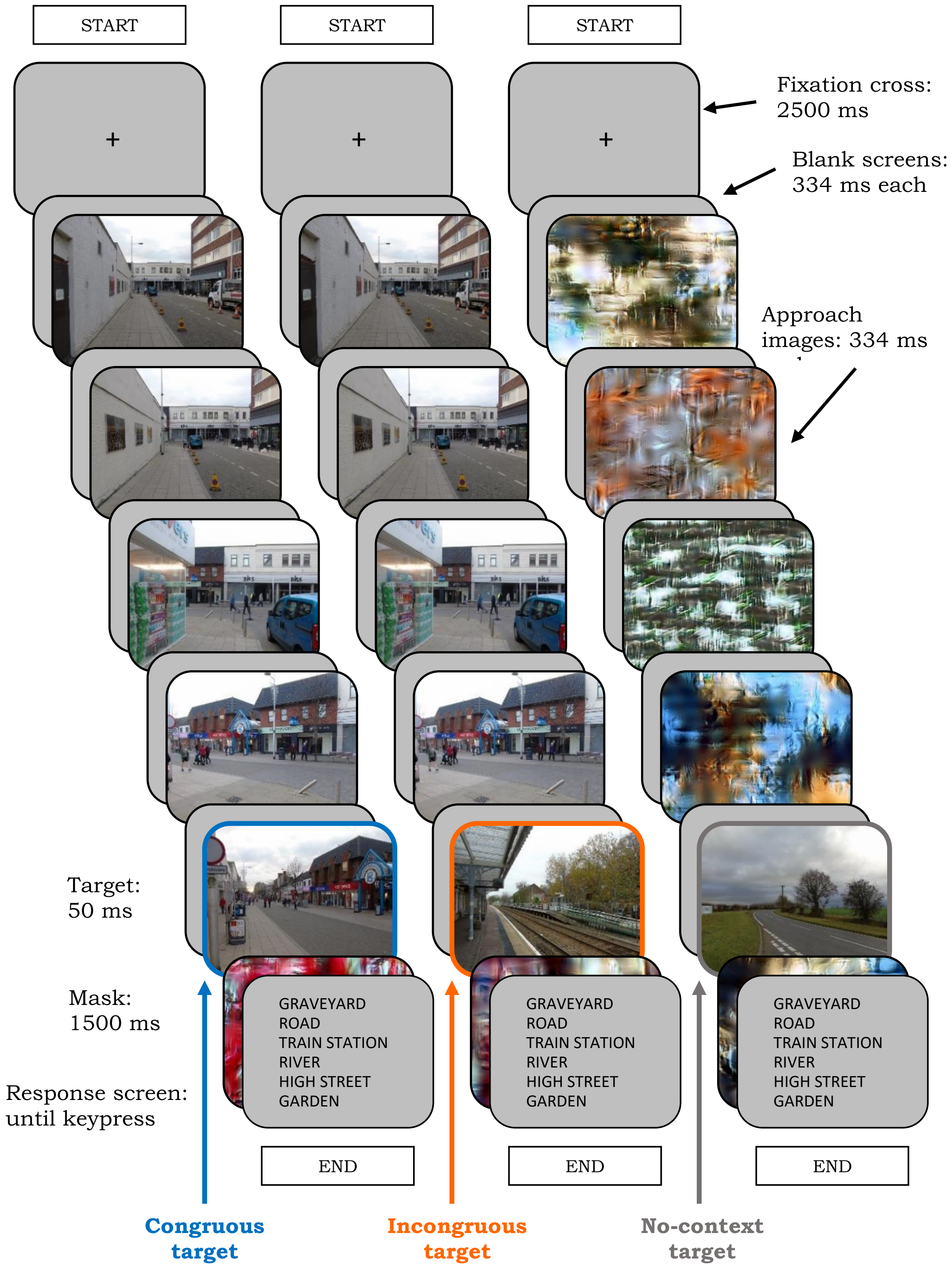
Experiment 3 maintained the approach-destination congruency manipulation of previous experiments but did not include the manipulation of approach sequentiality seen in Experiment 2. Alongside the previous congruency conditions, a third condition was added in which the approach images provided no semantic context to participants prior to destination-onset. This led to three within-participant conditions: Congruous; Incongruous; and No-context. Each condition consisted of 40 trials, including at least one, and no more than two, series for each of the scene categories. A Latin Square design was employed, so that each series alternated across all conditions within the three versions of the experiment. For each version, the destination images for those series selected to constitute Incongruous trials were randomly reallocated amongst each other, in line with the principles of previous iterations. The presentation order of trials was randomised independently for each participant. Target duration was again not manipulated in Experiment 3, as the presentation of targets was set at 50 ms for all participants. See Figure 6 for a schematic of the experimental protocol.

**Participants**

Participants were recruited through the online international participant pool, Prolific (www.prolific.co), and received a small payment for taking part. Demographic screeners were used to ensure all participants were adults who lived in the UK, US, Canada, Australia or New Zealand, and were fluent speakers of English. This filtering was to ensure that all participants would both be able to fully understand the task instructions and would be familiar with the types of sceneries used in the experiment. Although our original power calculation suggested that a sample size of 30 was sufficient for investigating medium-sized effects, we decided to increase this by 50% for Experiment 3. This was due to it being the first online experiment we had conducted, and so we had some concerns whether the level of engagement shown by some of those participating remotely might be too low to be included in the analysis. Therefore, 45 participants took part in Experiment 3 ($M_{\text{age}}$ = 33.53, $SD_{\text{age}}$ = 11.62; 30 Females, 15 Males; 37 Right-handed, 7 Left-handed, 1 Ambidextrous).

**Figure 6**

*Schematic of the Protocol for Experiment 3*



**Stimuli**

Experiment 3 used the same image set and masks as the previous experiments, although in this iteration some of the mask-images were repurposed to act as leading images in the No-context condition (as set out below). The response screens were again reconfigured, using the same randomisation procedures as before.

**Procedure**

The procedure followed a similar routine as Experiment 2, with some minor alterations necessary for the experiment to be run online. The experiment was programmed using Testable (www.Testable.org) and, due to constraints imposed by the software, the number of images displayed per trial needed to be reduced. This was achieved in two ways. Firstly, only four approach images were presented per trial, as we removed the first approach image from each series (i.e., the image most geographically distant from the destination). Secondly, destination images were followed by a single mask rather than a set of five dynamic masks. The duration of these individual masks was extended to 1500 ms to ensure a suitable disruption to processing from target-offset was maintained. The previous experiments each used 600 mask-images, and so 120 of these were randomly selected to again be used as masks in Experiment 3. A further 160 were then randomly selected to serve as leading images in the No-context condition. The order of presentation of these images, both within series and across trials, was also randomised. These randomisation procedures were followed for each of the three Latin Square versions, with the proviso that a mask-image could not be used as a leading image and a mask within a single version, and that all 600 mask-images were used across the experiment as a whole.

Two further minor alterations were included to ensure the smooth running of the experiment, due to the inevitable reduction in researcher oversight during an online study. Firstly, a fixation cross was displayed in the centre of the screen prior to the start of each series. Secondly, selection of a response was made by navigating a cursor to the chosen textbox, rather than by pressing a number on a keypad.

**Results and Discussion**

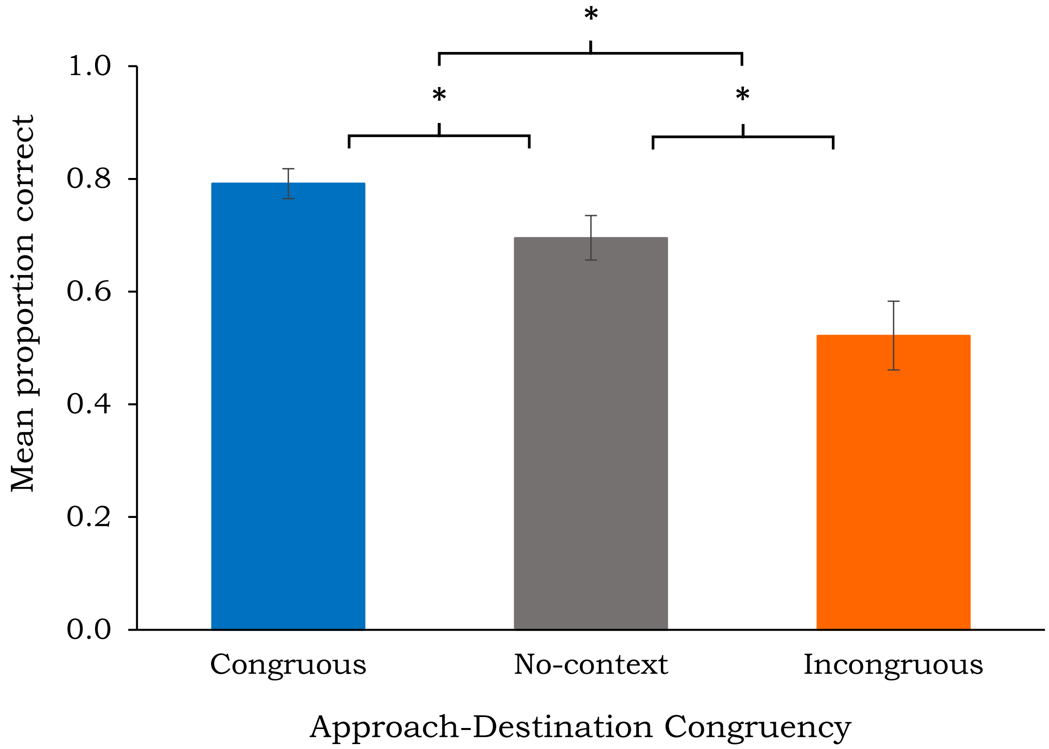
All participants scored within three standard deviations of the mean for each of the experimental conditions, and so all 45 were included in the analysis. For each participant, the difference in accuracy scores for the Congruous compared to the No-context condition, and the No-context compared to the Incongruous condition, were calculated. These values revealed the data to be normally distributed, displaying no issues with skewness or kurtosis, as confirmed through Shapiro-Wilk tests (all *p*s > .3). Consequently, a one-way (Congruency: Congruous; Incongruous; No-context) repeated measures ANOVA with planned comparisons was chosen for the analysis.

Mauchly's test indicated that the assumption of sphericity had been violated, $X^2(2) = 16.00$, $p < .001$, and so degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (Greenhouse & Geisser, 1959). There was a significant effect of the type of approach image on categorisation performance, $F(1.53, 67.18) = 78.37$, $p < .001$, $\eta p^2 = .64$. This represents a large effect. As can be seen in Figure 7, the proportion of correct responses was greatest in the Congruous condition (*M* = 0.79, *SE* = 0.01), followed by the No-context condition (*M* = 0.70, *SE* = 0.02), and with weakest performance in the Incongruous condition (*M* = 0.52, *SE* = 0.03). Follow-up paired samples t-tests, with Bonferroni-adjusted p values, revealed that the mean performance difference between Congruous and No-Context trials, 0.10, 95% CI [0.07, 0.13], was significant $t(44) = 6.21$, $p < .001$, $r = .68$, as was the difference between Congruous and Incongruous trials, 0.27, 95% CI [0.22, 0.32], $t(44) = 10.19$, $p < .001$, $r = .84$. Both represent large effects. Furthermore, the mean performance difference between No-context and Incongruous trials, 0.17, 95% CI [0.13, 0.22], was also found to be significant $t(44) = 7.84$, $p < .001$, $r = .76$, again representing a large effect.

A clear pattern of results emerged in Experiment 3, with significantly contrasting levels of categorisation performance seen across each of the three conditions. As with previous iterations, the ability to categorise destination scenes was greater when preceded by semantically congruous, rather than incongruous, approaches. Further to this, increased performance was also apparent on Congruous trials as compared to those where pre-target context was absent, thus confirming our contention that semantic congruity leads to a facilitation of gist processing.

**Figure 7**

*Scene Categorisation Accuracy for Congruous, Incongruous and No-context Trials*



*Note.* Error bars represent 95% CIs. * denotes *p* < .05.

In addition, we found that performance on Incongruous trials was significantly below that of the No-context condition, revealing that participants' ability to categorise a destination scene was inhibited if preceded by an unrelated scenic context. This appears to be in agreement with previous research showing scenes containing unexpected features are more difficult to extract meaning from (Greene et al., 2015), as well as recent work demonstrating interference to object recognition as a result of contextual violations within object-scene pairs (Lauer et al., 2020). In sum, this pattern of results clearly shows that approach images were eliciting expectations as to the likely identity of an upcoming target scene, resulting in a benefit to gist processing if expectations were realised but, alternatively, resulting in a cost to processing if violated.

## Experiment 4

The initial set of behavioural experiments revealed that providing semantic information leads to increased performance on subsequent scene processing. Building on these results, we turned to an investigation of the neural signature. To that end, in Experiment 4 we investigated the event-related potential (ERP) correlates of scene processing and the role of expectations on gist extraction. This investigation was exploratory in nature as, to date, we are unaware of any prior use of this methodology for the examination of the role of sequential, naturalistic leading images on subsequent scene-gist processing. However, as set out below, previous research allowed for inferences to be made as to the ERP components most likely to be correlated with the effect of congruency on scene processing.

Using such a methodology provides the opportunity to better understand the timing of expectation-related alterations to gist processing. While manipulation of target presentation duration in behavioural studies can help point towards the general speed of an effect, it alone cannot determine the point of occurrence for expectation-induced violations to scene processing; despite utilising dynamic backward masking, a complete cessation to target-image processing at offset is unlikely. Event-related potentials, on the other hand, offer a more precise means by which to uncover the temporal points at which differences in brain activity emerge as a function of scene congruency, allowing for conclusions to be drawn as to the potential swiftness of any top-down influence.

Therefore, the first ERP component selected for investigation was the P2. Arising rapidly within Parieto-occipital regions – at around 200 ms after target onset – this component has been proposed as the earliest known marker for scene-specific processing (Harel et al., 2016), affected by changes in global scene properties but not top-down observer-based goals (Hansen et al., 2018). However, the exact influence of top-down information on the P2 remains unclear. While evidence indicates early components such as this are sensitive to low-level visual information such as salience (Straube & Fahle, 2010), as well as object identification (Viggiano & Kutas, 2000), the influence of higher-level processes is less well determined. For example, differences in ERPs at around 200 ms have been found when identifying the presence of objects within briefly presented natural images, potentially reflecting decision-related activation (Thorpe et al., 1996; VanRullen & Thorpe, 2001). As a result, a lack of agreement exists, both in terms of whether the P2 is altered by top-down processing at all and, if so, what form of top-down processing might hold influence. Furthermore, it has been suggested the P2 may in fact index an intermediary processing stage, somewhat bridging perceptual and higher-order processes, such as segmentation and categorisation, respectively (De Cesarei et al., 2013).

In terms of the current study, the above implies predictions relating to the P2 must be tentative. We can contend, however, that if congruency-based differences in activation were shown to exist within the earliest indicant of scene-specific processing (Harel et al., 2016), this would be representative of expectations influencing early perceptual processing. More broadly, finding such activation differences would signify that the P2 component is open to influence from top-down information. Indeed, top-down modulation of the P2 as being related to the semantic processing of scenes has been proposed before (Federmeier & Kutas, 2002).

As well as determining the timing of initial alterations to gist processing, ERP analysis can help elucidate the mechanisms underlying expectation-related performance changes as cognitive processing continues. In other words, investigating activation changes across subsequent scene-related ERP components can help reveal the manner in which scene congruity might affect gist processing. Previous scene processing research has shown two later components as being susceptible to experimental manipulation. The first of these – the N400 – has long been associated with semantic processing, with its amplitude observed as being inversely proportional to semantic expectancy (e.g., Kutas & Hillyard, 1984) and more generally to the ease with which conceptual information can be retrieved (Van Petten & Luka, 2006). For this component a certain level of consensus has been reached: across both central and anterior sites, increased negativity within the N400 time window has been related to scene-object semantic violations in static images (Ganis & Kutas, 2003; Mudrik et al., 2010; Võ & Wolfe, 2013) and within video clips (Sitnikova et al., 2003; Sitnikova et al., 2008). N400 effects have also been found to be sensitive to the semantic association between pairs of sequential pictures (Barrett & Rugg, 1990), and to violations of semantic expectation in language comprehension studies (Holcomb, 1993; Van Petten, 1995). A similar pattern of N400 changes across conditions within the current study would, therefore, indicate that differential behavioural performance derived from congruency-based manipulations in the behavioural experiments was due to semantic violations, rather than simply violations of expected low-level visual information.

The second of these later components, again potentially revealing in terms of the mechanisms responsible for the effect of expectations on processing, is the P600. Like the N400, this component was initially described in language comprehension studies, where syntactic errors creating a need for sentence reanalysis were observed to elicit increased positivity within posterior regions at ~600 ms (Hagoort et al., 1993; Osterhout & Holcomb, 1992). This was irrespective of whether sentences were experienced through visual or auditory modalities (Hagoort & Brown, 2000; Osterhout & Holcomb, 1993). Similarly, within scene processing research, increased positivity at the P600 has been reported as reflecting reanalysis prompted by mis-located objects (Võ & Wolfe, 2013). There, increased late positivity was found when appropriate objects were positioned in inappropriate places within a scene (such as a dishtowel on a kitchen floor), irregularities proposed by the authors as reflecting syntactic – rather than semantic – violations. However, a lack of agreement should be noted regarding the functional role of the P600. For instance, It has been suggested that this increased late positivity may not exclusively represent syntactic violations, as its sensitivity to semantic information has also been demonstrated (Gunter et al., 1997; Gunter et al., 2000; Kuperberg, 2007; Sitnikova et al., 2003).

Furthermore, such changes to late positive components have not always been observed when objects break syntactic rules within scenes (e.g., Demiral et al., 2012). To confuse matters further, while Võ and Wolfe (2013) did not find alterations to the P600 when inappropriate objects were placed in appropriate locations – taken by the authors as evidence of the dissociation between the effects of semantic and syntactic violations – this form of semantic violation was shown to elicit a reduction in P600 amplitude in previous work (Mudrik et al., 2010). It is possible this inconsistency across studies is rooted in contrasting methodological choices, with one allowing for expectations to be generated due to the context-scene appearing prior to the target object (Võ & Wolfe, 2013), and the other avoiding this through simultaneous presentation of targets and their associated scenes (Mudrik et al., 2010).

There is still much debate as to the comparability between language and scene processing in general, particularly in terms of whether the processing of words and pictures shares a common semantic system (e.g., Federmeier & Kutas, 2002), and questions remain for both paradigms in relation to the nature of the P600. However, perhaps the most reproducible findings regarding this component have been through the use of 'garden path' sentences in linguistic studies, whereby violation of the expected structure of a sentence creates the need for reanalysis of the preceding sequence of words (e.g., Frazier & Fodor, 1978; Osterhout & Holcomb, 1992). Accordingly, display of similar congruency-related changes to the P600 in the current study would suggest the violation of expectations, created through approach images, resulted in the need for reanalysis of incongruous targets. In other words, just as a garden path sentence might build an inaccurate expectation as to the grammatical structure of a sequence of words, which is subsequently violated, a sequence of images depicting a journey is likely to build expectations as to the eventual destination, only for this to be violated on presentation of an incongruous target scene.

Based on the above, predictions as to the pattern of results can be made, although they must remain speculative due a lack of consensus across previous research. Firstly, the association between the N400 and semantic incongruity, observed in studies of language and scene processing, leads us to expect greater N400 amplitudes across central and anterior regions for Incongruous trials, as compared to Congruous trials. Secondly, due to violations in thematic coherence, we expect to observe increased P600 amplitude across posterior sites for Incongruous trials, as compared to Congruous trials. Finally, due to debate remaining as to the influence of top-down factors on the P2 component, we do not make predictions as to whether Incongruous trials will elicit increased positivity in posterior regions during this time-window. However, if such changes were observed, we would take this as signalling the violation of expectations was able to influence the earliest stages of scene processing, including the integration of visual properties.

## Design

To maintain consistency throughout the study the experimental protocol mirrored previous iterations closely, although with certain alterations necessary to improve the suitability of the trial routine for use with electroencephalography. The 120 experimental trials were split equally across two conditions of approach-destination congruity, and there was no manipulation of approach sequentiality. Sixty image-series were randomly selected to serve as Incongruous trials, with their destination images randomly redistributed amongst themselves. The same restrictions were applied to the randomisation procedure as previous experiments. A counterbalanced version of the protocol was created, and these two versions were employed in a Latin Square across the course of the experiment to ensure no unintended bias was introduced due to the allocation of trials to congruency conditions. See Figure 8 for a schematic of the experimental protocol.

## Participants

Experiment 4 included 26 Psychology students, again recruited through the research pool and given course-related credits for taking part ($M_{\text{age}}$ = 20.31, $SD_{\text{age}}$ = 2.77; 20 Females, 6 Males; 19 Right-handed, 7 Left-handed). This is slightly below the number suggested from our initial power calculation (~30 participants) but is standard for ERP research. Furthermore, the relatively large congruency effects seen in the previous iterations allowed for confidence in the experiment remaining sufficiently powered despite this reduction. None had participated in any of the behavioural experiments, and so all were unfamiliar with the stimuli and naïve to the purpose of the study. One participant was removed as their comprehension of the task could not be assured (incorrectly responding to 77% of Incongruous trials), and another removed due to excessive high-frequency noise across multiple channels. Analyses were conducted on the remaining 24 participants ($M_{\text{age}}$ = 20.42, $SD_{\text{age}}$ = 2.86; 18 Females, 6 Males; 18 Right-handed, 6 Left-handed). All participants reported as having no history of neurological disorders.

**Stimuli**

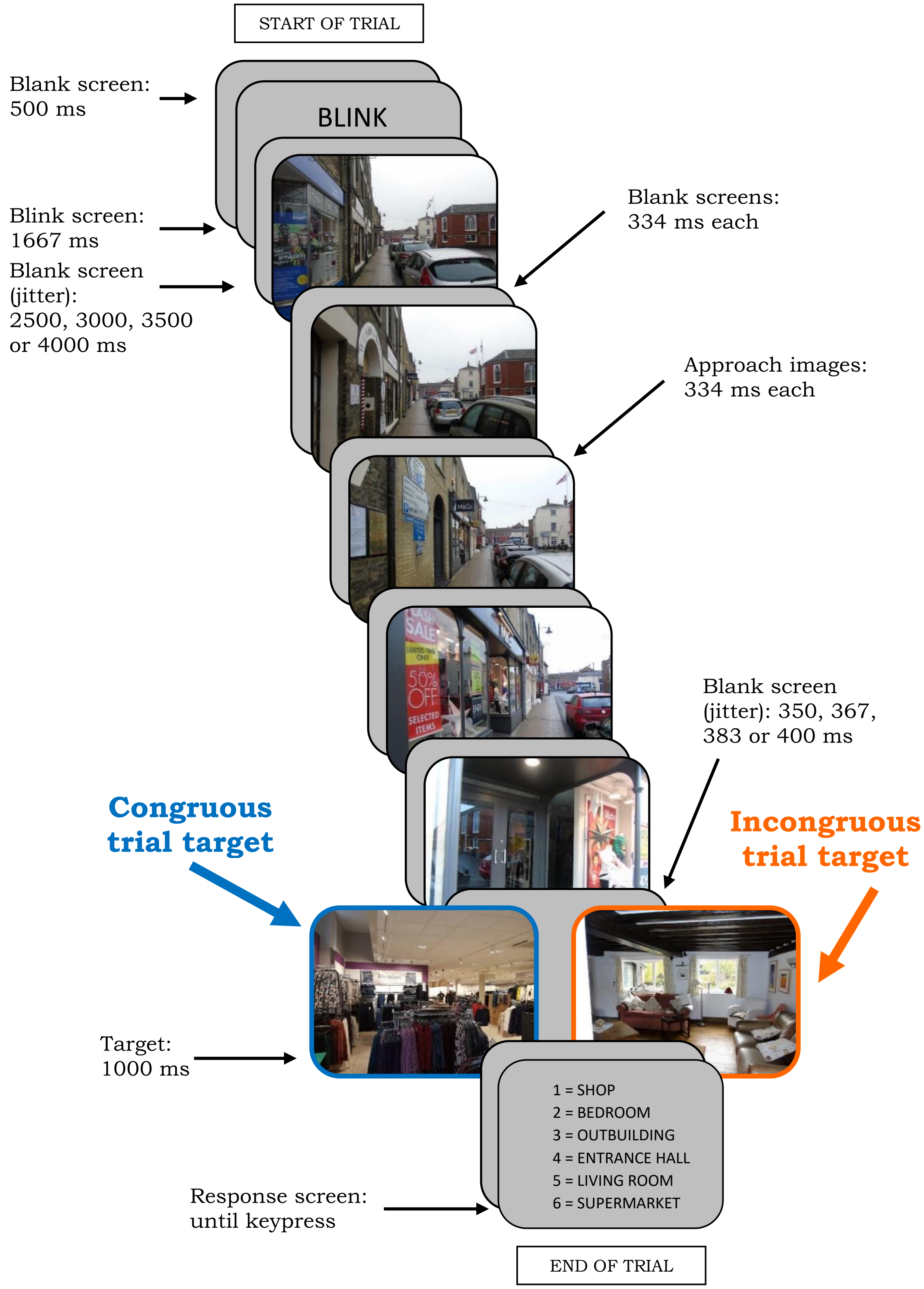
The same image set was again used, although all masks were removed for Experiment 4. Response screens were reconstructed using the same guidelines as previous experiments.

**Procedure**

Each trial began with a ‘blink’ screen, followed by a blank screen including a jitter (duration: 2.5, 3, 3.5 or 4 seconds). This was to protect the ERPs from the potential systematic influence of slow baseline drifts coinciding with the routine. The jitter was pseudo-randomised to ensure a different blank screen duration prior to each of the four approach-series per scene category. These initial screens gave participants the opportunity to get comfortable prior to the presentation of each trial, with the aim of reducing the number of movement-based artefacts within the subsequent ERPs. A second, shorter jitter (duration: 350, 367, 383 or 400 ms) was also introduced to the last blank screen prior to target presentation, to shield against artefacts caused by participants being able to predict the exact onset time of the target. This jitter was pseudo-randomised in the same manner as

**Figure 8**

*Schematic of the Protocol for Experiment 4*



before, and was evenly distributed across the two congruency conditions. There was no manipulation of target duration, with presentation length set at 1 second. This extended duration served two purposes: firstly, as only correctly answered trials were used in the analysis it sustained a high level of categorisation performance and, secondly, it protected against noise within the ERP caused by the offset of the stimulus or the onset of the response screen. No masking was used, with the target followed by a blank screen prior to a 6AFC response screen.

**Data Acquisition**

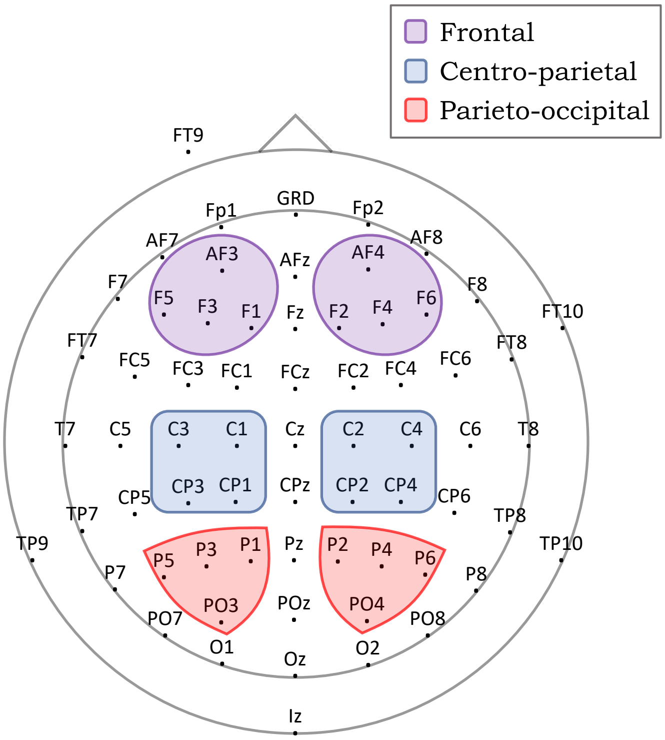
The EEG was recorded using a Brain Vision 64-channel active electrode system, embedded within a nylon cap (10/20 system). Electrode FT9 was removed from the cap and placed under the left eye to monitor blinks and eye movements. The signal was acquired at a 1000 Hz sampling rate with FCz used as the online reference (see Figure 9).

