Investigating the Neural Bases of Crossmodal Correspondences

By

Kelly R. McCormick

Doctor of Philosophy

Psychology

Lynne C. Nygaard, Ph.D.
Advisor

Krish Sathian, M.B.B.S., Ph.D.
Advisor

Lawrence Barsalou, Ph.D.
Committee Member

Daniel Dilks, Ph.D.
Committee Member

Harold Gouzoules, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Investigating the Neural Bases of Crossmodal Correspondences

By

Kelly R. McCormick

M.A., Emory University, 2012

Advisors: Lynne C. Nygaard, Ph.D. and Krish Sathian M.B.B.S., Ph.D.

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Psychology

2019

Abstract

Investigating the Neural Bases of Crossmodal Correspondences

By Kelly R. McCormick

Across diverse test populations, people have been found to reliably associate stimulus features in different sensory modalities. Such 'crossmodal correspondences' have been found for a number of audiovisual domains. For example, people associate higher-pitched sounds with higher spatial elevation, compared to lower-pitched sounds which are associated with lower spatial elevation. Despite the pervasiveness of crossmodal correspondences, the neural and cognitive mechanisms underlying these phenomena are currently unknown. The overarching goal of this dissertation research has been to better understand the neural systems that provide a basis for such mappings. Three functional neuroimaging experiments were conducted to investigate the neural basis of three audiovisual correspondences: auditory pitch-visuospatial elevation, pseudoword-object shape, and pitch-object size. To identify systems sensitive to each of these correspondences, we contrasted blood-oxygen-level dependent (BOLD) activity for multimodally congruent and incongruent audiovisual stimulus couplings. In addition, three independent localizer tasks were employed to functionally define neural systems involved in multisensory integration, magnitude, and semantics, all of which have been theorized to play a role in crossmodal mappings.

The results of the pitch-elevation experiment (Chapter 2) did not indicate involvement of the functionally localized systems, although activation overlapped with the semantic control condition, possibly reflecting phonological processing. However, patterning of the congruency-related activity was consistent with a possible basis in multisensory attention. The results of the pseudoword-shape experiment (Chapter 3) provided no evidence for semantic mediation, and limited evidence for processes relating to multisensory integration and magnitude estimation as possible underlying mechanisms. Again countering our predictions, support was found for a relationship between pseudoword-object shape mapping and multisensory attention and/or phonological processing. Behavioral results in the experiment on the pitch-size experiment (Chapter 4) were heterogeneous making it difficult to meaningfully interpret neural activity. Together, these experiments provide new insight into the neural basis of the pitch-elevation and pseudoword-shape correspondences and offer a novel and generalizable approach that may be used in future research. In addition to enriching our understanding of crossmodal correspondences more generally, these findings are important for understanding the mechanisms underlying sound-symbolic mappings in language and how words and sounds are mapped to meaning in the brain.

Investigating the Neural Bases of Crossmodal Correspondences

By

Kelly R. McCormick

M.A., Emory University, 2012

Advisors: Lynne C. Nygaard, Ph.D. and Krish Sathian M.B.B.S., Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Psychology
2019

Acknowledgments

First, I would like to thank my dissertation committee, Lynne Nygaard, Krish Sathian, Larry Barsalou, Danny Dilks, and Harold Gouzoules. Their thoughtful input and feedback throughout the course of this dissertation research has helped to shape not only the project but has also shaped me as an intellectual. I specifically thank my advisors, Lynne Nygaard, and Krish Sathian for their enduring support and collaboration.

I am particularly indebted to Lynne Nygaard and the members of Nylaab for fostering such a positive and productive learning environment. My labmates, Christina Tzeng, Eva Lewandowski, and Lauren Clepper-Defife have contributed to this project immeasurably, offering input throughout the experiment design process. The research in this dissertation was supported by undergraduate research assistants from Nylaab. I am most grateful to Tami Feng, Valentin Lazar, Sara List, Jee Young Kim, Will Kelly, for their contributions to my projects.

Thanks to Krish Sathian and to my labmates for welcoming me into their lab. Thanks to Simon Lacey for collaboration, camaraderie, and explanations of cricket. My utmost appreciation goes to Randall Stilla for training me in both conducting MRI scanning, and then processing and analysis of neuroimaging data. None of the experiments in this dissertation would have been possible without Randall's contributions and guidance. Both Randall and Simon have contributed immensely to our publications based on these studies, including figures I have included in this thesis.

I would also like to acknowledge Marty Sereno, whose amazing courses in systems neuroscience and symbolic systems captivated me as an undergraduate and ultimately put me on the course for this dissertation research.

My heartfelt thanks to the friends and family who have shared in the journey. Thanks to Adam Long for supporting these efforts with caffeine, perspective, belays, and editing. Thanks to Mary and Joe Hutson, Eva Salina Primack, Mike McCormick, Melanie and Hux Pincus, Tiffany Hopkins, Nawaf Almoosa, Peter Kiley, Jess Locke, Francesca Peacocke, Tom Hennessey, Samrawit Yisahak, Lado Botchorishvili, Kelly and Tyler MacNiven, The Kahn family, Yula Paluy, Esther Papies, Umay Suanda, Zanna Clay, Sean Arlt, Daniel Schultz, Tim Plotke and the dirtbags of Hidden Valley, and the denizens of Glittertown. Thank you all for being a source of inspiration an encouragement.