**Processing**

Offline processing and analyses were conducted using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014), running under Matlab 9.2.0 (R2017a, Mathworks, Natick, MA). Trials with incorrect responses were removed from the continuous EEG (3.89% of Congruous trials and 5.01% of Incongruous trials across participants). Ocular artefact correction took place through Independent Component Analysis (ICA) to identify blinks and lateral eye movements. These artefacts are located at anterior electrodes and can be identified based on their characteristic shapes (frequent clear spikes or step-like functions, respectively). Therefore, removal of these components was conducted manually by simultaneously comparing the continuous EEG to the time-course of the Independent Components. This led to removal of 41 Independent Components across the sample as a whole, with no more than two components removed for any single participant. Re-referencing to the average of the TP9 and TP10 electrodes (which approximate to the location of the mastoids) was computed offline (e.g., Cohn & Foulsham, 2020). Any channels suffering from persistent high-frequency noise were interpolated using the mean signal from the surrounding electrodes (mean percentage of channels interpolated across participants: < 1%). After removal of DC trends, an IIR Butterworth filter was applied for high- and low-pass filtering the data with half-amplitude cut off values of 0.01 Hz and 80 Hz, respectively (12 dB/oct; 40 dB/dec). The EEG was segmented into epochs of 1 second, from 200 ms before to 800 ms after target-scene onset. The length of the baseline used to correct epochs was the 200 ms immediately preceding target onset. Epochs contaminated with excessive artefacts were identified, and rejected, by setting a peak-to-peak voltage threshold of 100 µV across a moving window of 200 ms with a window step of 50 ms. This resulted in the rejection of 6.94% of Congruous trials and 7.10% of Incongruous trials across participants.

**Figure 9**

*Map of Electrode Placement Including the ROIs*



*Note.* FT9 was removed from the cap and placed on the left cheekbone to monitor blinks.

The amplitudes of the P2, N400 and P600 were measured as the mean of all data points between 175-250 ms, 300-500 ms and 500-700 ms, respectively. These specific components were chosen as the P2 has previously been suggested as the earliest indicator of scene selectivity (Harel et al., 2016), while the N400 and P600 have been associated with semantic and syntactic integration, respectively (e.g., Friederici et al., 1993; Hagoort & Brown, 2000; Holcomb, 1993; Mudrik et al., 2010; Van Petten, 1995; Võ & Wolfe, 2013).

The time windows chosen are commonly used as boundaries for investigating the N400 (e.g., Ganis & Kutas, 2003; Guillaume et al., 2016; Mudrik et al., 2010) and P600 (Angrilli et al., 2002; Cohn et al., 2014; De Vincenzi, 2003). Less standardisation exists regarding the P2, however, with previous research involving the processing of scenes employing time windows ranging anywhere between 140 to 320 ms post-stimulus onset (see, for example, De Cesarei et al., 2013; Ferrari et al., 2017; Harel et al., 2020; Yuan et al., 2007). We, therefore, determined our window of interest based on visual inspection of the grand average ERP. As a result, a window of 175-250 ms was selected as it covered the 220 ms timepoint previously identified as showing maximal amplitude for scene processing (Harel et al., 2016), while offering as large a span as was achievable without incorporating elements of the proximal P1 and P3 components.

Key electrode sites were grouped into three regions of interest (ROIs), each incorporating eight electrodes (split equally across hemispheres). A Centro-parietal ROI included electrodes C1/C2, C3/C4, CP1/CP2 and CP3/CP4, a Parieto-occipital ROI comprised electrodes P1/P2, P3/P4, P5/P6 and PO3/PO4, and a Frontal ROI contained electrodes F1/F2, F3/F4, F5/F6, and AF3/AF4 (see Figure 9). The posterior ROI was selected as Parieto-occipital regions are associated with maximal amplitude of the P600 (e.g., Gouvea et al., 2010) and the P2 (e.g., Hansen et al., 2018). The more central and anterior ROIs were chosen as the amplitude of the N400 has previously been found to be maximal at Centro-parietal regions (e.g., Ganis & Kutas, 2003), while the processing of semantic information related to images, as compared to text, has often been shown to elicit a Frontal negativity during the 300-500 ms temporal window (e.g., Ganis et al., 1996; Holcomb & McPherson, 1994; Mudrik et al., 2014).

**Results**

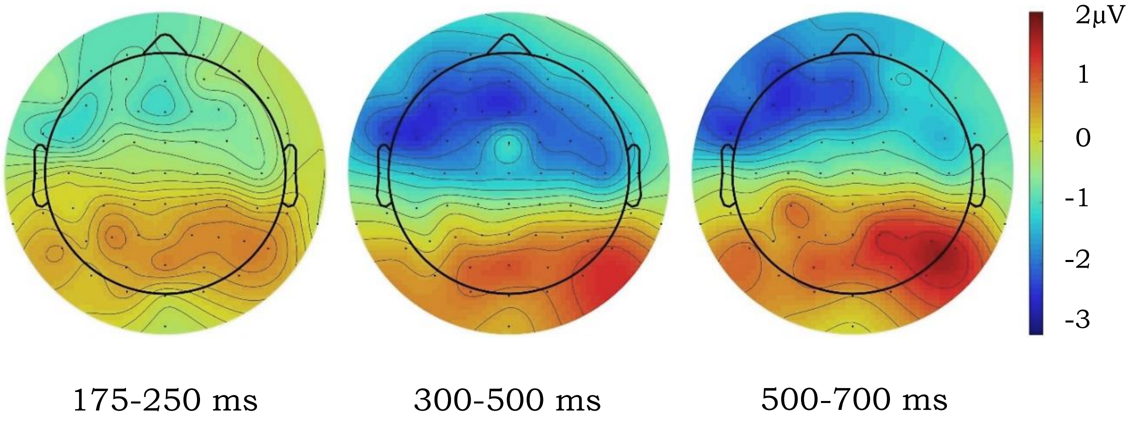
Analysis was conducted on the mean amplitudes for each time-period of interest using 2 (Hemisphere: Left; Right) x 3 (Region: Centro-parietal; Parieto-occipital; Frontal) x 2 (Congruency: Congruous; Incongruous) repeated-measures ANOVAs. Where Mauchly’s test revealed possible violations of the sphericity assumption Greenhouse-Geisser corrected values are reported (Greenhouse & Geisser, 1959). Significant interactions were followed up with paired t-tests where appropriate. See Appendix C for a summary of the statistical analyses conducted.

### *175-250 ms Window*

A three-way ANOVA revealed no main effect of Congruency (*p* = .842). There was also no three-way interaction (*p* = .552), nor a Hemisphere x Region interaction (*p* = .396), nor a Hemisphere x Congruency interaction (*p* = .424). There was, however, a significant Region x Congruency interaction *F*(2, 46) = 15.68, *p* < .001, ηp$^2$ = .41. See Figure 10 for scalp maps of voltage differences across conditions. In terms of this interaction, follow-up paired t-tests revealed no significant effect of congruency within the Centro-parietal ROI (*p* = .893). There was a significant effect within the Frontal region, *t*(23) = 2.54, *p* = .018, *r* = .47, due to there being a significantly more negative mean

**Figure 10**

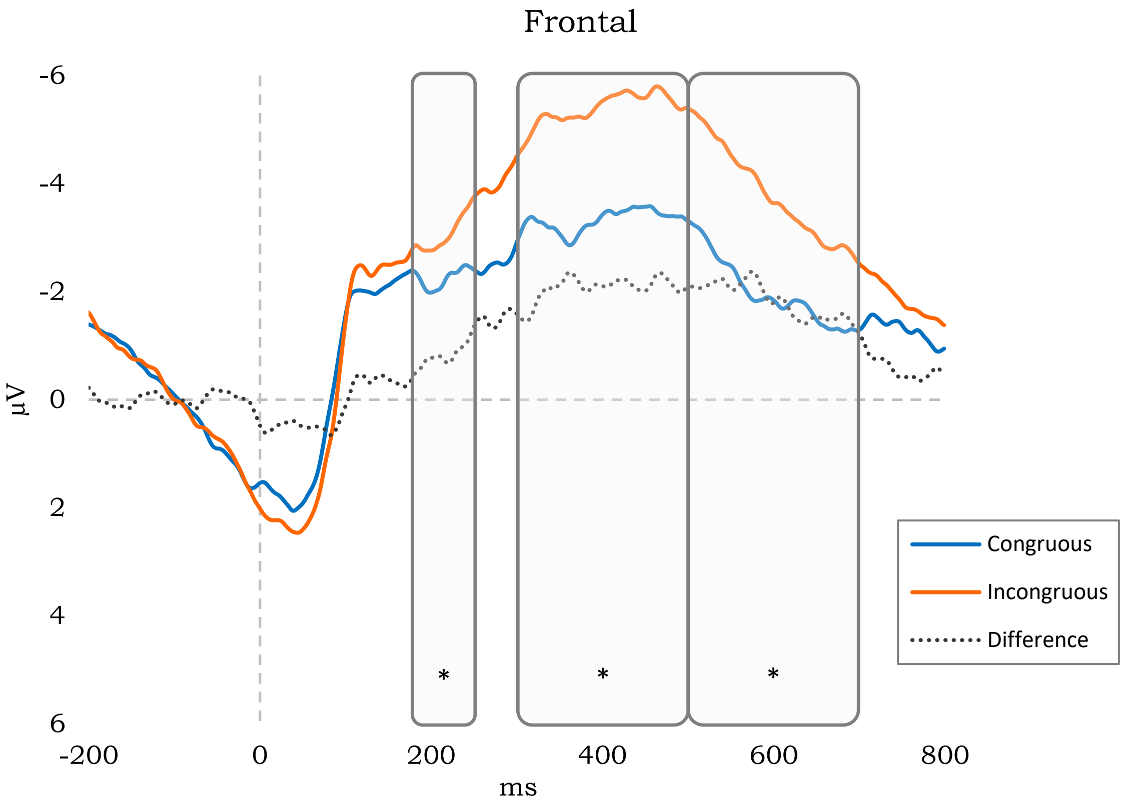
*Scalp Maps of the Mean Voltage Difference Between the Congruency Conditions for Each of the Time Windows Under Investigation*



*Note.* Blue colours indicate the difference is negative, while red colours indicate the difference is positive. Scalp maps represent Incongruous minus Congruous amplitudes.

**Figure 11**

*Grand-averaged ERPs for the Frontal Region, Collapsed Across Hemispheres*

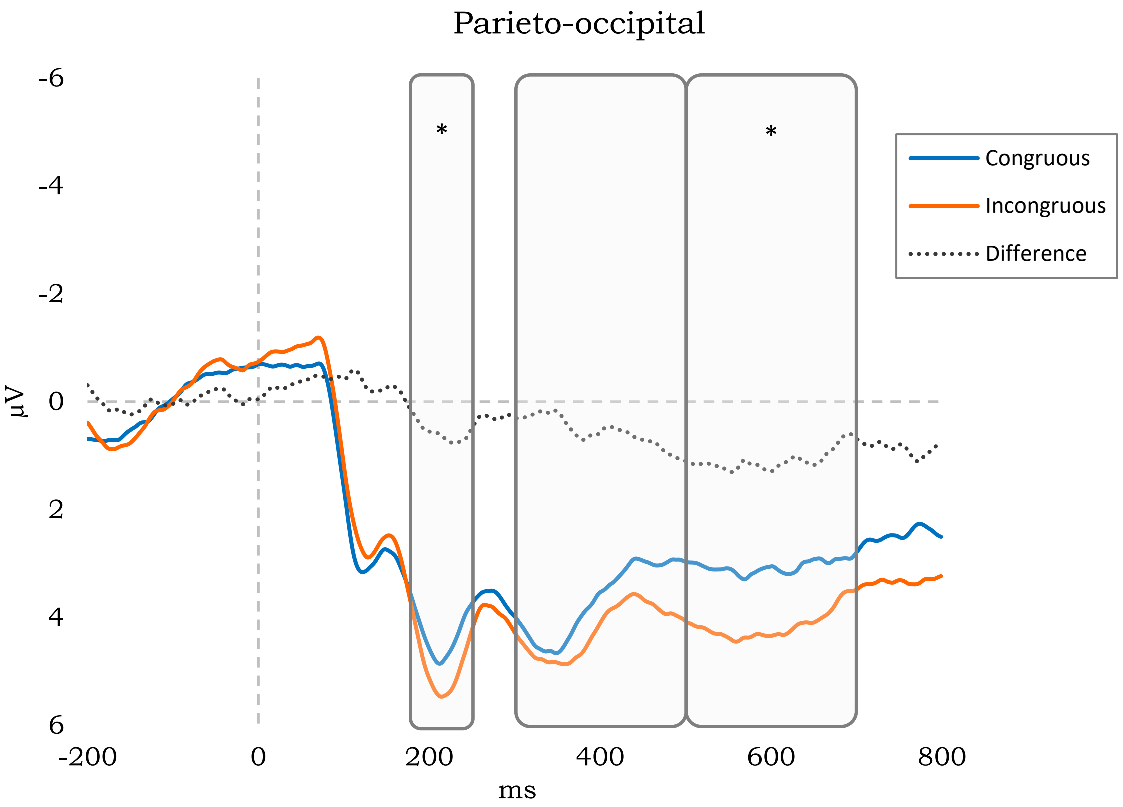


*Note.* Blue lines represent amplitudes for Congruous trials and orange lines represent amplitudes for Incongruous trials. Dotted line represents the difference wave (Incongruous minus Congruous). Waveforms low-pass filtered at 30Hz for display purposes (*n* = 24). Grey boxes represent the three time-windows of interest. * denotes *p* < .05.

amplitude for Incongruous trials (*M* = -3.06 µV) than Congruous trials (*M* = -2.27 µV). See Figure 11 for the grand-averaged Frontal ERPs. This represents a medium-to-large effect. There was also a significant effect within the Parieto-occipital region, *t*(23) = -2.08, *p* = .048, *r* = .40, representing a medium-sized effect. This was due to there being a significantly more positive mean amplitude for Incongruous trials (*M* = 4.90 µV) than Congruous trials (*M* = 4.33 µV) within Parieto-occipital areas (see Figure 12 for the grand-averaged Parieto-occipital ERPs). Additionally, we re-ran our analysis at slightly more lateral posterior sites, in regions where maximal P2 changes have previously been shown (e.g., Harel et al., 2016; Harel et al., 2020; Hansen et al., 2018). This confirmed our finding of congruency-related changes to the P2 component (see Appendix D for further details).

**Figure 12**

*Grand-averaged ERPs for the Parieto-Occipital Region, Collapsed Across Hemispheres*



*Note.* Blue lines represent amplitudes for Congruous trials and orange lines represent amplitudes for Incongruous trials. Dotted line represents the difference wave (Incongruous minus Congruous). Waveforms low-pass filtered at 30Hz for display purposes (*n* = 24). Grey boxes represent the three time-windows of interest. * denotes *p* < .05.

### *300-500 ms Window*

A three-way ANOVA revealed a main effect of Congruency, *F*(1, 23) = 6.16, *p* = .021, ηp$^2$ = .21, due to there being significantly more negative mean amplitudes for Incongruous trials (*M* = -0.85 µV) than Congruous trials (*M* = -0.02 µV) during this time-window. There was no three-way interaction (*p* = .136), nor a Hemisphere x Region interaction (*p* = .274), nor a Hemisphere x Congruency interaction (*p* = .117). There was, however, a significant Region x Congruency interaction *F*(1.46, 33.65) = 32.92, *p* < .001, ηp$^2$ = .59. In terms of this interaction, follow-up paired t-tests revealed no significant effect of congruency within the Parieto-occipital region (*p* = .129). However, there was a significant effect within the Centro-parietal region, *t*(23) = 2.40, *p* = .025, *r* = .45. This represents a medium-to-large effect. This was due to there being a significantly more negative mean amplitude for Incongruous trials (*M* = -1.39 µV) than Congruous trials (*M* = -0.41 µV) within the Centro-parietal region during this time-window (see Figure 13 for the grand-averaged Centro-parietal ERPs). There was also a significant effect within the Frontal region, *t*(23) = 5.37, *p* < .001, *r* = .75, representing a large effect. This was due to there being a significantly more negative mean amplitude for Incongruous (*M* = -5.40 µV) than Congruous trials (*M* = -3.33 µV).

***500-700 ms Window***

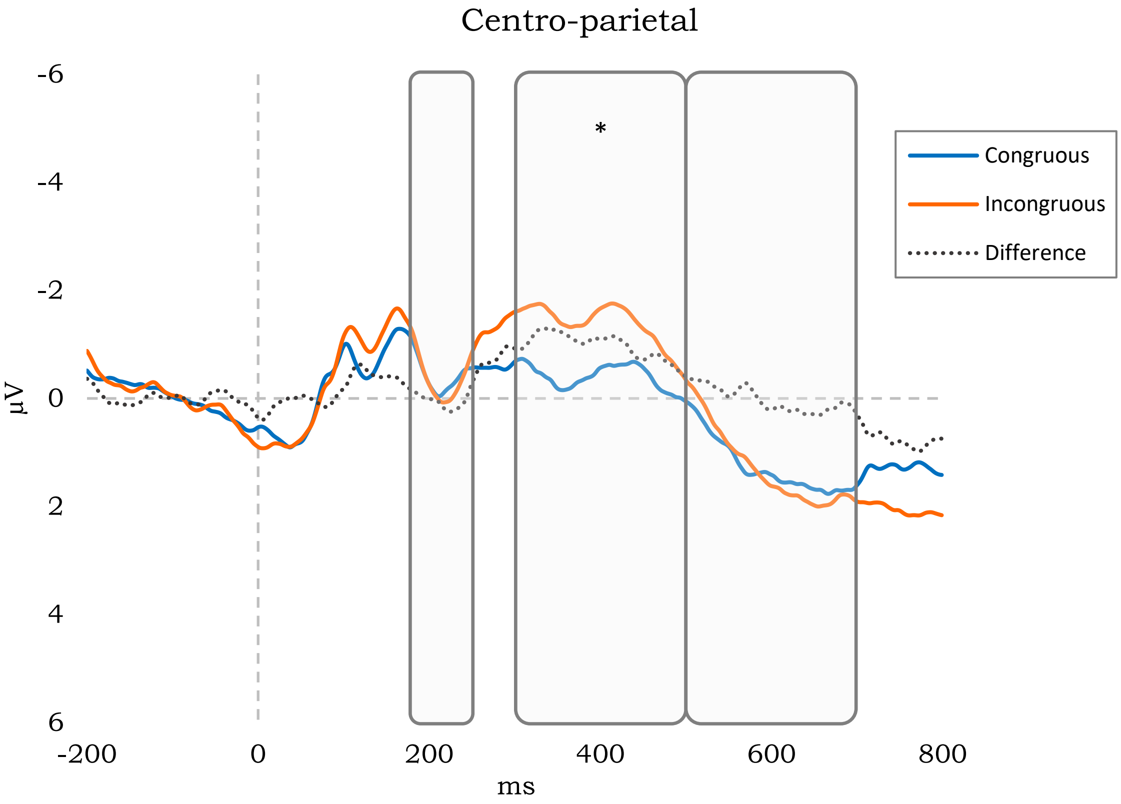
A three-way ANOVA revealed no main effect of Congruency (*p* = .553). There was also no three-way interaction (*p* = .056), nor a Hemisphere x Region interaction (*p* = .656). There was, however, a significant Hemisphere x Congruency interaction, *F*(1, 23) = 5.72, *p* = .025, ηp$^2$ = .20.

Follow up paired t-tests for this interaction, collapsed across region, revealed no significant congruency-related difference in amplitude in either the left *(p* = .185) or right hemisphere (*p* = .958). There was also a significant Region x Congruency interaction *F*(2, 46) = 34.05, *p* < .001, ηp$^2$ = .60. In terms of this interaction, follow-up paired t-tests revealed no significant effect of congruency within the Centro-parietal region (*p* = .972), but did find a significant effect within the Parieto-occipital region, *t*(23) = -2.41, *p* = .025, *r* = .45. This represents a medium-to-large effect. This was due to there being a significantly more positive mean amplitude for Incongruous trials (*M* = 4.15 µV) than Congruous trials (*M* = 3.06 µV) within the Parieto-occipital region during this time-window. There was also a significant effect of congruency within the Frontal region, *t*(23) = 4.57, *p* < .001, *r* =

.69, representing a large effect. This was due to there being significantly more negative mean amplitudes for Incongruous ($M$ = -3.83 µV) than Congruous trials ($M$ = -1.99 µV).

**Figure 13**

*Grand-averaged ERPs for the Centro-Parietal Region, Collapsed Across Hemispheres*



*Note.* Blue lines represent amplitudes for Congruous trials and orange lines represent amplitudes for Incongruous trials. Dotted line represents the difference wave (Incongruous minus Congruous). Waveforms low-pass filtered at 30Hz for display purposes ($n$ = 24). Grey boxes represent the three time-windows of interest. * denotes $p$ < .05.

### ERP Results Summary and Discussion

Compared to Congruous trials, Incongruous trials displayed a significantly more positive mean amplitude for the P2 across the Parieto-occipital region, a significantly more negative mean amplitude for the N400 across the Centro-parietal and Frontal regions, and a significantly more positive mean amplitude for the P600 within the Parieto-occipital region. We also found significantly more negative amplitudes for Incongruous trials within the Frontal region across the early and late windows of interest.

The congruency-related amplitude changes seen in the P2 component help firm our contention that observer expectations were affecting gist processing, as this suggests top-down information was having an influence while perceptual processing was still ongoing. The sensitivity of the P2 to top-down information is still debated (e.g., Hansen et al., 2018), although previous work has relied on the presentation of individual images. It may be that changes to this early component seen here result from participants being able to generate expectations prior to target onset, thus providing the opportunity for more immediate top-down influence once the destination image is presented.

Amplitude changes across conditions were also apparent in the N400. There is a level of consensus that this component is a neural marker for semantic processing (see, for a review, Kutas & Federmeier, 2011), and so this finding supports our assertion that expectations were influencing higher-level cognitive processes rather than simply the extraction of low-level visual information. Furthermore, changes to the N400 have repeatedly been demonstrated for violations to semantic expectations within language (e.g., Van Petten, 1995) and across object-scene pairs (e.g., Demiral et al., 2012), and so the current findings suggest that an equivalent effect exists for violations between separate scenes. We found these N400 amplitude changes within the Centro-parietal as well as the Frontal region, in line with research showing the semantic processing of pictorial stimuli elicits a more anteriorly located negativity during this temporal window (e.g., Ganis et al., 1996). Further to this, we saw morphological dissimilarities within the ERPs across these two regions. In particular, the congruency-related amplitude differences in the anterior region spanned all three time-windows investigated, meaning the effect of approach-destination congruency was apparent in Frontal sites within 175-250 ms from target onset.

As with the N400, changes to the P600 again suggest an influence of higher-level processes on gist extraction. There is still debate, within both scene and language research, as to whether alterations to the P600 are a reflection of difficulties with semantic (e.g., Mudrik et al., 2010; Sitnikova et al., 2003) or syntactic (Hagoort & Brown, 2000; Võ & Wolfe, 2013) processing, or indeed an integration of both (Friederici & Weissenborn, 2007). Additionally, different aspects of syntactic processing have been proposed as being reflected in the P600 (e.g., Friederici et al., 2002; Gouvea et al., 2010; Kaan et al., 2000). For example, increased positivity at P600 has been elicited during 'garden path' sentences, which require a re-interpretation of expectations while reading a sentence due to an atypical grammatical format (Osterhout & Holcomb, 1992). Some equivalence to the current study is apparent, whereby expectations are built during a progression of approach images only to require re-evaluation once violated by the appearance of an incongruous destination.

In sum, differences were found in a putative scene-selective ERP component, related to integrating visual properties (P2), as well as later components related to contextual integration including semantic and syntactic coherence (N400 and P600, respectively).

## General Discussion

We conducted a series of behavioural and ERP experiments, involving the presentation of 'approach' images prior to target scenes. In so doing, we hoped to better understand the manner by which scenes are processed outside the laboratory, namely as elements of a progression of contextual information rather than simply isolated images. This allowed us to investigate whether semantic information was derived from the advancement through environments, and whether this generated 'on-line' expectations able to facilitate the processing of subsequent scenes. Experiments 1a and 1b investigated the effect of expectations on the processing of conceptual gist within scenes, through the manipulation of 'approach-destination' congruency, as well as the time-course of the effect through manipulation of target display duration. Experiment 2 then manipulated the sequentiality of these pre-target series, in order to investigate the influence of spatiotemporal coherence on gist processing, while Experiment 3 introduced a baseline condition to disentangle the separate roles of facilitation and interference on gist processing. Finally, in Experiment 4 we employed electroencephalography to chart the neural correlates associated with the manipulation of scene congruency.

As predicted, across experiments we found a benefit for categorising scenes when semantically congruous with lead-up images. Also in line with predictions, Experiment 1a revealed an advantage that was greatest at shorter target durations, where the opportunity to process visual information was most limited. This pattern of results was mirrored in Experiment 1b, where congruous target scenes only appeared on a quarter of trials, indicating the effect was not simply based on the frequency of conditions and that participants' predictions as to an upcoming destination were being driven by an automatic mechanism. Next, Experiment 2 revealed that the performance advantage seen for Congruous trials was based on approach images providing a semantic context for upcoming targets. While a main effect of approach-image sequentiality was found, an increase in categorisation ability for Congruous-Sequential compared to Congruous-Disordered trials only neared significance, contrary to our predictions. Then, Experiment 3 confirmed that providing participants with semantically congruous approach images led to a facilitation of gist processing, as hypothesised, and demonstrated reduced performance compared to baseline when trials were incongruous in nature.

Finally, Experiment 4 investigated the neural correlates of predictability on rapid scene processing, showing an effect across all tested ERP components. For Incongruous trials, the P2 and P600 showed significantly greater mean amplitudes within the Parieto-occipital region, while a significantly more negative mean amplitude for the N400 was seen within the Centro-parietal and Frontal regions. Furthermore, Incongruous trials were also associated with a significantly more negative amplitude across the early and late time-windows within Frontal sites. Taken together, this meant we found congruency-related changes within the earliest known indicant of scene-specific processing (P2), within the component classically proposed as an index of semantic expectancy as well as the retrieval of conceptual information (N400), and within the component associated with both semantic and syntactic processing (P600). We will begin by addressing the findings from the behavioural experiments, before moving on to discuss potential interpretations for the task-related alterations to brain activity seen here.

Firstly, the condition-based differences in categorisation performance within Experiments 1a-b reveal that an observer's expectations can alter scene processing. Importantly, the most substantial differences were found at target durations of 50 milliseconds and below, indicative of expectations influencing the earliest stages of processing. It appears, therefore, that top-down information has a role in modulating the extraction of scene gist. These results are perhaps not surprising when we consider the considerable quantity of research demonstrating an influence of expectations on the subsequent processing of the environment (Endsley & Garland, 2000; Langham et al., 2002; Mahon, 1981). Furthermore, such results appear to mirror the finding that the processing of objects is facilitated when contextually related to the scenes in which they are embedded (Antes et al., 1981; Biederman et al., 1973; Boyce & Pollatsek, 1992, Davenport & Potter, 2004; Underwood, 2005; Võ & Henderson, 2011). This has not only been found during the simultaneous presentation of scenes and their objects, but also when a scene is presented and then removed from view prior to the target object being displayed (Palmer, 1975).

So, just as the rapid processing of semantic information can influence the processing of objects, our results show the same is true for scenes. While 'scene priming' effects have been reported before (e.g., Sanocki & Epstein, 1997), such facilitation is likely based on low-level information, such as the priming of visual features (Brady et al., 2017) or the maintenance of basic scene layout in memory (Oliva & Torralba, 2001). The findings seen here, on the other hand, relate to the processing of scenes at the semantic level, as further discussed below, and so propose an expansion of the concept of scene priming. Just as a scene can be semantically primed by text displayed prior to its presentation (Reinitz et al., 1989), it appears that similar facilitation is possible when a target scene is preceded by a semantically relevant context. Such a finding adheres to the typical mechanisms underlying visual processing, where the constraints of cognitive capacity drive us to look for predictable patterns within the environment, from which to form expectations that lower the demands of subsequent processing (e.g., Bartlett, 1995; Gregory, 1997). In addition, it lies in agreement with recent findings that demonstrate pre-target narrative sequences are able to affect subsequent scene processing (Smith & Loschky, 2019).

The relationship between expectations and gist processing builds on complimentary work concerning improbable scenes (Greene et al., 2015). That research uncovered increased difficulty in understanding the meaning of atypical scenes, pointing to a disruption in gist processing when scenes diverge from what an observer expects. Such findings strongly point to a role of top-down information in rapid scene understanding, although there is an important distinction to our work. The violation of expectations within single scenes – such as a boulder inside a room (Greene et al., 2015) – would potentially result from inconsistencies between the bottom-up signal and a template stored in long-term memory. On the other hand, our study showed the effect of violating predictions based on the on-line flow of information: the introduction of approach images meant predictions could be formed prior to target onset, potentially resulting in the pre-activation of templates expected to be required for matching against the stimulus. Such pre-activation provides the opportunity for a stored representation to be available prior to the appearance of the target-derived signal, conceivably resulting in more rapid matching or, through predictive coding mechanisms, allowing for the detection of inconsistencies at an earlier processing level due to pre-emptive changes in error thresholds (Rauss et al., 2011).

These results stand in opposition to 'forward sweep' models, which assume minimal top-down modulation of gist processing (Itti et al., 1998; Potter et al., 2014; Rumelhart, 1970). However, we do not contend this necessarily rejects the primacy of bottom-up visual factors in scene perception. Across each of our behavioural experiments the accuracy with which scenes were categorised far exceeded chance level, even when approach images were incongruous with destinations. In other words, some degree of gist processing was still possible when no relevant semantic information was provided prior to destination-scene onset. Therefore, we propose that feature extraction mechanisms may well be capable of rapidly distinguishing a great deal of information within complex natural scenes (e.g., Potter et al., 2014), but that these mechanisms are susceptible to influence from higher-level processing. This might particularly be the case when, as in our design, antecedent information is provided before gist processing begins, thereby allowing for the formation of expectations prior to a scene being encountered. So, we cannot comment on the processing of individual, segregated scenes, as we did not investigate this. What we do contend, however, is that under conditions which better reflect functioning outside the laboratory it appears that top-down information, in the form of expectations, affects conceptual gist processing.

Secondly, the results from Experiment 2 show the performance advantage for Congruous trials was largely due to the provision of contextual information. More specifically, the facilitation of gist processing through expectations appears to be driven by the observer being provided with semantic information, in the form of an environmental setting, from which more accurate predictions can be formed. Perhaps surprisingly, the sequentiality of the approach images did not significantly influence performance, and so we found no evidence that the spatiotemporal coherence of series, or the generation of a perceived flow of movement, had a bearing on participants' categorisation ability. This was counter to our predictions, based on recent work identifying an important role for the narrative coherence of image sequences (Cohn et al., 2012; Foulsham et al., 2016; Smith & Loschky, 2019).

While we did find evidence for a significant Congruency-Sequentiality interaction, as well as a significant main effect of sequentiality, the expected effect within Congruous trials only neared significance. This suggests that – within the particular constraints imposed by our design – if an effect of sequentiality does exist its influence is much reduced as compared to the effect of congruency. However, this is not necessarily true under all circumstances, and there is an important distinction from previous work. The approach images adopted here were within relatively close proximity to their eventual destination, and so the narrative created by these series is not comparable to the narrative created across the panels of a comic strip (e.g.., Cohn et al., 2012; Foulsham et al., 2016), nor the strings of images depicting a journey from one distinct location to another that is spatially distant (Smith & Loschky, 2019). So, the generated story of "I am approaching a shop" may have remained unaltered irrespective of whether the leading images were sequential or not, something that would be unlikely for the more complex narratives within previous designs. However, while the narrative of our sequences may not have been overly disrupted by being disordered, this manipulation certainly disrupted the appearance of linear movement. As a consequence, our results find there to be comparatively minimal additional benefit from the creation of 'perceived flow' (e.g., Gibson, 1966).

A final point to make relating to the results of Experiment 2 is that they help alleviate any concern regarding the origin of the congruency-based changes in performance. Our design meant some target scenes necessarily contained low-level similarities with the most proximal leading images. For instance, on a congruous (and sequential) trial the final leading image of an approach to a shop may allow for some information relating to the final scene to be pre-empted because, say, the general dimensions of the store could be determined by a partial view through the window. For this reason, it might be argued that the categorisation performance changes across congruency conditions were due to the priming of low-level visual features prior to target onset, whether these would be similarities in terms of constituent features (Shafer-Skelton & Brady, 2019), layout (Sanocki & Epstein, 1997) or 'spatial envelope' (Oliva & Torralba, 2001). However, in Experiment 2 the disordered nature of approach images meant targets were immediately preceded by an image more spatially distant to the destination, with the implication that the similarities in low-level information across the final leading image and the target scene were, by definition, reduced. Despite this, the effect of congruency remained.

While the preceding behavioural experiments provided clear support for an effect of expectations on gist processing, it was important to investigate the manner of such influence. As the previous iterations did not contain a control condition it remained open to question whether gist processing was being facilitated by congruous approaches, inhibited by incongruous approaches, or a mixture of both. Therefore, in Experiment 3 we introduced a 'No-context' condition where approach images were replaced with images of coloured patterns, allowing us to maintain the same trial structure while removing any pre-target semantic information. By doing so, this condition served as a measure of baseline performance, in terms of gist processing ability in the absence of antecedent contextual information, to which the congruency conditions could be compared. As predicted, we saw significantly increased categorisation ability on Congruous trials when compared to baseline performance, confirming that contextual information facilitated subsequent gist processing. Such a finding was expected due to the well-understood mechanisms of visual processing, where increased efficiency is achieved through utilisation of learned regularities to generate expectations as to the current environment (Chaumon et al., 2008; Fiser et al., 2016; Gregory, 1997; W. Li et al., 2004; Rock, 1997; Ullman, 1980).

While we did not make predictions as to whether a cost to processing on Incongruous trials would be apparent, Experiment 3 also demonstrated interference to gist extraction when participants were provided with inappropriate contextual information. This appears to be in agreement with previous gist processing research, where improbable scenes – i.e., those which contain unexpected features – were found by participants to be more difficult to extract meaning from, as compared to typical scenes (Greene et al., 2015). It is similarly in line with a recent investigation using object-scene pairs, which showed not only contextual facilitation of object processing but also interference to performance as a result of semantic violations within pairs (Lauer et al., 2020).

The exact mechanisms governing such interference remain open to interpretation, although it seems reasonable to suggest that the deficit in performance results from an attempt to match an unexpected bottom-up signal to an inappropriate, internally generated representation. This may be in the form of predictive coding mechanisms, whereby a significant disparity between expectations and ascending signal leads to prediction errors substantial enough to force reanalysis of the sensory input (e.g., Barrett & Simmons, 2015; Macpherson, 2017; Talsma, 2015). Alternatively, the Scene Perception and Event Comprehension Theory (SPECT; e.g., Loschky et al., 2018) proposes that an observer creates an internal current event model while progressing through a narrative, which represents their understanding of what is happening in that moment. Within this framework, significant changes in situational continuity initiate an automatic cognitive shifting towards creation of a new event model, and this operation is associated with distinct processing costs (Loschky et al., 2018). In the current study, therefore, reduced performance may have resulted from the disruption to processing due to the break in contextual continuity within Incongruous series. On the other hand, the case could be made that participants continued to search for an associative link when confronted with the lack of coherence within Incongruous trials, resulting in a protracted cognitive load that affected low-level perceptual processes (Afiki & Bar, 2020), or even that violations to predictions invoked increased encoding of the current scene-image while actively suppressing retrieval mechanisms (Sherman & Turk-Browne, 2020).

Turning to Experiment 4, changes to the neural signature help elucidate both the time-course and means by which the violation of expectations affects processing. Firstly, the contention that observer expectations were influencing gist processing is further strengthened by the display of changes to early ERPs, specifically the congruency-related amplitude differences in the P2 component. Changes in amplitude appearing so soon after target onset suggest an influence of top-down information while perceptual processing was still ongoing, similar to that proposed for object recognition (Bar, 2003; Fenske et al., 2006). While the P2 has previously been advanced as a marker for scene processing (Harel et al., 2016), there is debate as to whether this component is sensitive to top-down influence. For example, recent research found no top-down modulatory effect (Hansen et al., 2018), at least in relation to observer-based goals. Conversely, some forms of early higher-order influence have been implied, as changes to amplitude at ~200 ms post-stimulus have been observed with tasks involving the detection of objects within natural scenes, potentially reflecting decision-related activation (Thorpe et al., 1996; VanRullen & Thorpe, 2001), and tasks that manipulated the emotional nature of scene-images, argued as being driven by motivational systems (Schupp et al., 2006).

It is possible that different forms of top-down information are integrated at different temporal points, or simply that modulations to such early ERP components are more apparent under certain experimental designs than others. It may be that changes to the P2, found here, result from the use of antecedent information. Our use of approach images allowed for an expectation of the upcoming target category to be formed prior to its onset, meaning that this top-down information was available to facilitate processing from the moment the destination scene was presented. This is a clear departure from a task that involves a single image, whereby bottom-up input, perhaps in terms of low spatial frequency information (e.g., Bar et al. 2006), has to first be employed at scene onset to form expectations and only then is available as a tool for the ongoing evaluation of the incoming signal. As a result, it appears reasonable that a design eliciting expectations prior to target-onset would be able to more swiftly affect early ERP components such as the P2, as compared to single-image designs. Moreover, this likely better reflects processing during day-to-day life, where we constantly generate expectations as to the setting we are to encounter next (e.g., Bartlett, 1995).

Secondly, condition-related changes in the magnitude of the N400 component suggest the processing of Incongruous trials was affected by perceived semantic violations within those series.

Evidence has repeatedly indicated that the N400 is a neural marker for semantic processing, with increased negativity in this temporal window being related to difficulties with semantic integration (Barrett & Rugg, 1990; Kutas & Hillyard, 1980; McPherson & Holcomb, 1999). Our display of increased negativity at the N400 mirrors previous work related to the semantic violation of object-scene pairs. Such effects have been observed both when a scene is presented prior to target-object presentation, thereby allowing for *a priori* expectations as to the identity of the upcoming object to be formed (e.g., Demiral et al., 2012; Ganis & Kutas, 2003; Võ & Wolfe, 2013), as well as during simultaneous presentation of objects and scenes, meaning expectations as to object appropriateness cannot be formed prior to onset (e.g., Mudrik at al., 2010). However, while this previous work investigated violations to the semantic relationship between single scenes and their objects, our results show a comparable neural signature resulting from semantic violations between scenes. Additionally, that the experimental manipulation led to alterations in the N400 again makes it improbable that effects were due to confounds based on the repetition of low-level features. This component reflects a later stage of processing (e.g., S. Wang et al., 2017), its association with semantic integration, across multiple modalities, is much replicated (for a review, see Kutas & Federmeier, 2011), and its origins have been localised to higher-order brain regions such as those involved in semantic unification processes (e.g., Lau et al., 2008; L. Wang et al., 2012).

The processing of semantic information related to images, as opposed to text, has often been shown to elicit a more anterior negativity during this temporal window (e.g., Ganis et al., 1996; Holcomb & McPherson, 1994; Kutas et al., 2006), and our results reflect this. However, while both Frontal and Centro-parietal sites here displayed typical N400 effects, in terms of increased negativity for Incongruous trials, the pattern of amplitude changes are morphologically dissimilar across regions. Notably, the congruency-based amplitude changes in anterior sites began to emerge earlier (~200 ms) and were sustained for a far greater period of time (until at least 750 ms after target onset), with significantly more negative amplitudes for Incongruous trials across all three time-windows. There is minimal research regarding similar late effects at anterior sites, and explanations have ranged from it being related to late processes of semantic evaluation (Mudrik et al., 2014) or as attributable to reactivation of the prior context (Brothers et al., 2020).

On the other hand, investigations of pre-N400 negativity across frontal regions have been more frequent. In particular, the earlier emergence of effects at anterior compared to central sites has repeatedly been observed in object-scene research, leading to the proposition that this reflects a separate component, namely the N300 (Barrett & Rugg, 1990; Demiral et al., 2012; McPherson & Holcomb, 1999; Truman & Mudrik, 2018). This has been offered as reflecting context effects at a perceptual level (e.g., Schendan & Kutas, 2002; Mudrik et al., 2010), immediately prior to the semantic processing indicated by the subsequent N400. Furthermore, the N300 appears to be sensitive to alterations in global stimulus features rather than to low-level visual elements (e.g., Schendan & Kutas, 2007), and recent work has suggested it may be an index of perceptual hypothesis testing at a scale of whole scenes and objects, such as template matching routines based on perceptual structure (Kumar et al., 2020). It has also been put forward that components prior to the N300 may reflect predictive coding mechanisms in relation to expected low-level visual features (Kumar et al., 2020). However, distinguishable N300 effects have often not been forthcoming (e.g., Demiral et al., 2012; Ganis & Kutas, 2003) and this dissociation between the N300 and N400 is still debated (see, for example, Draschkow et al., 2018; Willems et al., 2008).

It is important to note that our early window of interest (175-250 ms) preceded the window typically used for investigating the N300 (e.g., Kumar et al., 2020; Lauer et al., 2020), and so our intention is not to comment directly on the debate surrounding that particular component. What we do assert, however, is that – if the N300 is taken as indexing perceptual, rather than higher-order, processing – then our early effects across anterior regions should be similarly categorised. In other words, due the early amplitude changes within Frontal sites as well as the alterations to the P2 discussed above, we suggest that expectations generated prior to target presentation were able to influence the extraction of scene gist at the level of perceptual processing. Predictions as to the category of an upcoming scene are likely to contain predictions not just of its identity, but also its expected perceptual features. At one level an observer may expect to see a beach, but on another level they may be expecting a certain spatial layout (Sanocki & Epstein, 1997) or specific form of spatial envelope (Oliva & Torralba, 2001), or a certain array of colours (Castelhano & Henderson, 2008; Gegenfurtner & Rieger, 2000), textures (Renninger & Malik, 2004), edge-based information (Walther & Shen, 2014) or other low-level features (Shafer-Skelton & Brady, 2019). However, whether the expectation-based violations to processing seen here were related to global properties or to lower-level information remains open to debate.

In terms of the P600, the changes observed here help further elucidate the potential mechanisms underlying the effect of expectations on scene processing. As with the N400, previous scene-related studies investigating this component have focused on object-scene pairs, but findings have proved inconsistent. For instance, Mudrik and colleagues (2010) found that positioning inappropriate objects in appropriate places (a semantic violation, such as a chessboard – rather than a baking tray – being placed into an oven) led to a more negative amplitude at 600 ms, compared to scenes containing appropriate objects. Võ and Wolfe (2013), alternatively, found no alterations to the P600 with similar object-scene semantic violations, but did find an increased P600 when appropriate objects were presented in a position considered to be atypical (such as a dishtowel on the floor, as opposed to hanging on a nearby towel rail). The authors proposed that these images created syntactic – rather than semantic – violations, as the objects contravened structural rules while remaining semantically congruous with their scenes. Thus, they reported the P600 as reflecting syntactic violations to scene processing (Võ & Wolfe, 2013), and so there appears to be a lack of consensus regarding the types of context-based violation that lead to changes in this component. However, it may be the case that these differing results reflect sensitivity to different methodological choices across studies, such as whether the scene is presented prior to the object or simultaneously with it, and whether the object is in a position of stable rest or being acted upon by agents within the image.

The current study, on the other hand, found alterations to the P600 without such violations to object location or appropriateness. It may be the case, therefore, that these similar ERP patterns are reflecting different phenomena, as research has shown the P600 to be associated with different forms of syntactic anomaly (Gouvea et al., 2010). Increased positivity at the P600 for inconsistent syntax between scenes and objects may be akin to grammatical errors in sentences (e.g., Hagoort et al., 1993), whereas the increased positivity seen here might be more similar to that elicited by 'garden path' sentences (e.g., Osterhout & Holcomb, 1992). Although containing no grammatical errors, progression through such sentences reaches a point where re-interpretation of expectations is necessary, through parsing the word-sequence in a different way. A similar form of violation may be responsible for our P600 pattern, whereby the progression of sequential approach images built an expectation in the observer – much like the expectation created during progression through the words of a sentence – until the final, incongruous destination disrupted the assumed end-point and resulted in an attempted re-evaluation of meaning. So, it may not be the case that the P600 is exclusively within the purview of violations to syntax, as it could also be a marker of the sudden need for reanalysis elicited by the disruption to an expected sequence. Such an explanation remains speculative, and further work surrounding the similarities in neural signatures across scene processing and language comprehension is certainly warranted. Both the N400 and P600 in scene processing appear somewhat analogous to those from language comprehensions studies and, while the specific forms of 'grammar' involved in these differing tasks likely diverge, a strong case can be made for the existence of commonalities (e.g., Võ et al., 2019).

This is an inchoate area of research and alternative interpretations as to the mechanisms responsible for such effects are possible. What seems a reasonable proposition, however, is that antecedent information allowed for expectations as to the category of the upcoming scene to be automatically generated. These expectations could be used to pre-activate internal representations or templates of expected-category exemplars which then become available for matching against the target scene once presented. Certainly, there are many separate conceptualisations of perceptual hypothesis testing which could be applied to our findings (see, for a review, Clark, 2013), although where such matching might take place within the visual processing stream remains open to debate. As we found congruency-based alterations to the ERP across all time-windows of interest, this recommends that it may be unwise to envisage a singular temporal or cortical point at which top-down predictions affect processing.

On one hand, our finding of early expectation-based amplitude changes indicates that predictions did influence feature extraction mechanisms. This is in line with recent findings pointing to a role of top-down feedback in the earliest stages of perceptual processing. Research using fMRI and multivariate pattern analysis has shown that expectations as to an upcoming, non-complex visual image are able to evoke stimulus templates in the primary visual cortex (Kok et al., 2012; Kok et al., 2014), and specifically within those deep layers proposed as being responsible for sending feedback to upstream regions (Aitken et al., 2020). Relatedly, higher-level cognitive factors have been shown to affect neurons in early sensory cortex (Lamme & Roelfsema, 2000), while representations based on semantic content have been shown to influence the extraction of elementary image features (Neri, 2014). It should be noted, though, that the semantic control of early sensory processing is still debated (see, for example, Carandini et al., 2005; Heeger et al., 1996).

On the other hand, our pattern of results reveals the violation of expectations also had an effect at a more advanced level of the processing stream. This contention is based on a number of factors. Firstly, the pattern of neural responses relating to the later components closely mirrors those long-associated with higher-order processing (e.g., Barrett & Rugg, 1990; Demiral et al., 2012; Ganis & Kutas, 2003; Kuperberg, 2007; Kutas & Hillyard, 1980; Lau et al., 2008; McPherson & Holcomb, 1999; L. Wang et al., 2012). Secondly, previous work showing the importance of expectations on gist processing found considerable deficits in the processing of improbable real-world scenes even when matched to probable scenes in terms of their low-level visual features, strongly implying that expectations were affecting processing at a stage somewhat beyond the level of initial feature extraction (Greene et al., 2015). Lastly, the superordinate category of targets in the current study (in terms of interior / exterior distinction) was maintained during Incongruous trials, ensuring that there was similarity across the low-level information present. For instance, a retail store scene may contain much of the same general structure or non-localised amplitude information as that of a supermarket, in terms of openness, roughness, etc. However, it should be noted that the exact level of similarities in low-level information across both superordinate and basic level scene categories is debated (e.g., Banno & Saiki, 2015; Fei-Fei et al., 2007; Gerhard et al., 2013; Loschky & Larson, 2008; Oliva & Torralba, 2001).

Taken together, it appears that *a priori* expectations had a broad effect across multiple stages of scene processing. Indeed, the concept of having a specific point of effect is perhaps only valid if a linear hierarchy of visual processing is accepted, as opposed to a cognitive network displaying abundant re-entrant connections (e.g., Boehler et al., 2008; Bullier, 2001; Koivisto et al., 2011). It may be, therefore, more germane to think of predictions of an upcoming scene as influencing manifold areas within the hierarchy simultaneously, whereby expectations set a cortical 'state' deemed appropriate for processing the predicted upcoming signal across the whole network (Gilbert & Li, 2013). Such an account could be considered as fitting within predictive coding frameworks (e.g., Friston, 2010; Friston & Kiebel, 2009; Rao & Ballard, 1999). Internal representations, activated through expectations as to the upcoming scene category, could allow for top-down predictions to propagate across processing areas (e.g., Lewis & Bastiaansen, 2015). As such, regions are informed by predictions based on the approach images, where reanalysis becomes necessary if the bottom-up signal is fundamentally at odds with what was expected (e.g., Talsma, 2015). In other words, a significant discord between predictions and input may create a substantial prediction error that crosses a pre-determined threshold or criterion, forcing both a major update of the internal model and reprocessing of the sensory signal (e.g., Barrett & Simmons, 2015; Macpherson, 2017; Talsma, 2015). Importantly, under such a model, *a priori* expectations may alter prediction error thresholds not only in early visual areas but also within higher-order processing regions (e.g., Hindy et al., 2016; Huang & Rao, 2011; Lewis & Bastiaansen, 2015; Summerfield et al., 2006), thus potentially resulting in a situation where difficulties in matching become apparent across separate levels of abstraction, such as at a perceptual and conceptual level.

The current study opens several important lines for further investigation. The finding that low-level visual information apparent at stimulus onset is not the only influence on gist processing asks the question as to what other sources of influence might exist. These could range from differing forms of top-down communication, such as an observer's goals, to the role of other sensory information, such as potential cross-modal facilitation through the parallel presentation of visual scenes and their related sounds. The design employed here attempted to better reflect scene processing outside the lab, but there are limits to how immersive a series of static images can be. To take this a stage further, leading images could be replaced by video clips of journeys or, better still, the incorporation of VR technology could embed participants within pre-determined environments. An important question that remains concerns the precise nature of the spatiotemporal dynamics of expectation effects. Other methodologies might be able to offer insights, such as the use of transcranial magnetic stimulation for interrupting re-entrant communication, or the application of dynamic causal modelling to tease apart the respective roles of top-down and bottom-up information. Finally, there appears to be clear similarities with how expectations affect the processing of meaning across both scenes and language. However, more work is needed to uncover the true extent of these commonalities, such as whether this demonstrates a single, amodal semantic system in operation.

This study moved away from the traditional RSVP approach – and towards more ecologically valid scenarios – through the incorporation of 'approach' images prior to target-scene onset, and the findings presented here reveal an important role of expectations during scene processing. Specifically, predictions as to an upcoming scene, generated automatically, were able to facilitate processing when valid, and interfere with processing when invalid. Furthermore, the use of both behavioural and neuroimaging methods adds to our understanding of the temporal dynamics of rapid scene processing and indicates an influence of top-down communication on the extraction of conceptual gist. This runs contrary to models supposing exclusive analysis of low-level information as determining the processing of scene-gist, such as 'forward sweep' frameworks. In addition, we also put forward a case that *a priori* expectations are able to affect gist processing at both a perceptual and conceptual level. While the precise mechanisms by which expectations affect the processing of scenes are still to be discovered, we argue that semantically relevant antecedent information may allow for scene-category templates to be pre-activated across various areas within the visual hierarchy. Future insights may be forthcoming from research concerning predictive coding, which offers a potential framework for the utilisation of top-down information within the brief timeframes where gist processing takes place.

## STUDY 2

The findings of the preceding collection of behavioural and ERP experiments can be taken as strong support for factors outside immediate visual stimulation having influence over gist processing. Specifically, Study 1 showed that observer expectations as to an upcoming scene could influence subsequent gist processing, and facilitate both the extraction of visual features and semantic integration. As a result, those findings bring into question frameworks suggesting real-world gist processing is driven solely by immediate visual stimulation (Itti et al., 1998; Potter et al., 2014; Rumelhart, 1970). An important pursuit within that set of experiments was an attempt to better replicate processing outside the laboratory, by using 'approach' images to mimic progressions through environments, and so better reflect how scenes are confronted in the real world. Study 2 attempts to extend this investigation of factors outside of current visual stimulation, while continuing to adhere to the principle of increased ecological validity, by studying whether simultaneously presented associative sounds can influence gist processing.

In addition, Study 2 helps address two important limitations of the previous experiments. First, the experimental manipulation employed previously took place prior to target-scene presentation, and so only provides evidence for additional information being beneficial for gist processing if received before a scene is encountered. To address this, in Study 2 additional information was presented simultaneously with the visual scene, rather than preceding it. Secondly, the previous experiments incorporated exclusively visual information, and so do not reveal whether separate forms of sensory stimulation might further facilitate gist processing. We live in a multisensory world, however, and typically experience the sceneries we encounter through more than a single modality. In Study 2, therefore, this additional information was presented in the form of sounds rather than images. As such, Study 2 should be considered as a companion and extension to the previous experiments, as it attempts to (i) provide further evidence for the facilitation of gist processing, and specifically whether gist can be facilitated when additional information is provided at the same time as a visual scene is encountered, and (ii) investigate whether the rapid processing of visual scenes can be influenced by cross-modal means, namely auditory stimulation.

Showing such simultaneous facilitation likely presents a greater challenge than the facilitation seen from eliciting pre-emptive expectations in Study 1. There is a great deal of evidence, across many different topics of investigation, that the visual system reduces cognitive load by using predictions to assist with information processing (e.g., Bartlett, 1995; Chaumon et al., 2008; Fiser et al., 2016; Gregory, 1997; W. Li et al., 2004; Rauss et al., 2011; Rock, 1997; Ullman, 1980). As previously discussed, the findings from Study 1 suggest that the processing of scene gist appears to follow these same rules, whereby providing an observer with additional information leads to more accurate predictions being formed as to an upcoming stimulus, and thus ultimately more efficient subsequent processing. However, there is still much debate as to whether simultaneous information can affect gist processing, with various research finding scene perception to be unaffected by interference from dual-task conditions (Fei-Fei et al., 2005; F. Li et al., 2002; Rousselet et al., 2002). Indeed, it has been claimed that semantic information is extracted from a scene in timeframes too brief for focussed attention to have a significant involvement in gist processing (e.g., Biederman, 1972; F. Li et al., 2002; Potter, 1975; Thorpe et al., 1996), taken as support for models positing awareness without attention (e.g., Lamme, 2004; Tononi & Koch, 2008). Conversely, however, other work has suggested that scene perception can be inhibited under certain task conditions (e.g., Evans & Treisman, 2005; Walker et al., 2008), and that gist extraction does indeed require the engagement of attention (Cohen et al., 2011; Mack & Clarke, 2012; Wolfe et al., 2011). So, there appears to be some evidence that simultaneous information can impair gist processing, but the possibility remains that such processing is too rapid to be facilitated by such information. In other words, gist processing might operate with a level of performance at ceiling, whereby ability can be interfered with but not enhanced. The current study aims to address this question, by investigating whether simultaneously-presented sources of information lead to improvements to gist extraction.

The second challenge is related to the modality with which information is presented. Only visual information was provided to participants in Study 1, and so only provides evidence that visual information can help with the processing of subsequent visual information. However, as set out below, there is evidence to predict that cross-modal facilitation of processing might be apparent. For example, simultaneously presented consistent sounds have been shown to affect the visual processing of objects, facilitating learning (Barenholtz et al., 2014) and recognition (Giard & Peronnet, 1999; Molholm et al., 2004), and biasing interpretation (Smith et al., 2007). So, previous work has been suggestive of an enhancement to visual stimulus processing through multimodal means, but to date few attempts have been made to extend this investigation of audio-visual interaction beyond objects to the study of scene processing (although, see Rummukainen & Mendonca, 2016; Rummukainen et al., 2014). Such a lack of investigation appears unwise when we consider that our experience of the world is multisensory, with separate senses providing complimentary information about our surrounding environment. While walking, we may hear that we are approaching a busy road before we see it, and while waiting to cross a street we may know it is safe to do so because of a beeping noise at the traffic lights. This is an issue that previous scene perception research has failed to tackle, and our current understanding of real-world scene processing is based almost exclusively on studies operating within visual frameworks (Epstein & Baker, 2019; Groen et al., 2017; Henderson & Hollingworth, 1999; Malcolm et al., 2016; Wolfe et al., 2011).

Therefore, Study 2 aims to address these two gaps in knowledge, further investigating potential influences on scene understanding that lie outside the traditional focus on bottom-up visual stimulation. To do this, a visual search paradigm that relied on rapidly understanding complex real-world scenes was employed, while the semantic congruency between a target object in the scene and a simultaneous sound was manipulated. Importantly, this sound was uninformative in terms of spatial location, and so is distinct from audio-visual studies related to attention orienting through spatial cueing (e.g., McDonald et al., 2000; Schmitt et al., 2000; Spence & Driver, 1997).

Visual search paradigms have previously been shown as an effective method for investigating scene processing (Malcolm & Henderson, 2010; Neider & Zelinsky, 2006; Spotorno et al., 2014, 2015). This is due to objects often being associated with certain scene categories (e.g., microwaves in kitchens, buoys in harbours), as well as specific locations within those scenes (e.g., microwaves on counters, buoys in the water), meaning a scene's context biases where we expect items to be. In other words, scenes have structural and semantic rules which govern our expectations as to the presence and location of objects within them (Malcolm et al., 2016). For example, if trying to locate a kettle we are likely to narrow our initial search to the kitchen counter, but if searching for a fire extinguisher we are likely to move our gaze to walls close to doorways. Therefore, a strong relationship exists between understanding the gist of the scene and our ability to search for items within that scene: we derive gist within just a short glimpse (Biederman et al., 1974; Castelhano & Henderson, 2007; Fei-Fei et al., 2007; Potter, 1975; Potter et al., 2014) and then rapidly use this information to direct our gaze to probable target locations, even within the first eye movement (Eckstein et al., 2006; Neider & Zelinsky, 2006; Spotorno et al., 2014; Torralba et al., 2006). It is reasonable to infer, therefore, that changes to the speed or efficiency of gist processing will have a subsequent, and relative, effect on an observer's ability to search for objects within that scene.

**Cross-modal scene processing**

It has long been understood that information from auditory and visual sensory systems interact to affect perception (e.g., Campbell & Dodd, 1980; McGurk & MacDonald, 1976; Sumby & Pollack, 1954), and there are clear behavioural advantages to such multisensory perception (e.g., Newell, 2004). Anatomical evidence has shown that V1 receives direct projections from auditory cortex (Falchier et al., 2002), that activity in this region is enhanced when presented with audio-visual – compared to just visual – stimulation (Romei et al., 2007), and that multisensory neurons within the superior colliculus are activated through the simultaneous presentation of pictures and sounds (Meredith et al., 1987). Cross-modal investigations have largely centred on studies of object recognition (e.g., Adams & Janata, 2002; Beauchamp et al., 2004, Laurienti et al., 2003), and have shown both the convergence of processing within multisensory cortices as well as direct cross-modal interaction between unisensory regions (for a review, see Amedi et al., 2005). Investigation of the audio-visual processing of natural scenes, on the other hand, has been sparse. Some recent exploratory work has attempted to address this multisensory nature of scenes, by investigating the processing of immersive audio-visual environments (Rummukainen & Mendonca, 2016; Rummukainen et al., 2014). While scene discrimination was not found to be significantly enhanced in the bimodal, compared to unimodal, condition, there did appear to be benefits of multimodal perception in making judgements about certain scenes characteristics (Rummukainen & Mendonca, 2016). Such findings of separate contributions to processing from the audio and visual modalities further suggest that unimodal investigation of real-world scenes is necessarily limited. Perhaps more promising has been work involving the playing of object sounds within complex scenes. As stated above, certain scenes become associated with certain sounds due to the objects that typically inhabit them, and so it would perhaps be surprising if these associated sounds were not able to influence the processing of one's environment. Put simply, if the visual characteristics of an object can be used to help disambiguate the scene category in which it is embedded (Brandman & Peelen, 2018), it seems reasonable to expect that the auditory characteristics of that same object could be similarly facilitative.

As discussed previously, scene gist can be thought of as an incomplete summary of a scene percept, containing initial information about spatial characteristics such as openness and navigability (Greene & Oliva, 2009) and central objects (Davenport & Potter, 2004; Fei-Fei et al., 2007), which can be matched against stored representations to guide recognition (Greene et al., 2014). This suggests two potential mechanisms by which target congruent audio information could facilitate search performance. The first is that semantically congruent target sounds could facilitate the processing of a scene's visual information. There is, for example, much evidence demonstrating that object identification can be facilitated if congruent audio information is also present, leading to both quicker reaction times in visual search studies (e.g., Iordanescu et al., 2008) as well as higher accuracy rates (Laurienti et al., 2004; Molholm, 2004). Furthermore, such facilitation has been demonstrated when the accompanying audio information is presented either before the visual object's appearance (Schneider et al., 2008), with simultaneous onsets (Molholm, 2004), or up to 300ms after the image (Chen & Spence, 2010). Correspondingly, recent work has suggested that similar facilitation is apparent in relation to complex natural environments – in terms of quicker search times within scenes when accompanied by corresponding target object sounds – both within static images (Mahzouni, 2019) and video clips (Kvasova et al., 2019). It may be, therefore, that just as auditory influences have been shown to enhance visual object processing early within the information-processing stream, potentially at the feature level of representation (Molholm, 2004), similarly early interaction between sensory inputs might be typical for scene processing. Take, for example, a bedroom scene in which participants have been asked to locate an alarm clock. A very brief visual presentation of the scene, accompanied by a sound consistent with a bedroom (an alarm clock bleeping), may allow a participant to gather more visual information about the layout of the scene (i.e., through a cross-modal facilitation of gist extraction), and use this additional information to guide search within the scene more efficiently.

An alternative explanation would be that such additional information is beneficial in terms of allowing for a more rapid matching of gist against stored representations, reducing uncertainty and resulting in more accurate processing. For example, processing a scene's gist depends on matching the percept against stored representations of typically occurring patterns developed over a lifetime of experiences (e.g., if there is a couch and a TV in an indoor space of a particular size, it will generally be a living room). When a scene is atypical (e.g., a boulder in a living room) gist processing is delayed (Greene et al., 2015), strongly suggesting that processing is not entirely driven by external properties, but is dependent on how easily those properties can be matched to prior experiences. Furthermore, there is robust evidence from behavioural studies for the interplay between scene and object processing (e.g., Biederman et al., 1982; Boyce & Pollatsek, 1992; Joubert et al., 2007;

Munneke et al., 2013), and recent research has demonstrated that contextual object cues can be used to facilitate scene layout representations in scene-selective regions (Brandman & Peelen, 2018). It seems reasonable to suggest, therefore, that such facilitation is not necessarily modality-specific, and could be achieved if contextual object cues were provided aurally. The inherent association between scene categories and the characteristic sounds of the objects that typically inhabit them (e.g., a city street has car sounds), means that consistent sounds may well provide additional information to be matched against stored representations, reducing ambiguity and potentially speeding up gist processing. It could also be the case, in line with the findings of Study 1, that any facilitation to gist processing is not exclusively based on one mechanism or another. Namely, associative sounds may allow for both the enhanced extraction of visual properties as well as improvements to representation matching.

Finally, there is further reason to believe that gist processing would be biased by real-world sound information due to the staggered timeline with which visual properties yielding gist information are processed and integrated (although, for debate in relation to classical hierarchical models for scene vision, see Groen et al., 2017). For instance, a scene's spatial characteristics are processed very early on (Greene & Oliva, 2009), while colour information is integrated at a later epoch (Castelhano & Henderson, 2008). Similarly, some evidence suggests that there is a centre-to-surround spatiotemporal integration process (Larson et al., 2014). As sound information has been repeatedly shown to affect very early stages of object processing in the sensory-specific cortical structures (Foxe et al., 2000; Giard & Peronnet, 1999; Schroeder & Foxe, 2005; Vetter et al., 2014) before the conceptual realisation stage (Iordanescu et al., 2008), it suggests that sound can similarly affect how visual properties are selected before a category meaning is determined.

### The current study

As is evident from the above, current research has arrived at a juncture. There is strong reason to surmise that semantic sound information would influence the processing of scenes, such as the repeated finding of audio facilitation of visual processing more generally (e.g., Beauchamp et al., 2004; Campbell & Dodd, 1980; McGurk & MacDonald, 1976), but with very little concrete investigation of this. Furthermore, research which has focussed on the audio-visual processing of scenes has not specifically examined gist extraction, and instead used visual stimuli with lengthy exposures (e.g., Kvasova et al., 2019), making determinations as to the temporal point of potential facilitation problematic. To address this inadvertence, Study 2 aimed to directly investigate the potential cross-modal facilitation to gist processing using the flash moving-window paradigm, previously shown to be a technique well-suited for assessing scene gist processing (see, for example, Castelhano & Henderson, 2007; Võ & Henderson, 2010, 2011). Participants were given the name of an object needing to be located, prior to a very brief presentation of a scene image on the monitor. This initial 'preview' was accompanied by either congruous, incongruous, or no auditory information pertaining to the target object. For example, in terms of a Congruous trial, if a car was to be found in a scene an audio file containing the sound of a running car engine began to play simultaneously with the onset of the scene preview. After a short delay the sound terminated and the scene image was presented again, now with a gaze-contingent window masking peripheral information, at which point the participant began their search for the target object. All sounds were played through speakers directly on either side of the monitor, and so provided no spatial information as to the location of the object within the scene. Any changes in performance across sound conditions were not, therefore, due to auditory cues directly highlighting an object's position. The 'No sound' condition served as a baseline, mimicking typical visual scene literature, and so differences in performance evident between this and either of the 'Sound' conditions would imply that our current understanding of scene processing is necessarily limited.

While recent work on object search within complex scenes has been informative, these previous studies have assessed performance exclusively using participant response times (Kvasova et al., 2019; Mahzouni, 2019), making impossible a disentanglement of whether consistent sounds were facilitating scene processing, object processing, identification processes or decision making.

Therefore, eye tracking methods were incorporated here to investigate task-related performance beyond reliance on manual response time. It was expected that Congruent sound conditions would lead to improved search ability when compared to No Sound conditions, due to previous work suggesting quicker reaction times under such conditions (Kvasova et al., 2019; Mahzouni, 2019). Furthermore, it was predicted that this would be based on more efficient gist processing, leading to quicker locating of target objects and quicker task completion. In addition, the time spent fixating target objects was also measured, to ensure that any potential changes to task performance were not simply the result of enhanced object-identification processes at fixation.

The presentation duration of the preview scene was also manipulated to further investigate the time-course of this potential multimodal advantage to gist processing. For example, an advantage for congruous scene-audio pairs at 50ms, but not at 100ms, would suggest this benefit is only apparent in circumstances where visual information is particularly limited (i.e., that at longer durations sufficient visual features are processed to render additional auditory information redundant). This would, in turn, highlight the importance of other sources of stimulation to the earliest instances of gist processing to help disambiguate meaning. In addition, the inclusion of a condition containing no visual preview served as a timing baseline, with the expectation that there would be no difference between the separate audio conditions when no preview image was displayed.

Finally, trials with incongruent sound were also used (e.g., searching for a cashpoint in a bank interior, with the sound of a lawn mower), but it was expected that this Incongruent sound condition would display similar performance levels in comparison to the No sound condition. This was due to recent studies of audio-visual processing providing no evidence for interference through presentation of incongruous sounds when searching through object arrays (Iordanescu et al., 2008; Iordanescu et al., 2010) or within complex natural scenes (Kvasova, 2019; Mahzouni, 2019). Similarly, while interference to gist processing has previously been demonstrated (e.g., Evans &

Treisman, 2005; Walker et al., 2008), this has often not been the case (Fei-Fei et al., 2005; F. Li et al., 2002; Rousselet et al., 2002). Indeed, a strong case has been made that the great efficiency with which gist extraction operates makes interference to this process unlikely under anything but especially taxing situations of divided attention (such as employed by Cohen et al., 2011), a level of distraction unlikely to be reached here by the playing of incongruous sounds. However, the potential for such interference remains, as object recognition has previously been shown to be degraded in the presence of incongruent auditory stimuli (Laurienti et al., 2004). So, if unhelpful information in the form of unrelated sounds is processed to a high level of understanding it could potentially lead to a deficit for effective gist processing, thus reducing performance ability. Therefore, by determining whether inconsistent sounds provide similar or inhibitory effects compared to no sound, the current study aimed to elucidate whether unhelpful information is rapidly discarded from processing or is automatically processed to a high level of understanding prior to causing an inhibitory response.

## Experiment 5

### Design

A 3 x 3 mixed design was used. Sound condition was a within-subjects variable (Congruous sound; Incongruous sound; No sound), and Preview Duration was a between-subjects variable (100; 50; 0 ms). The 0 ms condition acted as a control for the timing manipulation, whereby participants were provided with no preview of the scene prior to the start of search. Dependent variables were the time taken for participants to respond by pressing the keyboard to signify the target object had been found (Overall Response Time), the time taken to first fixate on the target object (First Fixation Time), and the total duration of fixations on the target object (Target Dwell).

Overall Response Time was chosen as a general measure of performance on the task, with shorter response times reflecting better search ability, and so replicates the measure used in previous multimodal scene-related research (e.g., Kvasova et al., 2019; Mahzouni, 2019). However, better performance may be the reflection of separate cognitive functions – such as efficiency of template matching or speed of decision making – thus leading to the inclusion of the two additional dependent variables. First Fixation Time is considered a more sensitive measure of search performance (Malcolm & Henderson, 2009), as it records the time taken before a participant locates and fixes their gaze on the target object, and so is not confounded by the time taken to decide on a physical response. Target Dwell reflects the time taken for the target object to be identified once fixated, and so is a measure of recognition speed and decision making. Taken together, therefore, these measures allow for investigation as to whether alterations in the duration of the preview scene and the consistency of sounds lead to changes in performance, and also for delineation regarding whether any such changes are related to improvements in participants' ability to locate objects due to enhanced scene gist processing.

Mixed ANOVAs were followed up with planned comparisons and paired-samples t-tests where appropriate, in order to investigate more fully the influence of both experimental manipulations on search performance.

**Participants**

An *a priori* power analysis (*G*Power*; Faul et al., 2009) suggested an estimated sample size of 34 participants per Preview Duration was required for medium sized effects. Aiming for equal numbers in each of the three Latin square versions of the experiment, our intention was to recruit a total of 108 participants, 36 for each of the Preview Duration conditions. However, testing was halted prior to completion, due to restrictions imposed by the Covid-19 pandemic. Thus, one hundred and five participants took part in Experiment 5 ($M_{\text{age}}$ = 20.55, $SD_{\text{age}}$ = 4.58; 86 Females, 19 Males; 95 Right-handed, 10 Left-handed). They were students and staff recruited through the University of East Anglia's research pool, as well as local residents from a volunteer research panel, who received either a small payment or course credits for participating. All reported having normal or corrected-to-normal vision and hearing. All experiments in Study 2 were approved by the Ethics

Committee at the University of East Anglia's School of Psychology, and all participants provided written informed consent prior to taking part in the study. There were 35 participants included in the 0 ms condition, 34 in the 50 ms condition and 36 in the 100 ms condition.

## Stimuli

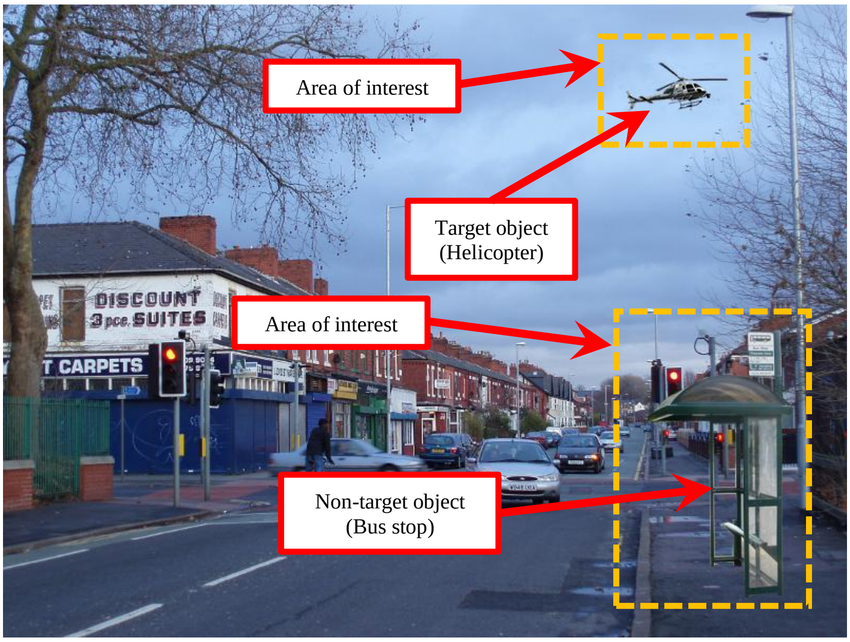
### *Visual Stimuli*

A set of 72 real-world scene images were collected from Google Images. These included a mixture of both interior and exterior locations, and consisted of familiar categories, such as bedrooms, parks, etc. All scene images were 800x600 pixels, and were presented in full colour. In addition, a set of 72 images of recognisable objects were also obtained from the internet to be used as targets. Each image was of a different object, and each was of a different size. Object width ranged from 0.34 to 9.13 degrees of visual angle (*M* = 2.72, *SD* = 1.46), and height from 0.17 to 10.17 (*M* = 2.61, *SD* = 1.78). Each scene image was then adapted by having one target positioned within it using Adobe Photoshop CC 2019 (version 20). The criteria used when doing so was that each object should be within a semantically related scene category (i.e., within a scene where it would typically be used or located), that it should be in a position and orientation that would be expected and that followed structural rules, that it was not occluded in any way, and that it was of an appropriate perceived size in terms of the perspective constraints of the scene.

Furthermore, target objects were placed at varying distances from the centre of the scene image, while ensuring that these were not visible within the gaze-contingent window at search-screen onset. The foveal window had a radius of 3.72 degrees of visual angle, while the distance of targets from the centre of the screen ranged from 7.04 to 19.50 degrees (*M* = 13.81, *SD* = 2.95). All selected object images had transparent backgrounds and no visible outer border. These criteria, therefore, ensured that objects appeared to be embedded within the scene images that they were paired with, so as to mimic as closely as possible the process of searching for an object within a real environment. See Figure 14 for an example of a trial image.

**Figure 14**

*Example of a Trial Image Used in Study 2*



*Note.* The original image is that of a street scene. To this, a target image (a helicopter in this instance) and a non-target image (a bus stop) were positioned within the scene. Objects were positioned in accordance with structural rules, and in locations where they might typically be found. The two area-of-interest boxes were not visible to participants. In Experiment 5 the preview scenes included the target and non-target objects, but this was not the case in Experiment 6. Scene image taken from www.photoeverywhere.co.uk.

To safeguard against low-level visual properties potentially confounding search performance, a set of 72 images of non-target objects was also obtained. A different non-target object was placed within each scene, following the same criteria as above, to safeguard against participants simply looking for photoshopped objects. These non-target images did not form part of the task, and participants were not informed of their presence. A rectangular interest area was drawn around each of the target and non-target images, not visible to participants, allowing for the software to automatically record the frequency and duration of fixations on these objects. The width of the interest area around targets ranged from 3.25 to 14.32 degrees of visual angle (*M* = 5.46, *SD* =

1.82), and height from 2.48 to 15.14 (*M* = 5.12, *SD* = 2.33). Therefore, across trials target and non-target objects varied in size, horizontal and vertical coordinates, and whether they were within the foreground or background of scenes. See Appendix E for a list of target and non-target object pairings.

***Auditory Stimuli***

A corresponding sound for each of the target objects was obtained from the internet (www.freesounds.org). These were characteristic sounds that typically represented the object, such as the sound of strumming for a guitar target, ticking for a wall clock, and so forth. See Appendix E for a list of the sound and object pairings. Each sound lasted 3000 ms and was played in its entirety. Sounds were played through speakers placed at the sides of the display monitor, and so gave no indication as to the location of the target within the scene. Object sounds were played in isolation, with no background noises. All sounds were root mean squared to between -18 and -20 dB (M = -18.68, SD = 0.52). See Appendix F for a list of individual sound levels. The first and last 10 ms of each sound was also enveloped, to remove any unintended clicking noise associated with the onset or offset of the audio file.

Previous research has provided conflicting evidence as to the appropriate timing of the auditory stimulus in cross-modal investigations. Sounds have been shown to enhance the visual processing of objects if presented before (Schneider et al., 2008), simultaneously with (Molholm, 2004), or after the image (Chen & Spence, 2010). In terms of scene processing, cross-modal effects have been demonstrated when the onset of sound precedes that of the image by 100 ms (Kvasova et al., 2019), but also with synchronous onsets (Mahzouni, 2019). Therefore, the current study chose to employ simultaneous audio-visual onsets, in line with its overarching aim to best represent typical functioning in everyday scenarios.

**Procedure**

The experiment was programmed using Experiment Builder. Images and audio were presented using a PC running Windows 7, with Logitech S-150 stereo speakers and a BenQ XL24IIT monitor running at 100Hz (800 x 600 resolution). Participants viewed the monitor from a distance of 60 cm, with scene stimuli subtending a Width x Height visual angle of 36.9° x 28.1°, maintained through use of a chin rest (with additional forehead support to minimise head movements). A keyboard was placed on the desk directly in front of the participant and within comfortable reach. The experimenter sat at a separate table to the side, monitoring participants' eye movements and search performance in real time.

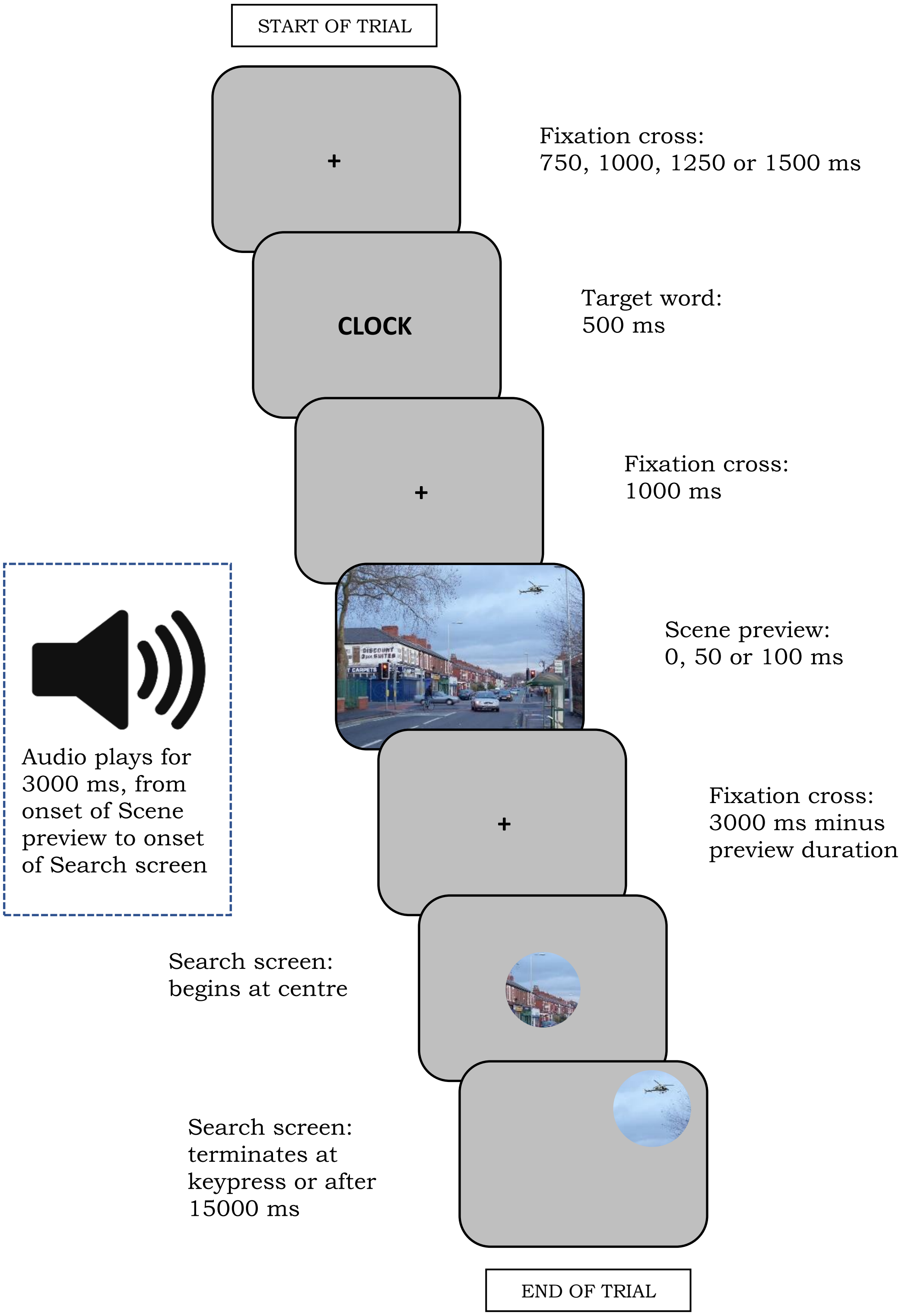
Prior to the start of the experiment, participants were given verbal instructions describing the task, read an information sheet and provided signed consent. They also read a list of the target objects presented in the study, and were given the opportunity to ask for clarification if unsure about the physical appearance of any of these items. The chin rest was adjusted to a comfortable height where necessary. Gaze position was calibrated using a 9-point calibration routine, followed by a 9-point validation routine, with additional drift correction performed at the beginning of each trial.

Each trial began with a black fixation cross displayed at the centre of a grey screen (see Figure 15 for a schematic of the experimental protocol). The duration of this screen varied pseudo-randomly and was displayed for either 750, 1000, 1250 or 1500 ms. A written word in black text was then presented at the centre of the screen for 500 ms, indicating the object needing to be found during that trial. The blank screen with fixation cross was then displayed for a further 1000 ms. Depending on which Preview Duration version the participant was allotted to, a scene image was then presented for either 100, 50 or 0 ms (i.e., no preview), before being replaced by another blank screen with fixation cross. Also dependent on the Preview Duration condition being sat, this blank screen lasted for either 3000, 2950 or 2900 ms (so that, irrespective of condition, there was always 3000 ms between the disappearance of the target word and appearance of the search screen).

In addition, an audio file commenced simultaneously with the onset of the scene preview

**Figure 15**

*Schematic Representation of the Experimental Protocol Used in Study 2*



*Note.* The comparative size of the gaze-contingent window has been enlarged for display purposes. The target and non-target objects are included in the scene preview, as would be displayed in Experiment 5.

image and played for 3000 ms. The sound would either be consistent or inconsistent with the object needing to be found, or no sound would be played. When the sound finished, the same scene image was presented again, although now obscured except for a circular area 3.72° of visual angle (radius) around the point of fixation. This gaze-contingent window moved in accordance with the movement of the participant's eyes as they searched the scene, thus eliminating the availability of peripheral information.

A trial was completed once the participant pressed the spacebar to signify that the designated target object had been found, or through timing out after 15 seconds of search. Participants sat a total of 72 trials, equally divided amongst the three audio conditions. A Latin square design was employed and incorporated three versions of the experiment, meaning that, over the course of the experiment as a whole, each scene was presented as part of the Congruous, Incongruous and No sound conditions. These versions were cycled through for each new participant, separated by Preview Duration condition. Trial order was randomised for each participant, and no scene images, object images or sound files were repeated at any point during the task. The experiment lasted approximately 30 minutes, and participants were informed that they could take a break between trials at any point. All participants were fully debriefed on completion.

**Results**

Certain criteria were used to process the data. First, only trials where the target object was fixated were included in the analysis (6.48% of trials removed). Secondly, any instances where the target was fixated at the start of the trial were discarded (i.e., the participant had moved their gaze from the central fixation cross prior to the onset of the search screen). This led to removal of a further 3.70% of trials. Thirdly, trials were only included if the participant had made a response within the 15 second presentation duration, even if the target had been fixated, leading to a further 0.61% of trials being removed. Finally, only trials where the participant pressed the spacebar while fixating the target, or within 300 ms of moving their gaze away from the target, were included. This was to ensure that the target object had been recognised as such, rather than a participant’s search continuing after the target had been fixated. This led to removal of a further 1.24% of trials. Overall, therefore, the analyses included 88.40% of the total number of trials sat in Experiment 5.

After commencement of testing, an issue with two of the stimuli became apparent. First, the ‘electric razor’ audio file was corrupted and so failed to play when expected. This was intended to play during a preview of either a bathroom scene (Congruous condition) or a street scene with the target object ‘bus’ (Incongruous condition), depending on the version of the experiment the participant sat. Therefore, both of these trials were removed from the analysis for all participants. Secondly, one scene depicted a recording studio with a microphone serving as the target object. However, it was later discovered that there was a second microphone in the background of the scene. Therefore, we removed both this trial as well as the ‘smoke alarm’ trial (the audio paired with the microphone image on Incongruous trials) from the analysis. Thus, in total we removed four trials, with the subsequent analysis being conducted on the remaining 68 trials (94%) of the experiment.

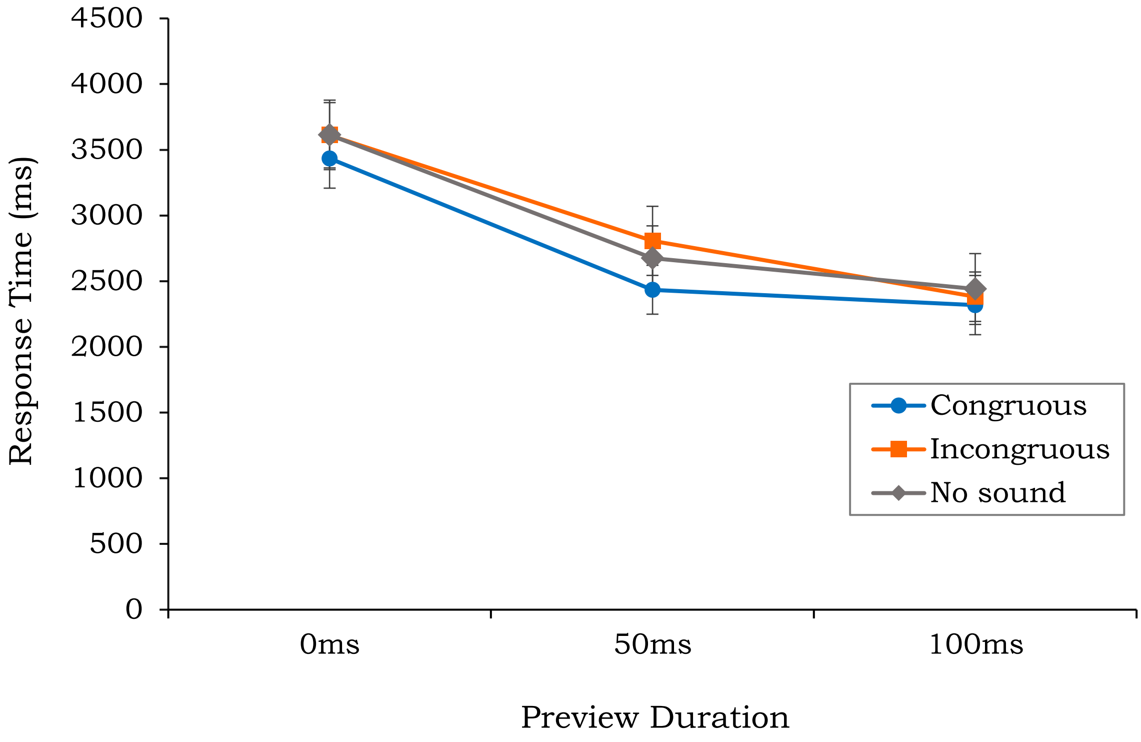
For each dependent variable, trials outside three standard deviations of the mean for that participant on a given Sound condition were removed. Normality was assessed by dividing skew and kurtosis values by their standard errors, with z-scores above 1.96 (corresponding to an alpha level of 0.05) judged as signifying a non-normal distribution. The distribution of the data was found to be positively skewed and leptokurtic for multiple conditions across separate dependent variables of both experiments. We first removed any participants with a mean score outside 3 standard deviations of the grand mean for that Sound condition, separately for each dependent variable. However, skewed distributions remained evident across both experiments, and so a logarithmic transformation (base 10) was applied to the entire Study 2 dataset. After transformation, no issues relating to normality or sphericity were found within the data. Therefore, parametric tests were chosen for the analysis. Significant effects within the mixed ANOVAs were followed up using Bonferroni-corrected pairwise comparisons.

***Overall response time***

Two participants scored outside of 3 standard deviations of the grand mean, and so were removed (one from the 0 ms condition, and one from the 50 ms condition). See Figure 16.

**Figure 16**

*Mean Response Time for Congruous, Incongruous and No Sound Trials as a Function of Scene Preview Duration*



*Note.* Error bars represent 95% CIs.

A 3 x 3 mixed ANOVA showed no Sound x Preview Duration interaction, *F*(4, 200) = 0.71, *p* = .588, ηp$^2$ = .01. There was, however, a main effect of Preview Duration, *F*(2, 100) = 40.45, *p* < .001, ηp$^2$ = .45. As expected, participants were on average significantly slower to complete trials in the 0 ms condition (*M* = 3552 ms) compared to either the 50 ms condition (*M* = 2639 ms), *p* < .001 (Mean difference = 913 ms; 95% CI [582 ms, 1244 ms]) or the 100 ms condition (*M* = 2380 ms), *p* < .001 (Mean difference = 1172 ms; 95% CI [848 ms, 1496 ms]). The difference between the 50 ms and 100 ms conditions only approached significance, *p* = .070 (Mean difference = 259 ms; 95% CI [-67 ms, 586 ms]). There was also a main effect of Sound, *F*(2, 200) = 4.96, *p* = .008, ηp$^2$ = .05. Pairwise comparisons showed the difference between Congruous sounds (*M* = 2729 ms) and Incongruous sounds (*M* = 2933 ms) was significant, *p* = .008 (Mean difference = -204 ms; 95% CI [-361 ms, -48 ms]), with shorter completion times on Congruous trials. The difference between Congruous sounds and No sounds (*M* = 2910 ms) only approached significance, *p* = .052 (Mean difference = -181 ms; 95% CI [-335 ms, -27 ms]), and there was no difference for Incongruous sounds compared to No sounds (*p* = 1). See Table 1 for mean scores across the two experiments.

**Table 1**

*Mean Times and Standard Deviations for Each Variable Across Experiment 5 and Experiment 6, Collapsed Across Preview Duration*

| Variable | Experiment 5 | | | Experiment 6 | | |
|---|---|---|---|---|---|---|
| | Congruous | Incongruous | No Sound | Congruous | Incongruous | No Sound |
| Overall response time | 2724 (791) | 2924 (841) | 2903 (902) | 3310 (953) | 3464 (935) | 3328 (859) |
| Time of first fixation on target | 2109 (777) | 2295 (783) | 2244 (803) | 2718 (982) | 2858 (957) | 2728 (900) |
| Total dwell time on target | 510 (116) | 517 (117) | 513 (116) | 500 (116) | 509 (112) | 519 (136) |

*Note*. All values in milliseconds. Standard deviations in parentheses.

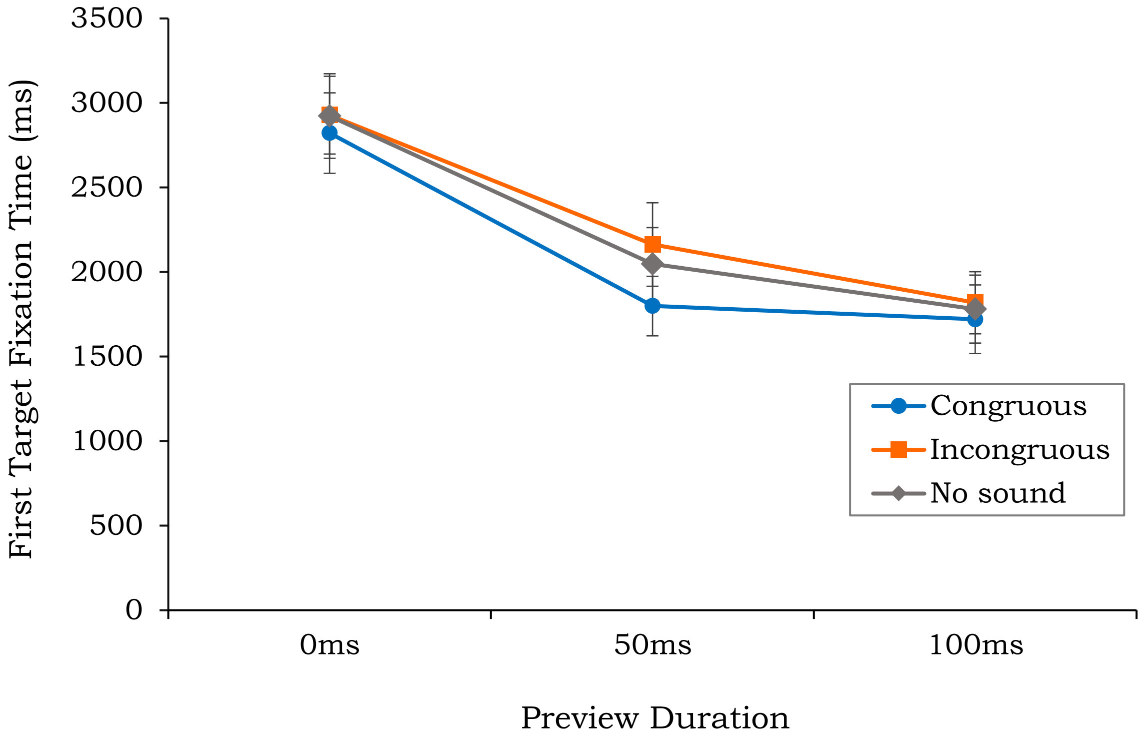
### ***Time taken to first fixate on target object***

Two participants scored outside of 3 standard deviations of the grand mean, and so were removed (one from the 0 ms condition, and one from the 50 ms condition).

A 3 x 3 mixed ANOVA showed no Sound x Preview Duration interaction, *F*(4, 200) = 0.99, *p* = .415, ηp$^2$ = .02. There was, however, a main effect of Preview Duration, *F*(2, 100) = 43.95, *p* < .001, ηp$^2$ = .47. As expected, participants were on average significantly slower to first fixate on the target object in the 0 ms condition (*M* = 2891 ms) compared to either the 50 ms condition (*M* = 2003 ms), *p* < .001 (Mean difference = 888 ms; 95% CI [587 ms, 1188 ms]) or the 100 ms condition (*M* = 1774 ms), *p* < .001 (Mean difference = 1117 ms; 95% CI [823 ms, 1412 ms]). The difference between the 50 ms and 100 ms conditions only approached significance, *p* = .077 (Mean difference = 230 ms; 95% CI [-67 ms, 526 ms]). There was also a main effect of Sound, *F*(2, 200) = 5.44, *p* = .005, ηp$^2$ = .05. Pairwise comparisons showed the difference between Congruous sounds (*M* = 2114 ms) and

**Figure 17**

*Mean Time to First Fixate on Target for Congruous, Incongruous and No Sound Trials as a Function of Scene Preview Duration*



*Note.* Error bars represent 95% CIs.

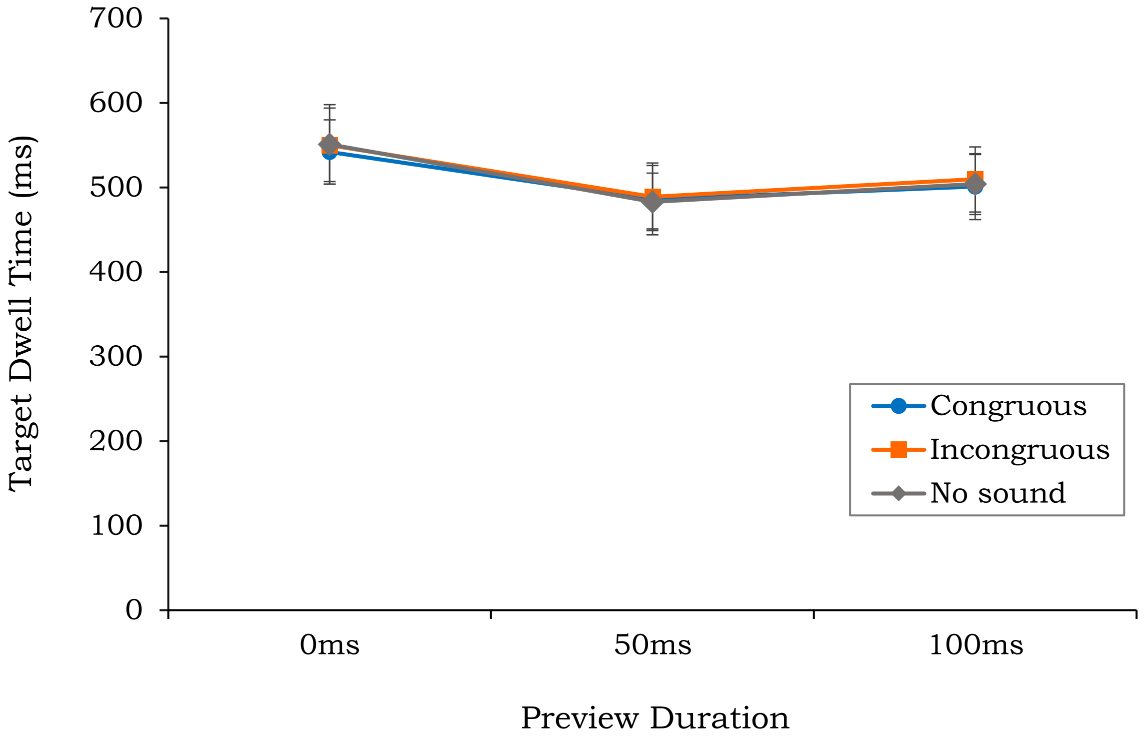
Incongruous sounds (*M* = 2303 ms) was significant, *p* = .003 (Mean difference = -189 ms; 95% CI [-331 ms, -47 ms]), with shorter completion times on Congruous trials. The difference between Congruous sounds and No sounds (*M* = 2251 ms) only approached significance, *p* = .093 (Mean difference = -137 ms; 95% CI [-287 ms, 14 ms]), and there was no difference for Incongruous sounds compared to No sounds (*p* = 1). See Figure 17.

***Total time spent fixating target (Target Dwell)***

Two participants scored outside of 3 standard deviations of the grand mean, and so were removed (both from the 50 ms condition). See Figure 18

**Figure 18**

*Mean Total Time Spent Fixating Target for Congruous, Incongruous and No Sound Trials as a Function of Scene Preview Duration*



*Note.* Error bars represent 95% CIs.

A 3 x 3 mixed ANOVA showed no Sound x Preview Duration interaction, *F*(4, 200) = 0.03, *p* = .998, ηp$^2$ < .01. There was, however, a main effect of Preview Duration, *F*(2, 100) = 3.37, *p* = .038, ηp$^2$ = .06. Participants fixated on the target object for significantly longer durations in the 0 ms condition (*M* = 548 ms) than the 50 ms condition (*M* = 486 ms), *p* = .040 (Mean difference = 62 ms; 95% CI [3 ms, 120 ms]). However, there were no significant differences between the 0 ms condition and the 100 ms condition (*M* = 505 ms), *p* = .236, or the 50 ms and 100 ms condition, *p* = 1. There was no main effect of Sound condition, *F*(2, 200) = 0.26, *p* = .773, ηp$^2$ < .01, suggesting that the sound being played did not affect the length of time that the target object was fixated.

**Discussion**

As expected, we found that task performance – in terms of both overall response time and the time taken to first fixate on the target – improved when a preview image of the scene was displayed compared to when no preview was provided, strongly suggesting that participants were able to extract a scene's gist during the preview and use this to aid their subsequent search. Somewhat surprisingly, the differences in task performance between the 50 ms and 100 ms previews only approached significance. This suggests that a preview of 50 ms was enough for gist to be sufficiently processed, and that there was minimal additional benefit when this duration was doubled.

We also found that the overall response time and the time taken to fixate on the target object were significantly shorter when Congruous sounds were played compared to when Incongruous sounds were played. While the time taken on both these measures was shorter in the Congruous than the No sound condition, these differences again only approached significance. No differences in either variable were found between the Incongruous and No sound conditions, suggesting that inconsistent sounds did not significantly interfere with performance.

Additionally, no differences were found across Sound conditions for target object dwell time, suggesting that the type of sound listened to did not affect the time taken to recognise the target object once fixated. On the other hand, target dwell time was influenced by the duration of the preview. Specifically, participants spent longer fixating the target when no preview image had been displayed as compared to a 50 ms preview, suggesting they took longer to recognise the fixated target when a preview had not been provided. Finally, no interaction effects were found for any of the three dependent variables, due to the similarities in performance across the 50 and 100 ms conditions, and the No sound and Incongruous conditions.

Together, these results are suggestive of changes to gist processing through cross-modal means, whereby hearing a congruous object sound results in more efficient scene processing as compared to when hearing an inconsistent sound. This could be due to semantically congruent target sounds either facilitating the processing of a scene's visual information or allowing for a more rapid matching of gist against stored representations. However, it was important to rule out the possibility that the pattern of findings was simply due to congruous sounds facilitating object guidance rather than scene gist processing, through participants being able to identify the position of the target object – independent of the scene background – on Congruous trials during the initial flash preview.

## Experiment 6

### Design

Experiment 6 followed the same design as the previous version, with the only alteration being that the target and non-target objects were absent from the scene preview. This was to ensure that the effects seen in Experiment 5 were the result of improvements to gist processing, rather than participants being better able to locate the target object during the preview image. Care had been taken to position all target objects outside at least 3.72 degrees of visual angle from the centre of the screen, so as not to be within the foveal range of participants' gaze during the preview image. Similarly, the preview images were presented for durations too brief for gaze to move from the central gaze position and a second fixation to be made (see, for example, Rayner, 1998).

However, the possibility remained that the pattern of results from Experiment 5 was due to congruous sounds causing target objects to 'pop out' within a participant's visual periphery, irrespective of the scene context. In other words, it was possible the findings from the previous iteration could be the result of object-based facilitation, and akin to previous work showing targets are found more rapidly within an object array when accompanied by a consistent sound (e.g., Iordanescu et al., 2008; Iordanescu et al., 2010). Therefore, in Experiment 6 we removed this possibility by excluding the target (and non-target) objects from the initial scene preview.

**Participants**

Using the previous power calculations, the same sample size was sought for the second experiment. Again, restrictions imposed by the Covid-19 pandemic limited recruitment to 101 participants for Experiment 6 ($M_{\text{age}}$ = 23.20, $SD_{\text{age}}$ = 10.73; 78 Females, 23 Males; 90 Right-handed, 11 Left-handed). They were students and staff recruited through the University of East Anglia's research pool, as well as local residents from a volunteer research panel, who received either a small payment or course credits for participating. All reported having normal or corrected-to-normal vision and hearing. All participants provided written informed consent prior to taking part in the study. There were 33 participants included in the 0 ms condition, 34 in the 50 ms condition and 34 in the 100 ms condition.

**Stimuli**

The same sets of images and audio files from the previous experiment were used. However, in Experiment 6 the initial visual preview of the scene did not include the target and non-target object.

**Procedure**

The same procedure as Experiment 5 was followed.

**Results**

The same criteria used to process the Experiment 5 data were again employed. First, only trials where the target object was fixated were included in the analysis (10.08% of trials removed). Secondly, any instances where the target was fixated at the start of the trial were discarded (i.e., the participant had moved their gaze from the central fixation cross prior to the onset of the search screen). This led to removal of a further 0.35% of trials. Thirdly, trials were only included if the participant had made a response within the 15 second presentation duration, even if the target had been fixated, leading to a further 0.54% of trials being removed. Finally, only trials where the participant pressed the spacebar while fixating the target, or within 300 ms of moving their gaze away from the target, were included. This led to removal of a further 1.20% of trials. Overall, therefore, the analyses included 88.05% of the total number of trials sat in Experiment 6.

The stimulus issues identified in Experiment 5 were corrected, and so all 72 trials were included in the analyses for Experiment 6. As with Experiment 5, a logarithmic transformation (base 10) was again employed to resolve issues with normality, also allowing for direct comparisons between the two experiments.

***Overall response time***

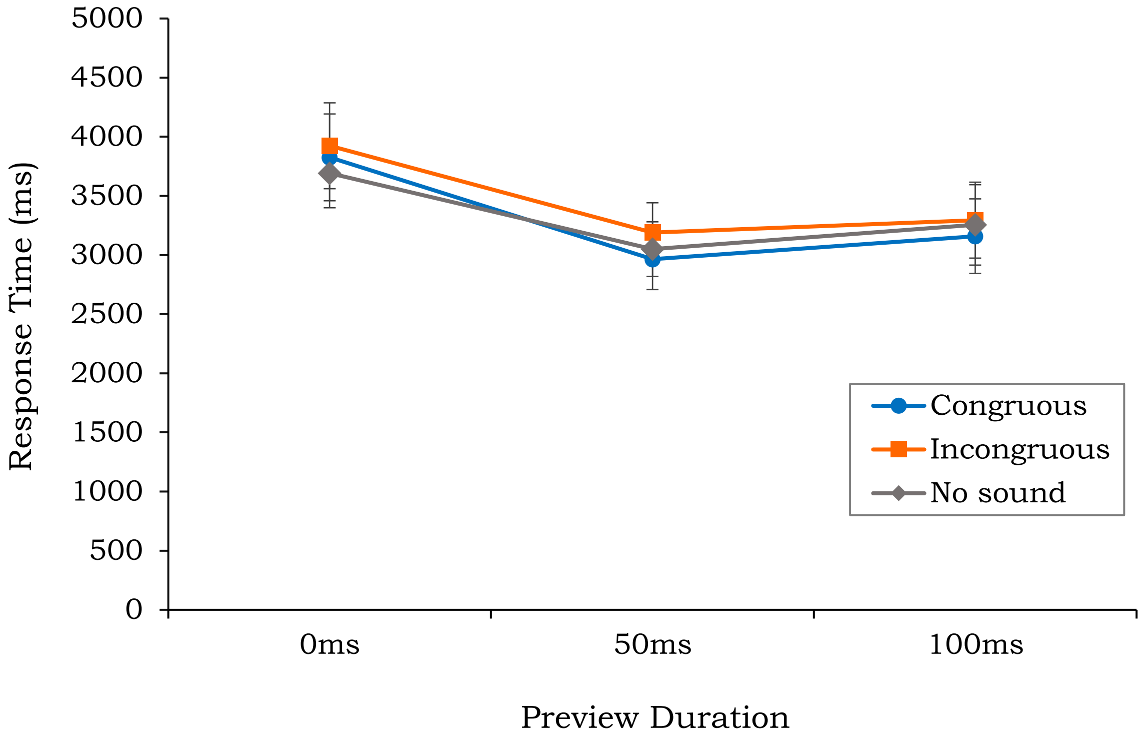
Two participants scored outside of 3 standard deviations of the grand mean, and so were removed (one from the 0 ms condition, and one from the 50 ms condition). See Figure 19.

A 3 x 3 mixed ANOVA showed no Sound x Preview Duration interaction, *F*(4, 192) = 0.50, *p* = .739, ηp$^2$ = .01. There was, however, a main effect of Preview Duration, *F*(2, 96) = 8.84, *p* < .001, ηp$^2$ = .16. As expected, participants were significantly slower to complete trials in the 0 ms condition (*M* = 3814 ms) compared to either the 50 ms condition (*M* = 3068 ms), *p* < .001 (Mean difference = 746 ms; 95% CI [282 ms, 1209 ms]) or the 100 ms condition (*M* = 3237 ms), *p* = .004 (Mean difference = 578 ms; 95% CI [117 ms, 1038 ms]). The difference between the 50 ms and 100 ms conditions was not significant, *p* = 1. There was also a main effect of Sound, *F*(2, 192) = 3.08, *p* = .048, ηp$^2$ = .03. Pairwise comparisons showed the difference between Congruous sounds (*M* = 3317 ms) and

Incongruous sounds (*M* = 3470 ms) was significant, *p* = .042 (Mean difference = -153 ms; 95% CI [-312 ms, 6 ms]), with shorter completion times on Congruous trials. There was no significant difference between Congruous sounds and No sounds (*M* = 3332 ms), *p* = 1, or for Incongruous sounds compared to No sounds (*p* = .189).

**Figure 19**

*Mean Response Time for Congruous, Incongruous and No Sound Trials as a Function of Scene Preview Duration*



*Note.* Error bars represent 95% CIs.

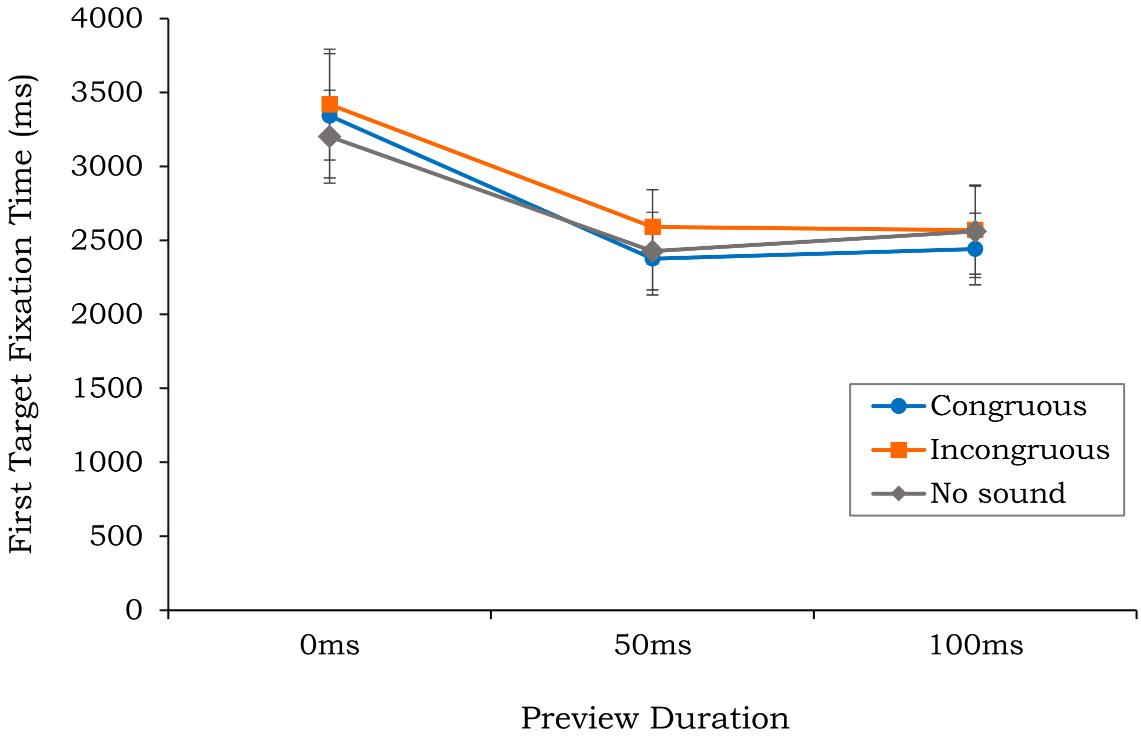
***Time taken to first fixate on target object***

One participant scored outside of 3 standard deviations of the grand mean, and so was removed (from the 100 ms condition).

A 3 x 3 mixed ANOVA showed no Sound x Preview Duration interaction, *F*(4, 194) = 0.49, *p* = .740, ηp$^2$ = .01. There was, however, a main effect of Preview Duration, *F*(2, 97) = 13.76, *p* < .001, ηp$^2$ = .22. As expected, participants were on average significantly slower to first fixate on the target object in the 0 ms condition (*M* = 3321 ms) compared to either the 50 ms condition (*M* = 2466 ms), *p* < .001 (Mean difference = 856 ms; 95% CI [396 ms, 1315 ms]) or the 100 ms condition (*M* = 2525ms), *p* < .001 (Mean difference = 796 ms; 95% CI [333 ms, 1260 ms]). The difference between the 50 ms and 100 ms conditions was not significant, *p* = 1. Unlike the findings from Experiment 5, the main effect of Sound only approached significance, *F*(2, 194) = 2.55, *p* = .081, ηp$^2$ = .03. See Figure 20.

**Figure 20**

*Mean Time to First Fixate on Target for Congruous, Incongruous and No Sound Trials as a Function of Scene Preview Duration*



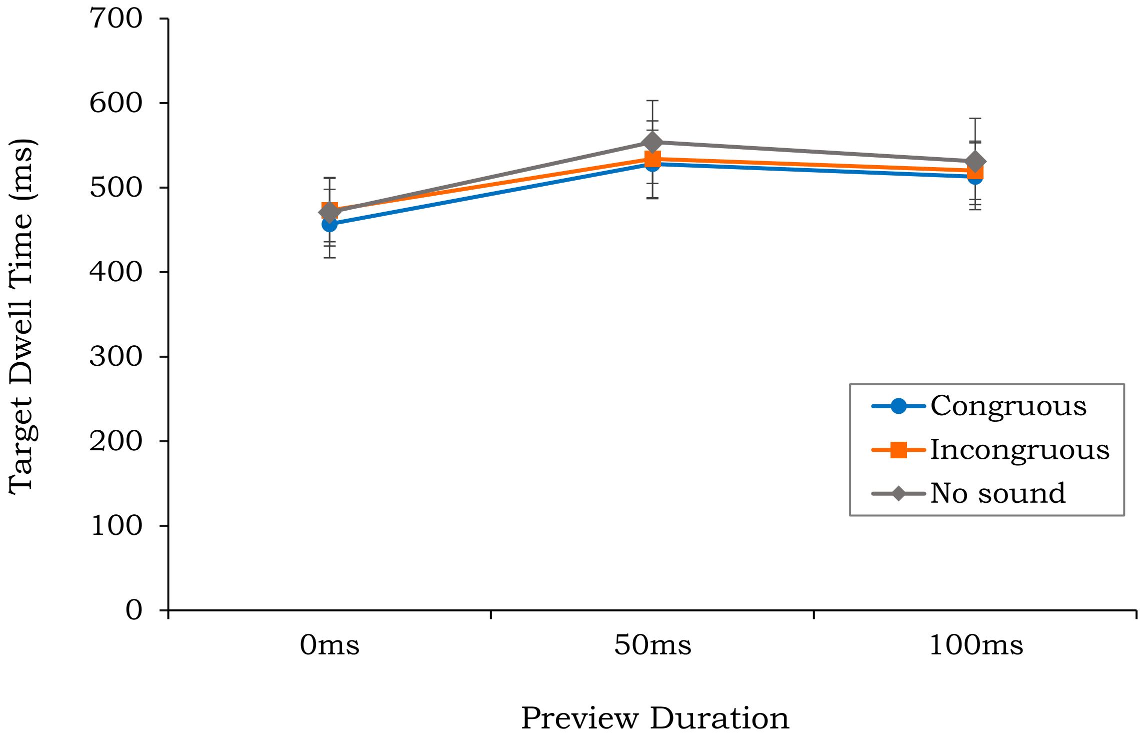
*Note.* Error bars represent 95% CIs.

***Total time spent fixating target (Target Dwell)***

Three participants scored outside of 3 standard deviations of the grand mean, and so were removed (one from the 0 ms condition, one from the 50 ms condition, and one from the 100 ms condition). See Figure 21.

**Figure 21**

*Mean Total Time Spent Fixating Target for Congruous, Incongruous and No Sound Trials as a Function of Scene Preview Duration*



*Note.* Error bars represent 95% CIs.

A 3 x 3 mixed ANOVA showed no Sound x Preview Duration interaction, *F*(4, 190) = 0.26, *p* = .903, ηp$^2$ = .01. There was, however, a main effect of Preview Duration, *F*(2, 95) = 4.53, *p* = .013, ηp$^2$ = .09. Participants fixated on the target object for significantly shorter durations in the 0 ms condition (*M* = 467 ms) than the 50 ms condition (*M* = 538 ms), *p* = .015 (Mean difference = -71 ms; 95% CI [-134 ms, -8 ms]. However, the difference between the 0 ms condition and 100 ms condition (*M* = 521 ms) only approach significance, *p* = .080 (Mean difference = -54 ms; 95% CI [-117 ms, 9 ms]), and there was no significant difference between the 50 ms condition and the 100 ms condition, *p* = 1. There was no main effect of Sound condition, *F*(2, 190) = 1.60, *p* = .204, ηp$^2$ = .02, suggesting that the sound being played did not affect the length of time that the target object was fixated.

**Discussion**

Experiment 6 served as a replication of the initial experiment, but with the absence of target objects from the flash preview scenes. As such, we aimed to determine whether congruous object sounds were influencing the processing of the scene image, rather than simply facilitating the processing of the target object in isolation. As with Experiment 5, we found that task performance – in terms of both overall response time and the time taken to first fixate the target – improved when a preview image of the scene was displayed compared to when no preview was provided, suggesting that participants were able to extract a scene's gist during the preview and use this to aid their subsequent search. Again, there were no significant differences in task performance between the 50 ms and 100 ms preview conditions. As before, this suggests that a preview of 50 ms was enough for gist to be sufficiently processed, and that there was little additional benefit when this duration was doubled.

In terms of Sound, and in line with Experiment 5, the overall response time was significantly shorter when Congruous sounds were played compared to when Incongruous sounds were played, but again not when compared to the No sound condition. As with Experiment 5, no differences were found between the Incongruous and No sound conditions, suggesting that inconsistent sounds did not significantly interfere with performance. However, while Experiment 5 found significant differences across Sound conditions in the time taken to fixate on the target, this only approached significance in Experiment 6.

As with the previous experiment, no differences were found across Sound conditions for target object dwell time, suggesting that the type of sound listened to did not affect the time taken to recognise the target object once fixated. On the other hand, target dwell time was again influenced by the duration of the preview. Surprisingly, however, this effect was in the opposite direction as that of Experiment 5, with participants spending less time fixating the target when no preview image had been displayed as compared to a 50 ms preview. This suggests that participants took longer to recognise the fixated target when a preview had been provided. As with Experiment 5, no interaction effects were found across any of the tested measures.

Together, the results of Experiment 6 are again suggestive of changes to gist processing through cross-modal means. A similar pattern of results was demonstrated across experiments, in terms of improved performance when a scene preview was provided and faster responses when congruous sounds were played. However, there were differences in terms of the relationship between target dwell time and preview duration, as well as the influence of the sound manipulation on the time taken to first fixate on the target. Therefore, to assess the nature of these similarities and differences, a comparison of the effects of the manipulations across the two experiments was undertaken.

## Comparison across experiments

To ascertain whether there were significant differences in the observed effects dependent on the presence or absence of the target object in the preview scene, we compared the pattern of results across Experiment 5 and Experiment 6. This was to ensure that the effects seen were driven by changes in scene – rather than object – processing ability. To do this, we created difference scores (Incongruous condition score minus Congruous condition score) for each participant on all three dependent variables. Such difference scores were chosen as both experiments had shown an effect of Sound when comparing the Congruous and Incongruous conditions. For each dependent variable, we then ran a 2 (Experiment: Experiment 5; Experiment 6) x 3 (Preview Duration: 0; 50; 100 ms) factorial independent ANOVA using these difference scores. No participants scored outside of 3 standard deviations of the grand mean for any of the separate conditions, and so all were retained in the analyses.

In terms of overall response time, no interaction was found, *F*(2, 200) = 0.45, *p* = .636, ηp$^2$ = .01, and there was no main effect of Experiment, *F*(1, 200) = 1.11, *p* = .294, ηp$^2$ = .01. The main effect of Preview Duration only approached significance, *F*(2, 200) = 2.738, *p* = .067, ηp$^2$ = .03.

In terms of the time taken to first fixate on the target, no interaction was found, *F*(2, 200) = 0.51, *p* = .603, ηp$^2$ = .01. There was also no main effect of Experiment, *F*(1, 200) = 1.77, *p* = .185, ηp$^2$ = .01. The main effect of Preview Duration was significant, *F*(2, 200) = 3.11, *p* = .047, ηp$^2$ = .03. Follow up pairwise comparisons with a Bonferroni correction showed that the difference between the 0 ms (*M* = 90 ms) and 50 ms condition (*M* = 333 ms) only approached significance, *p* = .064 (Mean difference = -242 ms; 95% CI [-517 ms, 32 ms]). There was no significant difference between the 0 ms and 100 ms condition (*M* = 99 ms), *p* = 1, or the 50 ms and 100 ms condition, *p* = .153.

In terms of the time spent fixating the target, no interaction was found, *F*(2, 200) = 0.14, *p* = .872, ηp$^2$ < .01. There was also no main effect of Experiment, *F*(1, 200) = 0.22, *p* = .640, ηp$^2$ < .01, or Preview Duration, *F*(2, 200) = 0.29, *p* = .745, ηp$^2$ < .01.

Taken together, therefore, the pattern of difference scores was similar across the two experiments. A minimal significant difference was found in terms of the influence of preview duration on the time taken to first fixate on the target object, but this was not apparent in the follow up comparisons. No other significant differences were found for any of the dependent variables across experiments. So, while there was some variation in results across the two iterations – most notably in the relationship between target dwell time and the presence or absence of a preview image – the similarities in terms of the sound manipulation across experiments suggest that the effect of congruous sounds seen in Experiment 5 was not solely driven by alterations to object processing.

## General Discussion

Study 2 investigated whether playing non-spatial object sounds affected scene gist processing. This was achieved by asking participants to search for target objects within scenes, while manipulating whether the sound being played during an initial preview of the scene was consistent with the object being searched for. In addition, the duration of this preview image was also manipulated, to be shown for either 100, 50 or 0 ms (i.e., no preview). In Experiment 5 the target objects were included in the preview image, whereas they were absent from the preview in Experiment 6. This allowed for a direct comparison between experiments, to determine the separate contributions of scene and object processing to subsequent search performance.

The results across Study 2 suggest that cross-modal stimulation affected gist processing, and specifically that experiencing congruous sounds led to more efficient processing as compared to when incongruous sounds were heard. In Experiment 5, the overall time taken to respond was found to be significantly shorter when congruous sounds were played. In addition, the time taken to first fixate on the target object was also significantly shorter in the Congruous condition, suggesting this cross-modal benefit was related to improvements in search efficiency rather than decision making. Furthermore, we did not find differences between the No sound and Incongruous conditions, suggesting no interference to processing from experiencing unhelpful sounds. Finally, sound did not influence the duration participants spent fixating the target object, suggesting this factor did not affect object recognition latencies. In Experiment 6, the overall response time was again found to be significantly shorter when congruous sounds were played, as compared to incongruous sounds. There was again no evidence for interference to processing from incongruous sounds, as no differences were found between this and the No sound condition. While the pattern of performance changes in terms of the time taken to first fixate on the target remained the same across experiments, with quicker times seen when congruous sounds were played, the difference between this and the Incongruous condition only approached significance in Experiment 6. Again, there was no evidence for sound influencing target object dwell time, suggesting that the type of sound listened to did not affect the time taken to recognise the target object once it had been fixated.

While the primary aim of Study 2 was an investigation of cross-modal effects on scene processing, findings in relation to the manipulation of preview image duration are of relevance to scene processing literature more widely. Firstly, in line with predictions, and research demonstrating the speed with which gist can be extracted from a scene (Malcolm et al., 2016), search performance across both experiments was significantly quicker when a brief scene preview was shown, compared to when no preview was displayed. Surprisingly, however, no performance differences were found between the 50 and 100 ms conditions of preview duration, suggesting the former was of sufficient duration for enough gist information to be extracted from the scene to facilitate subsequent search, and there to be little (if any) additional benefit from increasing this preview duration to 100 ms. Both experiments also revealed that the time spent fixating the target object was influenced by the length of the preview image. However, the pattern of results in relation to this differed across the study, with Experiment 5 finding longer target dwell times when no preview was shown, but with Experiment 6 finding such dwell times to be shortest in that same condition.

Lastly, a statistical comparison of the two experiments was conducted. This displayed the pattern of results as largely similar across experiments, with only minimal differences being found. Therefore, such a comparison helps strengthen the suggestion that the effects seen within Study 2 were driven, at least in part, by changes in the ability to process scene gist, rather than solely being related to object processing. Each of the above findings is addressed in more detail below, starting with the results in relation to the audio-visual nature of scenes, before a discussion of what the manipulation of scene preview reveals about gist processing more generally.

**Effects of cross-modal stimulation on gist processing**

In terms of the primary focus of Study 2, results across experiments suggested that cross-modal stimulation could affect gist processing, as hypothesised. Just as previous work has shown associative sounds to affect the visual processing of objects (Barenholtz et al., 2014; Giard & Peronnet, 1999; Molholm et al., 2004; Smith et al., 2007), our results suggest the same is true for the visual processing of scenes. Therefore, Study 2 extends our understanding of how perception is affected by the interaction of information from the auditory and visual sensory systems (e.g., Campbell & Dodd, 1980; McGurk & MacDonald, 1976; Sumby & Pollack, 1954). So, it appears that just as the visual characteristics of an object can be utilised to help disambiguate the scene category in which it is embedded (Brandman & Peelen, 2018), the auditory characteristics of that same object can be similarly influential. While recent research has suggested a cross-modal influence in the processing of scenes, to the author's knowledge this is the first time that semantically-associated sounds have been used to directly investigate alterations to gist processing. For example, while investigations using immersive audio-visual environments have suggested there to be benefits to processing through separate contributions from auditory and visual modalities, they have not used exposures below 100 ms (Rummukainen & Mendonca, 2016; Rummukainen et al., 2014).

Similarly, work involving object search within complex audio-visual scenes has not attempted to disentangle the separate mechanisms at play during cross-modal stimulation. For example, these have presented scenes for extended durations while asking participants to search within them (Kvasova et al., 2019; Mahzouni, 2019). As such, the finding of cross-modal effects in those studies cannot be exclusively attributed to rapid gist extraction, as the stage of processing influenced by sound remains unclear. Such an issue is further compounded by the reliance of those studies on behavioural measures. Put simply, presenting a scene to a participant and asking them to press a button once a target is found makes a determination impossible as to whether the sound manipulation is affecting early stages of visual processing or later stages of cognitive function such as decision making, and even whether the effects are related to alterations in scene processing, object processing, or both. Indeed, this has resulted in a lack of consensus as to the contributing factors, with explanations ranging from facilitated object processing (Kvasova et al., 2019) to changes in the allocation of attention (Mahzouni, 2019).

The use of the flash-preview moving window paradigm here, coupled with the recording of eye tracking measures, offers some answers. So, while our finding of quicker overall participant response time during congruous sound trials, compared to incongruous sound trials, mirrors that of previous studies (Kvasova et al., 2019; Mahzouni, 2019), we contend that this was due to different factors than those previously proposed. Firstly, the brief presentation durations of the scene preview, followed by a peripherally-obscured search screen, suggests that the audio condition was able to influence the earliest stages of visual processing, i.e., the processing of gist. Secondly, the finding of reduced times to first fixate on the target in trials with consistent sounds suggests that the effects were due to quicker search (e.g., Malcolm & Henderson, 2009), rather than speeded decision making or alterations to motor response times. Thirdly, that no sound-related alterations were found in the time participants spent fixating the target object prior to making a response suggests that the auditory stimulus did not significantly affect object recognition ability. For example, it might have been expected that congruous object sounds would lead to reduced target dwell times, by helping speed the extraction of featural information from the object once it had been fixated. However, Study 2 found no evidence to support this suggestion, potentially due to the fact that here the sound was played prior to when a participant would eventually fixate the object, contrary to those audio-visual studies where the sound is experienced while gaze is fixed on the target (e.g., Chen & Spence, 2010).

Taken together, therefore, the combination of findings across these separate measures offers strong support for auditory stimulation affecting scene gist processing. Additionally, the pattern of findings regarding the sound manipulation remained similar across experiments, irrespective of whether target objects were present in the scene preview. A statistical comparison of these experiments revealed only minimal differences, further suggesting that the cross-modal nature of effects was related to changes in the visual processing of scenes rather than objects. If, for example, the cross-modal influence was exclusively related to the early stages of object processing, then one would expect to see substantially divergent patterns of results dependent on the inclusion or exclusion of target objects from preview images. So, while there is much evidence that multimodal stimulation can affect object processing (e.g., Barenholtz et al., 2014; Giard & Peronnet, 1999; Molholm et al., 2004; Smith et al., 2007), we contend that such mechanisms cannot alone account for the findings across Study 2. Thus, it appears that real-world sound information within complex natural scenes can bias gist processing by affecting how visual properties are selected before a category meaning is determined, just as sound information has repeatedly been shown to affect the very early stages of object processing (e.g., Foxe et al., 2000; Schroeder & Foxe, 2005; Vetter et al., 2014). It may be, therefore, that the apparent cross-modal influence of associative sounds on search within a structured object array on a blank background (e.g., Iordanescu et al., 2008; Iordanescu et al., 2010) is functionally different from that of search within a complex natural scene.

There are two important points to make as to the specific pattern of the influence of sound across Study 2. Firstly, whether congruous sounds can be said to have facilitated gist processing is open to question. It was found across measures that there was significantly improved performance when congruous, compared to incongruous, sounds were played, but the differences between this condition and trials where no sound was heard tended to only approach significance. As a result, care has been taken not to claim that the results show facilitation of gist processing as compared to baseline performance. However, we contend that – although effect sizes might be smaller than expected – the pattern of similar performance across No sound and Incongruous conditions, with markedly better performance in the Congruous condition, is strongly suggestive of a trend within the data towards gist facilitation. Such a contention is in agreement with other research showing cross-modal facilitation without interference (e.g., Iordanescu et al., 2010; Kvasova et al., 2019), but more work is certainly warranted. Secondly, there was no evidence of interference to processing ability through the playing of unhelpful sounds, with no significant differences in performance between the No sound and Incongruous conditions being found across measures and experiments. This was in line with predictions, based on recent studies of audio-visual processing finding no such interference during search of object arrays (Iordanescu et al., 2008; Iordanescu et al., 2010) and complex scenes (Kvasova et al., 2019; Mahzouni, 2019). It appears, therefore, that the great efficiency with which gist extraction operates (Cohen et al., 2011; Fei-Fei et al., 2005; F. Li et al., 2002; Rousselet et al., 2002) meant that unhelpful auditory information was here rapidly discarded from processing.

Crucially, these results confirm that gist processing can be influenced by additional information presented simultaneously with the scene. Study 1 had demonstrated gist to be affected by observer expectations formed prior to scene onset, through utilisation of additional information to generate more accurate predictions as to an upcoming stimulus, and thus ultimately more efficient subsequent processing. Such a finding presented a challenge to forward-sweep models (e.g., Itti et al., 1998; Potter et al., 2014; Rumelhart, 1970), due to the suggested role of top-down input over the brief timeframes of rapid scene processing. However, this is perhaps a less direct critique of such models, as eliciting expectations may simply influence the pre-emptive selection of scene templates to be used for representation matching (see, for a review of perceptual hypothesis testing models, Clark, 2013). In other words, within such a scenario it could be argued that gist processing does follow a forward sweep of activation, at least up to the point of representation matching. The specific pattern of behavioural and ERP results from Study 1 suggests this not to be the case, with scene gist being influenced by top-down information in a widespread manner across the processing stream – including the initial stages of feature extraction – and the findings of Study 2 can be taken as further support for this asseveration. They suggest alterations to gist processing are also possible when semantically relevant additional information is provided at the same time as a visual scene is experienced, and thus demonstrate alterations to early scene processing are not related to the manipulation of expectations alone. While this has previously been suggested in terms of exclusively visual scenes (Greene et al., 2015), to the author's knowledge this is the first demonstration of simultaneously presented cross-modal alterations to scene gist processing.

**Effects of preview duration on gist processing**

A fundamental finding across Study 2 was the increase in participant performance when provided with a brief preview of the scene prior to search. Indeed, the results suggest that this manipulation had the most dominant effect on search ability, with participants' search ability being consistently and significantly better – irrespective of the audio condition – when a preview was displayed. This was to be expected, due to the substantial amount of previous work demonstrating that gist can be extracted from scenes at presentation durations of 100 ms and below (e.g., Biederman et al., 1974; Castelhano & Henderson, 2007; Eckstein et al., 2006; Fei-Fei et al., 2007; Neider & Zelinsky, 2006; Potter, 1975; Potter et al., 2014). Our results are confirmatory of such research, through demonstrating that observers were able to gather enough information from this brief glimpse of a scene to significantly improve subsequent visual search. While visual information relating to the scene – such as expansiveness or general structural layout – may assist search within it, by providing depth cues, position of the horizon, and so forth, it is an appreciation of the meaning of a scene which allows for more efficient search (see, for a discussion of the importance of semantic relevance within scenes, Henderson & Hayes, 2018). It is the author's contention, therefore, that the information gathered in this brief glimpse was both visual and semantic in nature, in what has been termed 'conceptual gist' (Oliva, 2005). In other words, and to use a previous example, understanding that a scene is a kitchen allows an observer to identify certain areas within the visual field as being kitchen counters, and ultimately use this knowledge to assist with locating the position of a microwave.

Unexpectedly, however, performance remained similar across scene previews of 50 and 100 ms. This was true for each of the measures investigated, and across both experiments. While a 50 ms preview was expected to be long enough to allow for gist extraction, it was hypothesised that a doubling of this duration would provide the opportunity to extract further details as to the scene, leading to enhanced performance. The results, contrarily, suggest that this was not the case. It is important to note, however, this should not be taken to mean that all available information was extracted within the first 50 ms, thus making any extension to the preview duration redundant. There is indeed much evidence that our understanding of a scene – and the amount of information gathered from it – develops over time (Castelhano & Henderson, 2008; Greene & Oliva, 2009; Larson et al., 2014). So, while further elongated preview durations may well have resulted in even more efficient search than demonstrated here, the findings of Study 2 suggest search was not significantly improved by whatever additional information was gathered by participants between 50 and 100 ms after scene onset. Such a proposition is in accord with models suggesting that a substantial amount of information relating to a scene is derived immediately and automatically, rather than in a step-by-step linear fashion (see, for example, F. Li et al., 2002).

As discussed above, a comparison across experiments revealed only minimal differences. A divergence of note, however, was an alteration to the pattern of target dwell times in relation to the manipulation of preview duration. This was unexpected but is also potentially revealing. These dwell times followed the predicted pattern in Experiment 5 – with longer fixation times on the target in the 0 ms condition compared to when a preview was shown – suggesting that participants took longer to process information relating to the target and its position when they did not have the opportunity to derive the layout and meaning of a scene prior to search. However, Experiment 6 saw the opposite pattern, with increased target dwell times when a preview was displayed. As this was not expected, an explanation is necessarily speculative. We suggest the most likely cause is due to a mismatch between the parafoveal information gathered by the participant during the preview and the visual information within the subsequent search screen. The distance target objects were placed from the centre of the screen makes it unlikely that participants could perceive the location of the target prior to search. However, it is more likely that they may have been able to gather spatial frequency information from the parafovea during the preview, and use this to determine which regions of a scene were ‘empty’, and thus unlikely to contain the object in question. In other words, a participant might make the judgement that a target is not in a certain region within the scene, due to the visual information extracted during the preview, and so subsequently – and unexpectedly – finding the object within that position during search would likely result in additional processing due to the unexpected nature of this event. A potentially revealing follow-up to test this proposition would be to have an unrelated object acting as a proxy in the target location during the scene preview, as would the gathering of responses from participants as to whether they were aware of the absence of targets from previews.

A final point needs to be acknowledged in relation to differences across the experiments. The overall response time and time taken before the target was fixated were both substantially longer in Experiment 6 than Experiment 5, and this was the case for all preview durations. As there was no alteration to the format of the 0 ms condition across the two versions of the experiment – with this intended to act as baseline – it appears that these extended durations were not due to experimental manipulation. The trial-by-trial data and the experiment logs were extensively reviewed, confirming there to be no programming errors or unintended alterations to the protocol between experiments. It was found, however, that during the intervening period between data collections that the Experiment Builder software was updated and replaced its previous image renderer, DirectDraw with a new application, OpenGL. It is worth noting that designs utilising a gaze-contingent window require constant updating of the display image and so even small changes to the speed and perceived smoothness with which these are rendered can have a significant impact on participants’ ability to conduct visual search. So, while this change may not be responsible for the timing issues, we consider this to remain the most likely candidate for the differences across experiments (for discussion of the superiority of DirectDraw in situations requiring fast display refresh see, Ward 2000).

**Future directions**

Study 2 has extended recent work investigating the audio-visual processing of scenes, and opened up several potential avenues for future research. The use of eye tracking methods has been revealing as to the nature of cognitive processing compared to behavioural studies (e.g., Kvasova et al., 2019; Mahzouni, 2019), and this could be further expanded. For example, here we have focussed on first fixation time and target dwell time, but there are many additional variables that might be revealing. For example, an examination of the mean fixation durations during search could signal whether the ease with which the image was being matched to an internal representation changed as a function of auditory condition. Alternatively, pupillometry could reveal whether the separate sound conditions were associated with alterations to cognitive load prior to the start of search (see, for example, Laeng et al., 2012). Further still, expansion of such methods could also incorporate EEG recording, which would be potentially revealing – as in Study 1 – in terms of the temporal dynamics of cross-modal influence, and thus allowing for further investigation of the effects on separate mechanisms such as those involved in feature extraction or template matching.

It is also important for future work to clarify whether manipulating the relative onsets of the auditory and visual stimuli leads to changes in the pattern of results. Here, it was chosen to play object sounds synchronously with the onset of the scene preview, as this was judged to be the closest approximation of processing outside the lab. However, this is not necessarily synonymous with daily experience. It may be the case, for example, that we tend to hear a scene before we see it, such as hearing voices and the clanking of cutlery from a busy canteen while walking towards it along a corridor. It may further be the case – as suggested by the finding of expectations facilitating gist processing in Study 1 – that sounds which allow us to predict the environment about to be encountered might be particularly beneficial to our processing of it. Indeed, previous research has suggested that the level of asynchrony between auditory and visual stimuli is a crucial factor to consider (e.g., Chen & Spence, 2010). For example, a direct investigation of stimulus onset asynchrony (SOA) on cross-modal processing of complex scenes suggested that effects were largest when the sound preceded the scene, somewhat tempered with simultaneous onsets, and absent when the sound followed the image (Mahzouni, 2019). Likewise, the other study investigating audio-visual scene processing of which the author is aware fixed the onset of the auditory stimulation at 100 ms prior to the visual scene appearing (Kvasova et al., 2019). Therefore, such an investigation of SOA might be particularly pertinent here due to the discrepancy between the durations of the preview image and audio playback, where the sound continued to play for a substantial time after the scene had disappeared. It may be the case, therefore, that having the sound onset at a point prior to the preview appearing may lead to a stronger cross-modal influence.

Our investigation of the interplay between auditory and visual stimulation has been necessarily exploratory. Here, it was chosen to use object sounds for auditory stimulation due to the previous suggestion of improved object search when associated object sounds are played (Iordanescu et al., 2008; Iordanescu et al., 2010), as well as the strong associations that exist between scenes and objects. However, it is also true that many scenes are associated with sounds that may not be considered as originating from a single object. For example, a football stadium is experientially connected with the sound of cheering crowds, Niagara Falls with the crashing of falling water, and a busy highway with the sound of traffic. Furthermore, scenes are often populated by separate distinct sounds, such as a harbour with the clanking of rigging, the lapping of the tide against the sea wall and the chatter of gulls. It is likely, therefore, that the association of sounds with scenes lies on a spectrum, ranging from a single object emitting a single sound (as here), to a rich selection of naturalistic sounds emanating from an environment. Recent work has started to address this (see, for example, Kvasova et al., 2019; Rummukainen & Mendonca, 2016; Rummukainen et al., 2014), but we are no doubt still at the early stages of this endeavour.

Finally, several of the methodological choices made here could be adapted in future work. For example, we included a No sound condition rather than a condition playing neutral sound (such as white noise). It may be argued that this is not a fair reflection of baseline performance, as it compares conditions of different levels of sensory stimulation. However, previous work has suggested there to be little difference between playing white noise and the absence of sound in the processing of natural scenes (Mahzouni, 2019). Similarly, our results clearly show only minimal differences between the No sound and Incongruous conditions. So, it appears that unhelpful auditory stimuli – whether it be white noise or incongruous sounds – do not influence processing, with performance remaining comparable to situations where no audio is present. Another methodological choice made here was to include a wide assortment of objects and scenes, in order to maintain generalisability in the findings. This meant that the level of association between scenes and objects varied a great deal. For example, a bowling ball and a bowling alley have a much stronger and more specific semantic connection than, say, a remote-controlled car and a child's bedroom. We see this variability as a strength in the design, but future work might choose to investigate this specificity of associations further. For example, a systematic manipulation of these associations would help determine whether a linear relationship existed between the semantic closeness of fit between an object and a scene and the amount of influence on gist processing.

Study 2 serves as an important extension to our current understanding of scene processing. While there has been much investigation as to the cross-modal processing of objects, similar research related to the audio-visual nature of scenes has been less forthcoming. Indeed, those previous studies which have involved the use of complex scenes have still ultimately tended to concern themselves with the audio-visual influence on object processing. The methodological choice made here, on the other hand, have allowed us to demonstrate that semantically related sounds can affect the efficiency with which the gist of a scene can be extracted. It appears, therefore, that just as experiencing semantically consistent sounds influences the rapid processing of an object's visual features, a similar influence of auditory stimulation may exist for the rapid processing of scenes. These findings can also be taken as extending those presented in Study 1. While that study displayed gist extraction to be affected through providing additional visual information, Study 2 showed this influence to not necessarily be modality specific. Furthermore, while Study 1 demonstrated the effect of *a priori* expectations on gist processing, Study 2 showed such effects were also possible when additional information was provided simultaneously with scene onset.

## Conclusion

Presented here has been a collection of behavioural, ERP and eye tracking experiments, aimed at furthering our understanding of how scenes are processed outside the laboratory. In Study 1 we used ‘approach’ images to elicit predictions in participants as to the identity of an upcoming scene, and in Study 2 the semantic congruency of simultaneously presented scene images and object sounds was manipulated in an investigation of cross-modal influence to gist processing.

These two studies have comprised a total of seven experiments, and the testing of over 500 participants. For Study 1, Experiment 1a and 1b manipulated approach-destination congruency as well as target presentation duration, in order to investigate the time-course of expectation effects on the processing of conceptual gist. These revealed a benefit for categorising scenes when they were semantically congruent with approach images and, furthermore, that this advantage was greatest at the briefest durations (i.e., when the opportunity to process visual information was most limited). Experiment 2 then investigated the influence of spatiotemporal coherence on gist processing, by manipulating the sequentiality of these pre-target series. This revealed that the increased ability shown by participants for Congruous trials was based on approach images providing a semantic context for upcoming targets. To disentangle the separate roles of facilitation and interference on gist processing, Experiment 3 introduced a baseline condition which replaced approach images with coloured patterns, and so provided no information from which to generate predictions as to the identity of an upcoming scene category. This confirmed that providing participants with semantically congruous approach images led to the facilitation of gist processing, compared to baseline, and that semantically incongruous approaches resulted in reduced performance.

The final experiment of Study 1 employed electroencephalography to chart the neural correlates associated with the manipulation of scene congruency. This revealed an effect of expectations on rapid scene processing across all tested ERP components. For Incongruous trials, the

N400 showed a significantly more negative mean amplitude within the Centro-parietal and Frontal regions, while significantly more positive mean amplitudes for the P2 and P600 were seen within the Parieto-occipital region. Furthermore, significantly more negative amplitudes were also associated with Incongruous trials across the early and late time-windows within Frontal sites. Together, therefore, Experiment 4 revealed congruency-related changes within the earliest known marker of scene-specific processing (P2), within the component suggested as indexing semantic expectancy and the retrieval of conceptual information (N400), and within the component associated with semantic and syntactic processing (P600). Such a finding of congruency-based alterations to the ERP across all time-windows of interest indicates it is unlikely that a singular temporal or cortical point exists at which top-down predictions affect processing. Rather, this pattern of amplitude changes suggests that *a priori* expectations had a broad effect across multiple stages of scene processing, both to early feature extraction mechanisms as well as to more advanced levels of the processing stream

Study 2 then investigated whether playing non-spatial object sounds influenced the processing of scene gist, across two eye tracking experiments. Participants searched for target objects within complex natural scenes, while the consistency of a sound being played during an initial preview of the scene and the object being searched for was manipulated. Additionally, the duration of this preview image was also varied, being displayed for either 100, 50 or 0 ms (i.e., no preview). The results across Study 2 revealed an effect of cross-modal stimulation on gist processing, with the playing of congruous sounds leading to more efficient processing as compared to when incongruous sounds were heard. In the first experiment of Study 2 – where target objects were included within the initial preview image – both the overall time to respond and the time taken to first fixate on the target were significantly shorter when congruous sounds were played. In addition, sound-related changes were not found in the total time participants spent fixating the target objects. Together, therefore, these results strongly suggest that the cross-modal benefit was founded on improvements to search efficiency, rather than due to changes in object recognition ability or decision-making strategies. Finally, we found no evidence of interference to processing from experiencing unhelpful sounds, as no differences in performance were identified between the incongruous and no sound conditions, suggesting that these sounds could be easily discarded from processing without cost to cognition.

The second experiment of Study 2 repeated the design of the previous iteration, except that target objects were absent from the initial scene preview. This allowed for a determination as to the separate contribution of scene and object processing to subsequent search performance. A similar pattern of results emerged, the correspondence of which was subsequently confirmed by a statistical comparison of the findings across experiments. These results suggested that the cross-modal effects were driven by changes in the ability to process scene gist rather than exclusively being related to object processing.

Together, this broad range of findings, spanning separate methodologies, suggests a number of theoretical implications. Perhaps the most considerable contribution of this work is the finding of influences to gist processing that lie outside of immediate visual processing. The 'approach-destination' congruency related changes seen in Experiments 1a-b demonstrated the effect of observer predictions on gist processing. As the most substantial differences were found at the briefest presentation durations – demonstrating an influence of expectations at the earliest stages of processing – this strongly suggests that top-down information has a role in modulating the extraction of scene gist. Furthermore, the results from Experiment 3 confirmed that this top-down influence led to the facilitation of gist processing. Such findings are a challenge to models which propose gist is exclusively based on a forward sweep of activation through the visual system (Itti et al., 1998; Potter et al., 2014; Rumelhart, 1970), and is in agreement with recent work demonstrating top-down alterations in the ability to extract meaning from rapidly presented scenes (Greene et al., 2015; Smith & Loschky, 2019). It also appears to challenge accounts which imply that gist processing operates at ceiling and takes place automatically, outside of attention (Biederman, 1972; F. Li et al., 2002; Potter, 1975; Thorpe et al., 1996). Our contention that expectations were influencing gist processing was further strengthened by changes to the early ERPs seen in Experiment 4, with these amplitude changes suggesting an influence of top-down information while perceptual processing was still ongoing (e.g., Bar, 2003; Fenske et al., 2006). In addition, the finding of changes to the N400 and P600 components can be taken as suggesting that the manipulation of approach-destination congruency was altering the semantic and syntactic processing of scenes.

However, the use of approach images across Study 1 meant that predictions could be formed prior to the onset of the target scene. As such, it is reasonable to suggest that templates expected to be required for matching against the upcoming stimulus could be preactivated, allowing either for a stored representation to be available prior to the appearance of the target-derived signal, or for pre-emptive changes in error thresholds at early processing levels through predictive coding mechanisms (e.g., Rauss et al., 2011), or both. It could be argued that such a scenario does not provide a direct challenge to forward sweep models, as the processing of a scene’s gist at onset may take place in a purely bottom-up manner within a system altered ahead-of-time by top-down influences. In other words, approach images may *a priori* set the boundaries within which the subsequent bottom-up information is to be processed. Study 2, however, showed this to be unlikely, as the findings clearly demonstrated gist to be affected by the congruency of information provided simultaneously with scene onset. Therefore, this strongly suggests that alterations to early scene processing are not only possible through the manipulation of expectations, but are also apparent when semantically relevant additional information is provided at the same time as the visual scene is encountered.

The mechanisms by which simultaneously presented information might influence gist are still unclear. One potential explanation is offered by the object processing literature, where evidence has revealed that top-down processes are initiated prior to completion of target recognition, with the suggestion that early activation of higher-order brain regions such as the orbitofrontal cortex facilitates the systematic analysis of bottom-up information (Bar et al., 2006). In other words, low spatial frequency information is passed rapidly to higher areas in prefrontal cortex – possibly using the magnocellular pathway – and is then used to form predictions as to the identity of the object being viewed. Consequently, this allows for the pre-activation of a limited set of object representations which are subsequently matched against the continuing flow of bottom-up information (Bar et al., 2006).

Further to this, though, Study 2 displayed that the influence on gist processing was possible cross-modally. As far as the author is aware, this is the first time semantically associated sounds have been used to directly investigate alterations to gist. So, while some recent work has investigated the audio-visual nature of complex scenes, they have not been explicitly concerned with the role of auditory stimulation on the extraction of gist (e.g., Kvasova et al., 2019; Rummukainen & Mendonca, 2016). The findings across Study 2 strongly suggest that real-world sound information within complex natural scenes can bias gist processing by affecting how visual properties are selected before a category meaning is determined. It appears therefore, that just as simultaneously presented consistent sounds have been shown to affect the visual processing of objects (Barenholtz et al., 2014; Giard & Peronnet, 1999; Molholm et al., 2004; Smith et al., 2007), the same may well be true for gist processing. It appears, therefore, that just as experiencing semantically consistent sounds influences the rapid processing of an object's visual features, a similar influence of auditory stimulation may exist for the rapid processing of scenes.

Perhaps at the most fundamental level, these findings further confirm the speed and efficiency with which gist processing takes place. Across all experiments, participants showed that a 50 ms presentation of a scene was sufficient for gist to be extracted, to both allow for the scene's basic level category to be derived and to improve subsequent search ability within it. One important divergence across the two studies was in relation to performance changes between 50 and 100 ms presentation durations. For example, the findings from those experiments in Study 1 where duration was manipulated showed a clear improvement in categorisation performance when a scene was displayed for 100 ms rather than 50 ms. In Study 2, conversely, search performance within a scene did not appear to be substantially affected by whether the initial preview of the scene lasted for 50 or 100 ms, suggesting that sufficient gist information was extracted from the scene to facilitate subsequent search during this shorter duration. It appears, therefore, that a doubling of presentation time did affect categorisation ability, but not search performance. However, at present it is unclear whether this discrepancy is related to the variability in the tasks participants were asked to perform across studies, the antecedent compared to simultaneous nature with which additional information was provided, the unimodal compared to cross-modal presentation formats, and so forth. It appears, therefore, that further testing would be warranted.

The findings presented here have opened several important avenues for future work, and these have been discussed above. Perhaps two of these are most pressing. Firstly, the desire here to better reflect scene processing outside the lab could be extended much further. For example, moving away from a reliance on individual static scene images appears to be a crucial step in improving the ecological validity of research, and so future work should strongly consider the use of video clips, immersive audio, VR technology, and even the testing of participants within natural environments. Secondly, the demonstration of influences to gist processing begs the question as to what other sources of influence might exist. For example, different forms of top-down information may have differing effects, such as that based on observer goals (e.g., navigation), the role of protagonists within scenes (such as interpersonal relationships or actions towards objects), etc. Similarly, these results ask as to what the separate contributions of top-down and bottom-up information are. The introduction of new methods such as transcranial magnetic stimulation would most likely be enlightening, by charting performance changes in relation to the targeted – both temporally and spatially – interruption of re-entrant communication (see, for example, Camprodon et al., 2010; de Graaf et al., 2012; de Graaf et al., 2014; Koivisto et al., 2011).

In sum, therefore, it is the author’s contention that the work described over the preceding pages offers a crucial new step in furthering our understanding of scene gist processing. While the remarkable speed and efficiency with which gist is extracted may seemingly make ‘forward sweep’ models appealing, strong evidence for the influence of top-down and cross-modal stimulation has here been provided. It appears that it is time to move away from the traditional paradigm of testing ability using individual visual scene images, to allow for further investigation of other real-world factors that might similarly affect how we rapidly extract the meaning of a scene. Our collective aim must ultimately be, of course, to better understand how we process the scenes we encounter within the continuous experiential flow of our daily lives, rather than simply determining how quickly a single two-dimensional image on a monitor can be recognised.

## References

Adams, R. B., & Janata, P. (2002). A comparison of neural circuits underlying auditory and visual object categorization. *Neuroimage*, *16*(2), 361-377.

Afiki, Y., & Bar, M. (2020). Our need for associative coherence. *Humanities and Social Sciences Communications*, *7*(1), 1-11.

Aitken, F., Menelaou, G., Warrington, O., Koolschijn, R. S., Corbin, N., Callaghan, M. F., & Kok, P. (2020). Prior expectations evoke stimulus templates in the deep layers of V1. *bioRxiv*.

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, *166*(3), 559-571.

Angrilli, A., Penolazzi, B., Vespignani, F., De Vincenzi, M., Job, R., Ciccarelli, L., ... & Stegagno, L. (2002). Cortical brain responses to semantic incongruity and syntactic violation in Italian language: an event-related potential study. *Neuroscience letters*, *322*(1), 5-8.

Antes, J. R., Penland, J. G., & Metzger, R. L. (1981). Processing global information in briefly presented pictures. *Psychological research*, *43*(3), 277-292.

Banno, H., & Saiki, J. (2015). The processing speed of scene categorization at multiple levels of description: The superordinate advantage revisited. *Perception*, *44*(3), 269-288.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of cognitive neuroscience*, *15*(4), 600-609.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., ... & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, *103*(2), 449-454.

Barenholtz, E., Lewkowicz, D. J., Davidson, M., & Mavica, L. (2014). Categorical congruence facilitates multisensory associative learning. *Psychonomic bulletin & review*, *21*(5), 1346-1352.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01).

Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and cognition*, *14*(2), 201-212.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419-429.

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*(5), 809-823.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(4043), 77-80.

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of experimental psychology*, *97*(1), 22.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, *14*(2), 143-177.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of experimental psychology*, *103*(3), 597.

Blondin, F., & Lepage, M. (2005). Decrease and increase in brain activity during visual perceptual priming: An fMRI study on similar but perceptually different complex visual scenes. *Neuropsychologia*, *43*(13), 1887-1900.

Boehler, C. N., Schoenfeld, M. A., Heinze, H. J., & Hopf, J. M. (2008). Rapid recurrent processing gates awareness in primary visual cortex. *Proceedings of the National Academy of Sciences*, *105*(25), 8742-8747.

Boyce, S. J., & Pollatsek, A. (1992). Identification of objects in scenes: the role of scene background in object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 531.

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1160.

Brandman, T., & Peelen, M. (2018). Object cues facilitate the multivariate representations of scene layout in human fMRI and MEG. *Journal of Vision*, *18*(10), 1242-1242.

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, *1*(1), 135-160.

Bullier, J. (2001). Integrated model of visual processing. *Brain research reviews*, *36*(2-3), 96-107.

Caddigan, E., Choo, H., Fei-Fei, L., & Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categories. *Journal of vision*, *17*(1), 21-21.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology*, *32*(1), 85-99.

Camprodon, J. A., Zohary, E., Brodbeck, V., & Pascual-Leone, A. (2010). Two phases of V1 activity for visual recognition of natural images. *Journal of cognitive neuroscience*, *22*(6), 1262-1269.

Caplette, L., Gosselin, F., Mermillod, M., & Wicker, B. (2020). Real-world expectations and their affective value modulate object processing. *NeuroImage*, *213*, 116736.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., ... & Rust, N. C. (2005). Do we know what the early visual system does?. *Journal of Neuroscience*, *25*(46), 10577-10597.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753.

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human perception and performance*, *34*(3), 660.

Catherwood, D., Edgar, G. K., Nikolla, D., Alford, C., Brookes, D., Baker, S., & White, S. (2014). Mapping brain activity during loss of situation awareness: an EEG investigation of a basis for top-down influence on perception. *Human factors*, *56*(8), 1428-1452

Chaumon, M., Drouet, V., & Tallon-Baudry, C. (2008). Unconscious associative memory affects visual processing before 100 ms. *Journal of vision*, *8*(3), 10-10.

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.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.

Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological science*, *22*(9), 1165-1172.

Cohn, N., & Foulsham, T. (2020). Zooming in on the cognitive neuroscience of visual narrative. *Brain and Cognition*, *146*, 105634.

Cohn, N., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2014). The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension. *Neuropsychologia*, *64*, 63-70.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea) nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive psychology*, *65*(1), 1-38.

Crouzet, S. M., & Serre, T. (2011). What are the visual features underlying rapid object recognition?. *Frontiers in psychology*, *2*, 326.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological science*, *15*(8), 559-564.

De Cesarei, A., Mastria, S., & Codispoti, M. (2013). Early spatial frequency processing of natural images: an ERP study. *PloS one*, *8*(5).

De Fockert, J. W. (2010). Early top-down attentional modulation in visual processing. *Acta psychologica*, *135*(2), 112-113.

de Graaf, T. A., Goebel, R., & Sack, A. T. (2012). Feedforward and quick recurrent processes in early visual cortex revealed by TMS?. *Neuroimage*, *61*(3), 651-659.

de Graaf, T. A., Koivisto, M., Jacobs, C., & Sack, A. T. (2014). The chronometry of visual perception: review of occipital TMS masking studies. *Neuroscience & Biobehavioral Reviews*, *45*, 295-304.

De Vincenzi, M., Job, R., Di Matteo, R., Angrilli, A., Penolazzi, B., Ciccarelli, L., & Vespignani, F. (2003). Differences in the perception and time course of syntactic and semantic violations. *Brain and language*, *85*(2), 280-296.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9-21.

Demiral, Ş. B., Malcolm, G. L., & Henderson, J. M. (2012). ERP correlates of spatially incongruent object identification during scene viewing: Contextual expectancy versus simultaneous processing. *Neuropsychologia*, *50*(7), 1271-1285.

Draschkow, D., Heikel, E., Võ, M. L. H., Fiebach, C. J., & Sassenhagen, J. (2018). No evidence from MVPA for different processes underlying the N300 and N400 incongruity effects in object-scene processing. *Neuropsychologia*, *120*, 9-17.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological science*, *17*(11), 973-980.

Egeth, H. E., Leonard, C. J., & Leber, A. B. (2010). Why salience is not enough: Reflections on top-down selection in vision. *Acta psychologica*, *135*(2), 130.

Endsley, M. R., & Garland, D. J. (2000). Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, *1*(1), 3-21.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, *2*(10), 704-716.

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual review of vision science*, *5*, 373-397.

Epstein, R. A., Higgins, J. S., & Thompson-Schill, S. L. (2005). Learning places from views: variation in scene processing as a function of experience and navigational ability. *Journal of Cognitive Neuroscience*, *17*(1), 73-83.

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological science*, *22*(6), 739-746.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free?. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1476.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience*, *22*(13), 5749-5759.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160.

Federmeier, K. D., & Kutas, M. (2002). Picture the difference: Electrophysiological investigations of picture processing in the two cerebral hemispheres. *Neuropsychologia*, *40*(7), 730-747.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene?. *Journal of vision*, *7*(1), 10-10.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, *12*(6), 893-924.

Fenske, M. J., Aminoff, E., Gronau, N., & Bar, M. (2006). Top-down facilitation of visual object recognition: object-based and context-based contributions. *Progress in brain research*, *155*, 3-21.

Ferrari, V., Codispoti, M., & Bradley, M. M. (2017). Repetition and ERPs during emotional scene processing: A selective review. *International Journal of Psychophysiology*, *111*, 170-177.

Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M., & Keller, G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nature neuroscience*, *19*(12), 1658.

Foulsham, T., & Kingstone, A. (2017). Are fixations in static natural scenes a useful predictor of attention in the real world?. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *71*(2), 172.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision research*, *51*(17), 1920-1931.

Foulsham, T., Wybrow, D., & Cohn, N. (2016). Reading without words: Eye movements in the comprehension of comic strips. *Applied Cognitive Psychology*, *30*(4), 566-579.

Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2000). Multisensory auditory–somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Cognitive Brain Research*, *10*(1-2), 77-83.

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. *Experimental brain research*, *142*(1), 139-150.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291-325.

Friederici, A. D., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of psycholinguistic research*, *31*(1), 45-63.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive brain research*, *1*(3), 183-192.

Friederici, A. D., & Weissenborn, J. (2007). Mapping sentence form onto meaning: The syntax–semantic interface. *Brain research*, *1146*, 50-58.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211-1221.

Gagne, C. R., & MacEvoy, S. P. (2014). Do simultaneously viewed objects influence scene recognition individually or as groups? Two perceptual studies. *Plos one*, *9*(8), e102819.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*(2), 123-144.

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for “common sense”: An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience*, *8*(2), 89-106.

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, *10*(13), 805-808.

Gerhard, H. E., Wichmann, F. A., & Bethge, M. (2013). How sensitive is the human visual system to the local statistics of natural images?. *PLoS computational biology*, *9*(1).

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of cognitive neuroscience*, *11*(5), 473-490.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology Press.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350-363.

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. *Psychonomic bulletin & review*, *23*(5), 1566-1575.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition*, *12*(6), 878-892.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and cognitive processes*, *25*(2), 149-188.

Graham, D. J., & Field, D. J. (2009). Natural images: Coding efficiency. *Encyclopedia of Neuroscience*, *6*, 19-27.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2014). Visual noise from natural scene statistics reveals human scene category representations. *arXiv preprint arXiv:1411.5331*.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics*, *77*(4), 1239-1251.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464-472.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95-112.

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1358), 1121-1127.

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.*, *27*, 649-677.

Groen, I. I., Ghebreab, S., Prins, H., Lamme, V. A., & Scholte, H. S. (2013). From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, *33*(48), 18814-18824.

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160102.

Guillaume, F., Tinard, S., Baier, S., & Dufau, S. (2016). An ERP Investigation of object-scene incongruity. *Journal of psychophysiology*.

Gunter, T. C., Friederici, A. D., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of cognitive neuroscience*, *12*(4), 556-568.

Gunter, T. C., Stowe, L. A., & Mulder, G. (1997). When syntax meets semantics. *Psychophysiology, 34*(6), 660-676.

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech compared to reading: the P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia*, *38*(11), 1531-1549.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes*, *8*(4), 439-483.

Hansen, N. E., Noesen, B. T., Nador, J. D., & Harel, A. (2018). The influence of behavioral relevance on the processing of global scene properties: An ERP study. *Neuropsychologia*, *114*, 168-180.

Haque, R. U., Inati, S. K., Levey, A. I., & Zaghloul, K. A. (2020). Feedforward prediction error signals during episodic memory retrieval. *Nature communications*, *11*(1), 1-14.

Harel, A., Groen, I. I., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *Eneuro*, *3*(5).

Harel, A., Mzozoyana, M. W., Al Zoubi, H., Nador, J. D., Birken, T. N., Lowe, M. X., & Cant, J. S. (2020). Artificially-generated scenes demonstrate the importance of global scene properties for scene perception. *Neuropsychologia*, 107434.

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. *Proceedings of the National Academy of Sciences*, *93*(2), 623-627.

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10-10.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, *50*(1), 243-271.

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature neuroscience*, *19*(5), 665-667.

Hogendoorn, H., & Burkitt, A. N. (2018). Predictive coding with neural transmission delays: a real-time temporal alignment hypothesis. *bioRxiv*, 453183.

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology*, *30*(1), 47-61.

Holcomb, P. J., & McPherson, W. B. (1994). Event-related brain potentials reflect semantic priming in an object decision task. *Brain and cognition*, *24*(2), 259-276.

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580-593.

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, *72*(7), 1736-1741.

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

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254-1259.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision research*, *47*(26), 3286-3297.

Juan, C. H., & Walsh, V. (2003). Feedback to V1: a reverse hierarchy in vision. *Experimental brain research*, *150*(2), 259-263.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and cognitive processes*, *15*(2), 159-201.

Kadar, I., & Ben-Shahar, O. (2012). A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *Journal of vision*, *12*(13), 16-16.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision research*, *46*(11), 1762-1776.

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *Journal of Neuroscience*, *31*(7), 2488-2492.

Kok, P., Failing, M. F., & De Lange, F. P. (2014). Prior expectations evoke stimulus templates in the primary visual cortex. *Journal of Cognitive Neuroscience*, *26*(7), 1546-1554.

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265-270.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, *17*(1), 26-49.

Kumar, M., Federmeier, K. D., & Beck, D. M. (2020). The N300: An Index For Predictive Coding Of Complex Visual Objects and Scenes. *bioRxiv*.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, *1146*, 23-49.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621-647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203-205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161-163.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In *Handbook of psycholinguistics* (pp. 659-724). Academic Press.

Kvasova, D., Garcia-Vernet, L., & Soto-Faraco, S. (2019). Characteristic sounds facilitate object search in real-life scenes. *Frontiers in psychology*, *10*, 2511.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious?. *Perspectives on psychological science*, *7*(1), 18-27.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701-710.

Lamme, V. A. (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural networks*, *17*(5-6), 861-872.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, *23*(11), 571-579.

Langham, M., Hole, G., Edwards, J., & O'Neil, C. (2002). An analysis of 'looked but failed to see' accidents involving parked police vehicles. *Ergonomics*, *45*(3), 167-185.

Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 471.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920-933.

Lauer, T., Willenbockel, V., Maffongelli, L., & Võ, M. L. H. (2020). The influence of scene and object orientation on the scene consistency effect. *Behavioural Brain Research*, *394*, 112812.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental brain research*, *158*(4), 405-414.

Laurienti, P. J., Wallace, M. T., Maldjian, J. A., Susi, C. M., Stein, B. E., & Burdette, J. H. (2003). Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Human brain mapping*, *19*(4), 213-223.

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155-168.

Li, W., Piëch, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience*, *7*(6), 651.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596-9601.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in human neuroscience*, *8*, 213.

Loschky, L. C., Hutson, J. P., Smith, M. E., Smith, T. J., & Magliano, J. P. (2018). Viewing static visual narratives through the lens of the Scene Perception and Event Comprehension Theory (SPECT). *Empirical comics research: Digital, multimodal, and cognitive methods*, 217-238.

Loschky, L. C., & Larson, A. M. (2008). Localized information is necessary for scene categorization, including the natural/man-made distinction. *Journal of Vision*, *8*(1), 4-4.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*(4), 513-536.

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2019). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in cognitive science*.

Mack, A., & Clarke, J. (2012). Gist perception requires attention. *Visual Cognition*, *20*(3), 300-327.

Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and cognition*, *47*, 6-16.

Maguire, J. F., & Howe, P. D. (2016). Failure to detect meaning in RSVP at 27 ms per picture. *Attention, Perception, & Psychophysics*, *78*(5), 1405-1413.

Mahon, P. T. (1981). Report of the royal commission to inquire into the crash on Mount Erebus. *Antarctica, of a DC-10 aircraft operated by Air New Zealand Limited: Wellington, Government Printer*.

Mahzouni, G. (2019). *The Top-Down Influences of Characteristic Sounds on Visual Search Performance in Realistic Scenes* (Doctoral dissertation, San Jose State University).

Malcolm, G. L., Groen, I. I., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in cognitive sciences*, *20*(11), 843-856.

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, *9*(11), 8-8.

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, *10*(2), 4-4.

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond gist: Strategic and incremental information accumulation for scene categorization. *Psychological science*, *25*(5), 1087-1097.

Mandler, J. M., & Ritchey, G. H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(4), 386.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

McDonald, J. J., Teder-SaÈlejaÈrvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, *407*(6806), 906-908.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

McManus, J. N., Li, W., & Gilbert, C. D. (2011). Adaptive shape processing in primary visual cortex. *Proceedings of the National Academy of Sciences*, *108*(24), 9739-9746.

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, *36*(1), 53-65.

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience*, *7*(10), 3215-3229.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual–auditory object recognition in humans: a high-density electrical mapping study. *Cerebral Cortex*, *14*(4), 452-465.

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia*, *48*(2), 507-517.

Mudrik, L., Shalgi, S., Lamy, D., & Deouell, L. Y. (2014). Synchronous contextual irregularities affect early scene processing: Replication and extension. *Neuropsychologia*, *56*, 447-458.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological cybernetics*, *66*(3), 241-251.

Munneke, J., Brentari, V., & Peelen, M. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in psychology*, *4*, 552.

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision research*, *46*(5), 614-621.

Neri, P. (2014). Semantic control of feature extraction from natural scenes. *Journal of Neuroscience*, *34*(6), 2374-2388.

Newell, F. N. (2004). Cross-modal object recognition.

Nijhawan, R., & Wu, S. (2009). Compensating time delays with neural predictions: are predictions sensory or motor?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1891), 1063-1078.

Oliva, A. (2005). Gist of the scene. In *Neurobiology of attention* (pp. 251-256). Academic press.

Oliva, A. (2013). Scene Perception. Chapter in the New Visual Neurosciences, Eds John S. Werner and Leo. M. Chalupa.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive psychology*, *34*(1), 72-107.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive psychology*, *41*(2), 176-210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, *42*(3), 145-175.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785-806.

Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. *Language and Cognitive Processes*, *8*(4), 413-437.

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*, 519-526.

Peirce, J., & MacAskill, M. (2018). *Building experiments in PsychoPy*. Sage.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*(1), 49-70.

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*(4180), 965-966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, *2*(5), 509.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*(2), 270-279.

Prasad, S., & Galetta, S. L. (2011). Anatomy and physiology of the afferent visual system. In *Handbook of clinical neurology* (Vol. 102, pp. 3-19). Elsevier.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79-87.

Rauss, K., Schwartz, S., & Pourtois, G. (2011). Top-down effects on early visual processing in humans: A predictive coding framework. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1237-1253.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

Reinitz, M. T., Wright, E., & Loftus, G. R. (1989). Effects of semantic priming on visual encoding of pictures. *Journal of Experimental Psychology: General*, *118*(3), 280.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition?. *Vision research*, *44*(19), 2301-2311.

Rensink, R. A. (2000). Scene perception. *Encyclopedia of psychology*, *7*, 151-155.

Rock, I. E. (1997). *Indirect perception*. The MIT Press.

Romei, V., Murray, M. M., Merabet, L. B., & Thut, G. (2007). Occipital transcranial magnetic stimulation has opposing effects on visual and auditory stimulus detection: implications for multisensory interactions. *Journal of Neuroscience*, *27*(43), 11465-11472.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature neuroscience*, *5*(7), 629-630.

Rumelhart, D. E. (1970). A multicomponent theory of the perception of briefly exposed visual displays. *journal of Mathematical Psychology*, *7*(2), 191-218.

Rummukainen, O., & Mendonca, C. (2016). Reproducing reality: Multimodal contributions in natural scene discrimination. *ACM Transactions on Applied Perception (TAP)*, *14*(1), 1-19.

Rummukainen, O., Radun, J., Virtanen, T., & Pulkki, V. (2014). Categorization of natural dynamic audiovisual scenes. *PloS one*, *9*(5), e95848.

Sanocki, T. (2013). Facilitatory priming of scene layout depends on experience with the scene. *Psychonomic bulletin & review*, *20*(2), 274-281.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, 374-378.

Schendan, H. E., & Kutas, M. (2002). Neurophysiological evidence for two processing times for visual object identification. *Neuropsychologia*, *40*(7), 931-945.

Schendan, H. E., & Kutas, M. (2007). Neurophysiological evidence for the time course of activation of global shape, part, and local contour representations during visual object categorization and memory. *Journal of Cognitive Neuroscience*, *19*(5), 734-749.

Schmitt, M., Postma, A., & De Haan, E. (2000). Interactions between exogenous auditory and visual spatial attention. *The Quarterly Journal of Experimental Psychology Section A*, *53*(1), 105-130.

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental psychology*, *55*(2), 121-132.

Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W., & Lamme, V. A. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision*, *9*(4), 29-29.

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current opinion in neurobiology*, *15*(4), 454-458.

Shafer-Skelton, A., & Brady, T. F. (2019). Scene layout priming relies primarily on low-level features rather than scene layout. *Journal of vision*, *19*(1), 14-14.

Sherman, B. E., & Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences*, *117*(37), 22760-22770.

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of cognitive neuroscience*, *20*(11), 2037-2057.

Sitnikova, T., Kuperberg, G., & Holcomb, P. J. (2003). Semantic integration in videos of real–world events: An electrophysiological investigation. *Psychophysiology*, *40*(1), 160-164.

Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology*, *17*(19), 1680-1685.

Smith, M. E., & Loschky, L. C. (2019). The influence of sequential predictions on scene-gist recognition. *Journal of vision*, *19*(12), 14-14.

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & psychophysics*, *59*(1), 1-22.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. *BioRxiv*, 677237.

Spotorno, S., Malcolm, G. L., & Tatler, B. W. (2014). How context information and target information guide the eyes from the first epoch of search in real-world scenes. *Journal of Vision*, *14*(2), 7-7.

Spotorno, S., Malcolm, G. L., & Tatler, B. W. (2015). Disentangling the effects of spatial inconsistency of targets and distractors when searching in realistic scenes. *Journal of Vision*, *15*(2), 12-12.

Straube, S., & Fahle, M. (2010). The electrophysiological correlate of saliency: Evidence from a figure-detection task. *Brain research*, *1307*, 89-102.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, *26*(2), 212-215.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science*, *314*(5803), 1311-1314.

Summerfield, C., & Koechlin, E. (2008). A neural representation of prior information during perceptual inference. *Neuron*, *59*(2), 336-347.

Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, *9*, 19.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta psychologica*, *135*(2), 77-99.

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

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness-an update.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, *113*(4), 766.

Truman, A., & Mudrik, L. (2018). Are incongruent objects harder to identify? The functional significance of the N300 component. *Neuropsychologia*, *117*, 222-232.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, *247*(4940), 301-306.

Ullman, S. (1980). Against direct perception. *Behavioral and Brain Sciences*, *3*(3), 373-381.

Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral cortex*, *5*(1), 1-11.

Underwood, G. (2005). *Cognitive processes in eye guidance*. Oxford University Press.

Van Petten, C. (1995). Words and sentences: Event-related brain potential measures. *Psychophysiology*, *32*(6), 511-525.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and language*, *97*(3), 279-293.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *Journal of cognitive neuroscience*, *13*(4), 454-461.

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, *24*(11), 1256-1262.

Viggiano, M. P., & Kutas, M. (2000). Overt and covert identification of fragmented objects inferred from performance and electrophysiological measures. *Journal of Experimental Psychology: General*, *129*(1), 107.

Võ, M. L. H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current opinion in psychology*.

Võ, M. L. H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, *10*(3), 14-14.

Võ, M. L. H., & Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics*, *73*(6), 1742.

Võ, M. L. H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological science*, *24*(9), 1816-1823.

Walker, S., Stafford, P., & Davis, G. (2008). Ultra-rapid categorization requires visual attention: Scenes with multiple foreground objects. *Journal of Vision*, *8*(4), 21-21.

Walther, D. B., & Shen, D. (2014). Nonaccidental properties underlie human categorization of complex natural scenes. *Psychological science*, *25*(4), 851-860.

Wang, L., Jensen, O., Van den Brink, D., Weder, N., Schoffelen, J. M., Magyari, L., ... & Bastiaansen, M. (2012). Beta oscillations relate to the N400m during language comprehension. *Human brain mapping*, *33*(12), 2898-2912.

Wang, S., Yang, C., Liu, Y., Shao, Z., & Jackson, T. (2017). Early and late stage processing abnormalities in autism spectrum disorders: An ERP study. *PloS one*, *12*(5).

Ward, J. A. (2000). A software based low vision aid.

Willems, R. M., Özyürek, A., & Hagoort, P. (2008). Seeing and hearing meaning: ERP and fMRI evidence of word versus picture integration into a sentence context. *Journal of Cognitive Neuroscience*, *20*(7), 1235-1249.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, *73*(6), 1650-1671.

Wu, J. (2011). *Introduction to neural dynamics and signal transmission delay* (Vol. 6). Walter de Gruyter.

Yuan, J., Zhang, Q., Chen, A., Li, H., Wang, Q., Zhuang, Z., & Jia, S. (2007). Are we sensitive to valence differences in emotionally negative stimuli? Electrophysiological evidence from an ERP study. *Neuropsychologia*, *45*(12), 2764-2771.

# Appendix

## Appendix A

### *Creation of Image Series for Study 1*

The intention when constructing the series was to create progressions which mimicked movement through an environment towards a destination, while reducing instances of over-similarity across viewpoints and avoiding sudden 'jumps' in the progression. Accordingly, variations in geographical distances between approach images needed to be considered across series, mainly due to the differing constraints imposed by the superordinate categories. For example, the distance between points during a progression through a house to, say, a bedroom would be inherently shorter when compared to the points of progression towards a beach. An approach to a bedroom might begin with a view of a stairway across an atrium, followed by an image on the stairway, one at the top of the stairway turning onto a hallway, another at the mid-point of a hallway, and one turning the corner to show a bedroom doorway prior to the target being shown. In doing so, each transition of the approach would be accounted for, although the geographical distance covered would be relatively short. If the progression was towards a beach, on the other hand, then mirroring the distances between approach images from the bedroom series would result in five very similar viewpoints, almost indistinguishable from one another under the processing constraints of rapid presentation. As a consequence, in such instances we somewhat 'stretched out' the approach, so that it covered a greater geographical distance but at the same time maintained the principle of showing each transition in the journey, say from a carpark, down a pathway and between dunes before arriving at the beach. Again, care was taken to avoid sudden jumps in the narrative, so that the spatiotemporal relationship between successive leading images always remained apparent. This could be considered an attempt to instil a 'semantic flow' within each series, with each of the transitional points of the approach represented in a manner which maintained the sense of a progression throughout.

There are several other important points to note relating to the construction of series. Firstly, the destination scene could not be immediately determined from the earliest leading images. This was due to there being similar progressions across many series, in both interior-destination series (for instance, ‘bathroom’ and ‘bedroom’ targets would have similar approaches, involving stairways, hallways, etc.), and exterior-destination series (where many progressions shared similarities, such as traversing pavements, pathways and carparks). Furthermore, the eventual superordinate category of the target could not be anticipated at the start of the series: the approach images might represent a journey out in the open but with an indoor destination scene, or vice versa, such as walking across a garden before entering an outbuilding. Additionally, approaches frequently passed through other target categories. For example, images of a high street – a target category on some trials – might be passed through within the approach images of a series with a ‘shop’ target. It should be reiterated that this potential interplay across trials was at the category level, not the exemplar level, as no scenery (whether approach image or destination) was repeated at any point during the task.

Secondly, a balance had to be struck in terms of the final approach image representing a viewpoint geographically close enough to heighten expectations as to the destination, while trying to minimise the amount of similarity in low-level features across these two images. This was to ensure that performance was based on semantic prediction rather than simply on the repetition of low-level visual information. Therefore, while some features of a destination might be visible within the later approach images (such as the ocean on the horizon while progressing towards a ‘beach’ target, or the corner of a table and chair seen through a doorway prior to reaching a ‘dining room’ target) care was taken to maintain substantial differences in both the viewpoint and available visual features between the approach images and the destination scene. This practice was considered in line with the overarching tenet driving the construction of each series, namely that the progressions should mirror as closely as possible how individuals experience the environments in which they are embedded through the course of daily life.

Thirdly, the inclusion of people within images was kept to a minimum. It was not considered necessary to exclude pedestrians, shoppers, etc. from the sequences, as the aim was to represent environments in their usual state. However, care was taken to ensure that individuals within sceneries did not become a distraction from the experimental task, and so no images included people positioned close in the foreground or looking directly at the observer. Finally, all images (with the exception of multi-storey carparks) were of sceneries outside the county of the university's location, in an attempt to limit any potential confounds due to familiarity with the specific exemplars used.

**Appendix B**

*List of Scene Categories Used in Study 1*

ART GALLERY; BATHROOM; BEACH; BEDROOM; CARPARK; CHURCH; DINING ROOM; ENTRANCE HALL; FIELD; GARDEN; GRAVEYARD; HIGH STREET; KITCHEN; LIVING ROOM; MULTISTOREY CARPARK; OUTBUILDING; PARK; PETROL STATION; PUB; QUAY; RECYCLING AREA; RETAIL STORE; RIVER; ROAD; SHOP; SPORTS PITCH; SUPERMARKET; TAKEAWAY; TRAIN STATION; WOODS

## Appendix C

*Statistical Analyses from Experiment 4*

| Window | Factor | *df* | *F* | *t* | *p* | $\eta p^2$ | *r* |
|---|---|---|---|---|---|---|---|
| 175-250 ms | Hemisphere | 1, 23 | 10.21 | | .004* | .31 | |
| | Region | 1.17, 27.00 | 51.28 | | .000* | .69 | |
| | Congruency | 1, 23 | 0.04 | | .842 | .00 | |
| | Hemisphere*Region*Congruency | 1.61, 37.04 | 0.54 | | .552 | .02 | |
| | Hemisphere*Region | 2, 46 | 0.95 | | .396 | .04 | |
| | Hemisphere*Congruency | 1, 23 | 0.66 | | .424 | .03 | |
| | Region*Congruency | 2, 46 | 15.68 | | .000* | .41 | |
| | Paired t-tests (for R*C interaction) | | | | | | |
| | Frontal | 23 | | 2.54 | .018* | | .47 |
| | Centro-parietal | 23 | | -0.14 | .893 | | .03 |
| | Parieto-occipital | 23 | | -2.08 | .048* | | .40 |
| 300-500 ms | Hemisphere | 1, 23 | 6.39 | | .019* | .22 | |
| | Region | 1.21, 27.90 | 45.37 | | .000* | .66 | |
| | Congruency | 1, 23 | 6.16 | | .021* | .21 | |
| | Hemisphere*Region*Congruency | 2, 46 | 2.08 | | .136 | .08 | |
| | Hemisphere*Region | 2, 46 | 1.33 | | .274 | .06 | |
| | Hemisphere*Congruency | 1, 23 | 2.65 | | .117 | .10 | |
| | Region*Congruency | 1.46, 33.65 | 32.92 | | .000* | .59 | |
| | Paired t-tests (for R*C interaction) | | | | | | |
| | Frontal | 23 | | 5.37 | .000* | | .75 |
| | Centro-parietal | 23 | | 2.40 | .025* | | .45 |
| | Parieto-occipital | 23 | | -1.57 | .129 | | .31 |
| 500-700 ms | Hemisphere | 1, 23 | 6.87 | | .015* | .23 | |
| | Region | 1.52, 34.92 | 44.53 | | .000* | .66 | |
| | Congruency | 1, 23 | 0.36 | | .553 | .02 | |
| | Hemisphere*Region*Congruency | 2, 46 | 3.07 | | .056 | .12 | |
| | Hemisphere*Region | 2, 46 | 0.43 | | .656 | .02 | |
| | Hemisphere*Congruency | 1, 23 | 5.72 | | .025* | .20 | |
| | Paired t-tests (for H*C interaction) | | | | | | |
| | Left hemisphere | 23 | | 1.37 | .185 | | .32 |
| | Right hemisphere | 23 | | -0.05 | .958 | | .01 |
| | Region*Congruency | 2, 46 | 34.05 | | .000* | .60 | |
| | Paired t-tests (for R*C interaction) | | | | | | |
| | Frontal | 23 | | 4.57 | .000* | | .69 |
| | Centro-parietal | 23 | | -0.04 | .972 | | .01 |
| | Parieto-occipital | 23 | | -2.41 | .025* | | .45 |
| 175-250 ms | Hemisphere | 1, 23 | 6.57 | | .017* | .22 | |
| (Lateral P2) | Congruency | 1, 23 | 5.81 | | .024* | .20 | |
| | Hemisphere*Congruency | 1, 23 | 1.26 | | .274 | .05 | |

*Note.* The three windows of the main analysis were analysed with 2x3x2 ANOVAs. The additional analysis of the 175-250 ms window for the lateral Parieto-occipital region was analysed with a 2x2 ANOVA. * denotes *p* < .05

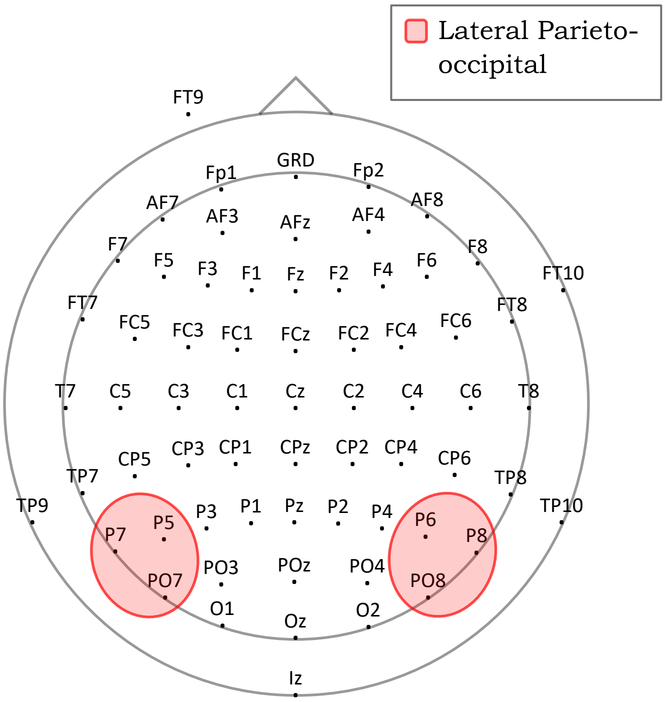
## Appendix D

### *Additional Analysis: Lateral P2*

In our initial analyses we found a significant effect of Congruency within the P2 time-window at posterior sites. However, in the interest of completeness we decided further investigation would be insightful. Previous scene processing research concerned with the P2 component has found effects to be maximal at sites more lateral than our initial ROIs (Hansen et al., 2018; Harel et al., 2016; Harel et al., 2020). Consequently, we created a Lateral Parieto-occipital ROI comprising six electrodes (split equally across hemispheres). The position of these regions was chosen to mirror previous work as closely as possible. Specifically, Harel and colleagues (2016; 2020) use a lateral region including eight electrodes across the two hemispheres (P5/P6, P7/P8, P9/P10 and PO7/PO8). Exact duplication of this setup was not possible, as instead of the electrode pair P9/P10 our array included TP9/TP10, which were located near the mastoids, and had been used as our re-referencing electrodes. Therefore, our lateral regions consisted of P5/P6, P7/P8 and PO7/PO8 (see Figure D1).

**Figure D1**

*Map of Electrode Placement Including the Lateral ROIs*



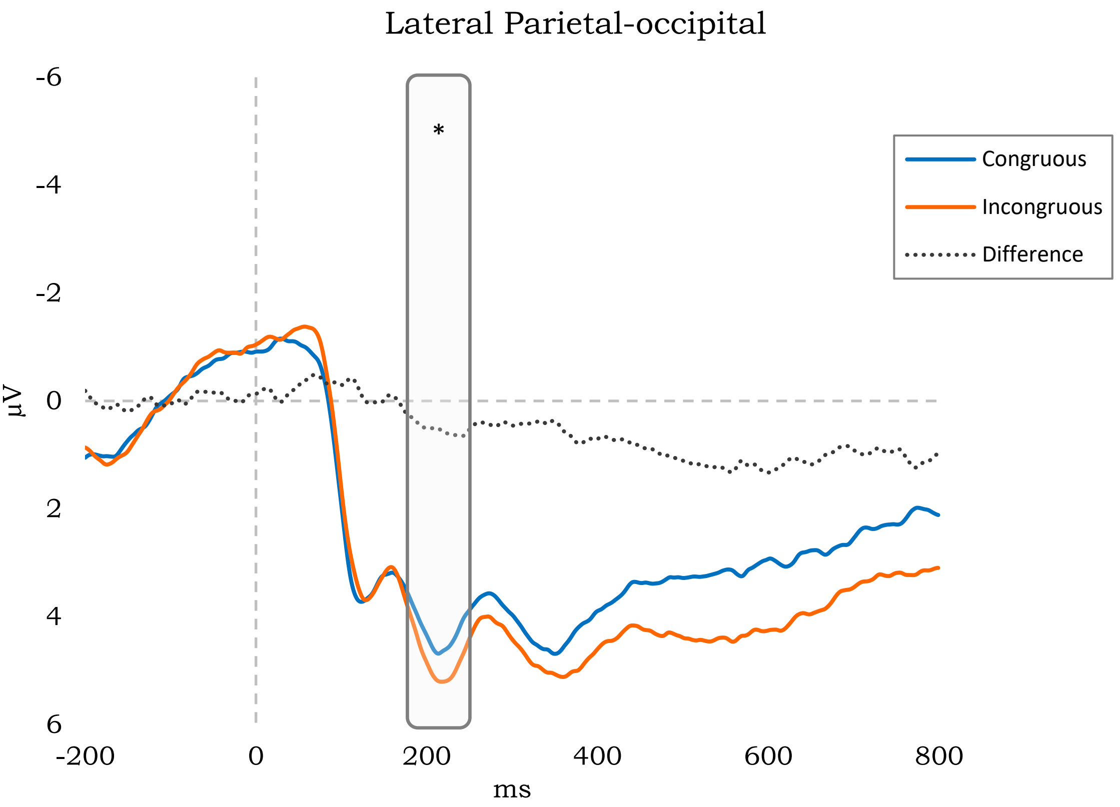
*Note.* FT9 was removed from the cap and placed on the left cheekbone to monitor blinks.

Analysis was conducted on the mean amplitudes for the same time-period as before (175-250 ms) using a 2 (Hemisphere: Left; Right) x 2 (Congruency: Congruous; Incongruous) repeated-measures ANOVA. This revealed a main effect of Congruency, $F(1, 23) = 5.81$, $p = .024$, $\eta p^2 = .20$, with more positive amplitudes for Incongruous ($M$ = 4.78 µV) than Congruous ($M$ = 4.26 µV) trials.

The Hemisphere x Congruency interaction did not reach significance (*p* = .274). See Figure D2 for grand averaged ERPs.

**Figure D2**

*Grand-averaged ERPs for the Lateral Parieto-occipital Region, Collapsed Across Hemispheres*



*Note.* Blue lines represent amplitudes for Congruous trials and orange lines represent amplitudes for Incongruous trials. Dotted line represents the difference wave (Incongruous minus Congruous). Waveforms low-pass filtered at 30Hz for display purposes (*n* = 24). Grey box represents the time-window of interest. * denotes *p* < .05.

## Appendix E

*List of Object and Sound Pairings Used in Study 2*

| Target object | Non-target object | Congruous Sound | Incongruous Sound |
|---|---|---|---|
| Ambulance | Flag | Ambulance | Gong |
| Beer can | Tent | Beer can | Sink |
| Boat | Wheelbarrow | Boat | Ambulance |
| Bowling ball | Beverage cup | Bowling ball | Soda can |
| Buzzsaw | Garage door cable | Buzzsaw | Fire extinguisher |
| Cash register | Stool | Cash register | Lawnmower |
| Clock | Tape measure | Clock | Drill |
| Drill | Dustpan | Drill | Clock |
| Electric toothbrush | Towel | Electric toothbrush | Beer can |
| Fire extinguisher | Kitchen roll | Fire extinguisher | Remote control car |
| Front desk bell | Suitcase | Front desk bell | Iron |
| Gong | Banner | Gong | Typewriter |
| Hand dryer | Hand soap | Hand dryer | Front desk bell |
| Iron | Coat hanger | Iron | Cash register |
| Lawnmower | Traffic cone | Lawnmower | Printer |
| Microwave | Dishtowel | Microwave | Watch |
| Printer | Remote control | Printer | Hand dryer |
| Remote control car | Picture frame | Remote control car | Microwave |
| Scoreboard | Football | Scoreboard | Bowling ball |
| Sink | Loofah | Sink | Tape recorder |
| Soda can | Hot air balloon | Soda can | Electric toothbrush |
| Tape recorder | Light fixture | Tape recorder | Buzzsaw |
| Typewriter | Cushion | Typewriter | Scoreboard |
| Watch | Slipper | Watch | Boat |
| Alarm clock | Cat basket | Alarm clock | Garbage truck |
| Barbecue | Chimney | Barbecue | Chainsaw |
| Blender | Child's toy | Blender | Alarm clock |
| Bongos | Birdhouse | Bongos | Smoke alarm |
| Bus | Fire hydrant | Bus | Electric razor |
| Cashpoint | Brochures | Cashpoint | Kettle |
| Chainsaw | Axe | Chainsaw | Dog |
| Dog | Hanging plant | Dog | Shower |
| Electric razor | Toilet brush | Electric razor | Bus |
| Film projector | Wall painting | Film projector | Helicopter |
| Frog | Hat | Frog | Record player |
| Garbage truck | Clock face | Garbage truck | Train |
| Hair dryer | Pedal bin | Hair dryer | Frog |
| Helicopter | Bus stop | Helicopter | Cashpoint |
| Kettle | Socks | Kettle | Motorcycle |
| Microphone | Calendar | Microphone | Washing machine |
| Motorcycle | Weathervane | Motorcycle | Hairdryer |

| Target object | Non-target object | Congruous sound | Incongruous sound |
|---|---|---|---|
| Record player | Yoga mat | Record player | Sprinkler |
| Saucepan | Rug | Saucepan | Microphone |
| Shower | Toilet roll | Shower | Bongos |
| Smoke alarm | Hammer | Smoke alarm | Film projector |
| Sprinkler | Wine bottle | Sprinkler | Barbecue |
| Train | Windmill | Train | Saucepan |
| Washing machine | Ladle | Washing machine | Blender |
| Aquarium | Shoe | Aquarium | Guitar |
| Bicycle | Streetlamp | Bicycle | Skateboard |
| Boiler | Bowl | Boiler | Spray bottle |
| Buoy | Blimp | Buoy | Fireplace |
| Car | Post box | Car | Aquarium |
| Ceiling fan | Ketchup bottle | Ceiling fan | Lighter |
| Cuckoo clock | Book | Cuckoo clock | Electric guitar |
| Electric guitar | Stage light | Electric guitar | Cuckoo clock |
| Extractor fan | Pitcher | Extractor fan | Bicycle |
| Fireplace | Open sign | Fireplace | Jackhammer |
| Frying pan | Pineapple | Frying pan | Rocking chair |
| Guitar | Teddy bear | Guitar | Vacuum |
| Harp | Lamp | Harp | Frying pan |
| Jackhammer | Sailboat | Jackhammer | Ceiling fan |
| Lighter | Wallet | Lighter | Radio |
| Mobile phone | Recycling bin | Mobile phone | Whistle |
| Radio | Dreamcatcher | Radio | Harp |
| Rocking chair | Cafetiere | Rocking chair | Telephone |
| Sewing machine | Chandelier | Sewing machine | Mobile phone |
| Skateboard | Picnic basket | Skateboard | Buoy |
| Spray bottle | Handbag | Spray bottle | Car |
| Telephone | Fire poker | Telephone | Extractor fan |
| Vacuum | Dartboard | Vacuum | Sewing machine |
| Whistle | Baseball glove | Whistle | Boiler |

## Appendix F

*Root Mean Squared Decibel Levels for Object Sounds Used in Study 2*

| Sound | RMS (dB) | Sound | RMS (dB) |
|---|---|---|---|
| Alarm clock | -18.3 | Lighter | -18 |
| Ambulance | -18.6 | Light flicker | -18.7 |
| Arcade game | -19.9 | Lion | -18.7 |
| Barbecue | -18.4 | Lorry | -18.1 |
| Bear | -18.1 | Microphone | -18.9 |
| Beer can | -19.2 | Microwave | -18.6 |
| Bees | -18.7 | Mobile phone | -18.1 |
| Bird | -18.5 | Motorbike | -18.3 |
| Blender | -18.8 | Mouse | -18.1 |
| Boat foghorn | -19.7 | Owl | -18.6 |
| Boiler | -19 | Parrot | -19.7 |
| Bongos | -18.1 | Pedestrian crossing | -18.4 |
| Bowling ball | -18.4 | Piano | -18.3 |
| Buoy bell | -18.1 | Pig | -18.6 |
| Bus | -18 | Printer | -18.5 |
| Car | -18.3 | Radio tuning | -18.4 |
| Car horn | -18.8 | Razor | -18.4 |
| Cash machine | -18.8 | Record player | -19.7 |
| Cash register | -18.7 | Rocking chair | -19 |
| Ceiling fan | -18 | Rollercoaster | -18.3 |
| Chainsaw | -18.1 | Rooster | -19.4 |
| Cow | -18.1 | Saucepan | -19.2 |
| Cuckoo clock | -18.7 | Saw | -18.8 |
| Desk bell | -18.2 | Scissors | -18.9 |
| Desk fan | -18 | Scoreboard | -18.5 |
| Dog | -18.7 | Seagull | -18.8 |
| Doorbell | -18.4 | Sewing machine | -18.6 |
| Doorknocker | -19 | Sheep | -18.3 |
| Drill | -18.3 | Shower | -19.1 |
| Duck | -18.6 | Sink | -19.6 |
| Electric Guitar | -18.4 | Skateboard | -18.3 |
| Extractor fan | -18.4 | Sleighbells | -19.6 |
| Film projector | -18.3 | Smoke alarm | -18.1 |
| Fire extinguisher | -18 | Soda can | -18.8 |
| Fireplace | -18.5 | Spray bottle | -19.9 |
| Fish tank | -18.3 | Spray can | -18.2 |
| Flag | -18 | Sprinkler | -19.4 |
| Frog | -18.9 | Steam train | -18.6 |
| Garbage truck | -18.4 | Swing | -18.9 |
| Gate creak | -18.4 | Tape recorder | -18.8 |
| Gong | -19.3 | Telephone | -18 |

| Sound | RMS (dB) | Sound | RMS (dB) |
|---|---|---|---|
| Guineapig | -19 | Toilet | -18.5 |
| Guitar | -19.4 | Toothbrush | -18.2 |
| Hairdryer | -18.4 | Toy car | -19.2 |
| Hand dryer | -18 | Train | -18.8 |
| Harp | -18.1 | Tram | -19.9 |
| Helicopter | -18.4 | TV static | -18.9 |
| Horse | -18.2 | Typewriter | -20 |
| Ice-cream van | -18.4 | Vacuum | -18.2 |
| Ice cubes | -19 | Wall clock | -19 |
| Iron | -19.6 | Washing machine | -19.4 |
| Jackhammer | -19.2 | Watch | -19.4 |
| Jacuzzi | -18.3 | Whistle | -18.4 |
| Kettle | -19.1 | Windchimes | -18.3 |
| Keyboard | -19 | Wineglass | -18.5 |
| Lawnmower | -19.9 | Woodpecker | -18.6 |
| Leaf blower | -19 | Xylophone | -18.9 |