Special thanks goes to my parents, Patrick and Cabell McCormick for their unwavering love and support. This stability has allowed me to take risks and pursue ideas that might not always sound like elements of a productive career path (*nonsense*, *arbitrariness*). I am profoundly grateful for my family.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Behavioral research has identified a host of crossmodal mappings or correspondences that are consistently shared across perceivers. In such mappings, a perceptual attribute in one sensory modality is reliably matched or associated with an attribute in another modality (Spence & Parise, 2012). Crossmodal mappings have been demonstrated between a number of auditory and visual dimensions. For example, lower-pitch sounds are consistently mapped to relatively larger object size, thicker shape, darker visual stimuli, amoeboid shapes, and lower positioning in vertical space, whereas higher-pitch sounds are mapped to smaller object size, thinner shape, brighter visual stimuli, more pointed shapes, and higher positioning in vertical space (Bonetti & Costa, 2017; Dolscheid, Hunnius, Casasanto, & Majid, 2014; Marks, 1974; Marks et al., 1987; Melara & O'Brien, 1987). However, to date we have a very limited understanding of the neural underpinnings of these phenomena.

How and why do we align certain sensory experiences across modalities, and what are the neural bases for crossmodal mapping phenomena? While there is a wealth of behavioral research examining cross-sensory associations, and extensive neuroscientific research on multisensory processing, these two lines of research have until now remained largely independent. A major goal of the present studies is to begin to merge these fields and converge on a productive account of crossmodal mappings in the brain. The present set of experiments explores the neural underpinnings of representations of crossmodal correspondences across the senses, and advances an approach for examining how neural activity reliably reflects individuals' mappings of congruency of stimuli across modalities.

Behavioral insights into crossmodal mappings

Behavioral researchers have documented numerous behavioral effects and perceptual biases produced by interaction of various senses (Bernstein & Edelstein, 1971; Marks, 1974; Marks, Hammeal, Bornstein, & Smith, 1987; Melara & O'Brien, 1987). Many studies have found crossmodal facilitation and interference effects in individuals making a behavioral response. Individuals are faster and more accurate in classifying a stimulus in one modality when it is coupled with a crossmodally-congruent (and task-irrelevant) stimulus in another modality compared to when it is coupled with an incongruent stimulus. For example, Melara and O'Brien (1987) employed a speeded-classification task to study the correspondence of auditory pitch and visuospatial elevation. They found a congruency effect; participants were faster to classify stimuli in the target modality (elevation of visual stimuli or auditory pitch) when the stimulus in the task-irrelevant modality was congruent with the task-relevant target stimulus. Gallace and Spence (2006) documented strong associations between auditory pitch and visual object size, with participants in their speeded-classification task facilitated in making responses when multisensory stimuli were congruently paired (e.g. high pitch sound paired with small object) compared to when pairings were incongruent. Other studies have demonstrated an association between waveform (square wave- or sinusoidal wave tones) and object angularity (Parise & Spence, 2012). In an implicit association task, Parise and Spence (2012) demonstrated a congruency effect for pairings of pitch and size as well as pairings of pseudoword and object shape.

Crossmodal correspondences have also been found to produce perceptual biases and illusory effects. In two studies, Parise and Spence (2008, 2009) tested the effect of crossmodal congruency on auditory capture, a phenomenon in which a visual stimulus presented immediately before or after an auditory stimulus will appear to be simultaneous with or 'captured by' the sound stimulus. This auditory capture effect can modulate perceptual sensitivity, either exaggerating or reducing the apparent lag between two stimuli presented in rapid succession, depending on crossmodal congruency.

Although there is a great deal of variation across test groups from different languages and cultures, some crossmodal mappings appear to be consistent across different populations (Bremner et al., 2013; Parkinson, Kohler, Sievers, & Wheatley, 2012). For example, Köhler (1929) found reliable mappings between certain pseudowords and object shapes, with individuals matching pointed shapes to pseudowords like *takete* and *kiki* and matching more rounded shapes to pseudowords like *maluma* and *bouba*. This finding has since been widely replicated across diverse subject populations, suggesting that certain language sounds are reliably associated with object shape irrespective of an individual's language background- and thus indicating that this mapping is likely rooted in cognitive mechanisms shared across cultures, rather than being idiosyncratic to any particular language (Bremner et al., 2013). Studying the distribution and diversity of these mappings provides clues as to underlying mechanisms.

Interestingly, individuals often share intuitions about crossmodal mappings that are conventionalized in other cultures but not their own. Linguistic research has shown that whereas many languages describe auditory pitch using the same terms as spatial verticality (e.g. 'high' and 'low'), other languages apply different sensory metaphors to describe pitch. For instance, in Farsi, high-pitched sounds are described using the term for 'thin', and low-pitch sounds using the term for 'thick'. Eitan and Timmers (2010) surveyed a cohort of Hebrew speakers, and found that they reliably matched high- and low-pitched sounds to concepts of 'thin' and 'thick', respectively, even though this was not a culturally-entrained or lexicalized metaphor in their culture. Similarly, Parkinson, Kohler, Sievers & Wheatley, (2012) reported that an isolated tribe in Cambodia demonstrated a preference for matching a rising pitch with a shape moving up and a falling pitch with a shape moving down, despite their lack of shared terms (e.g. 'high' and 'low') for these attributes (Parkinson et al., 2012). This has led researchers to ask whether mappings shared across language and cultures could be based in universals of perceptual or cognitive experience, or whether we could have an innate predisposition to associate certain domains. Do we come to associate these attributes because thin objects in the environment emit higher-pitched sounds and have higher resonating frequencies than thick objects, or are we prewired or otherwise predisposed to associate certain domains?

Research in infants has demonstrated that many crossmodal mappings are present early in life indicating that at least for some domains, humans may be predisposed to make certain crossmodal associations (Dolscheid, Hunnius, Casasanto & Majid, 2014; Mondloch & Maurer, 2004; Spector & Maurer, 2008, 2009; Walker et al., 2010). For example, Dolscheid et al. (2014) found that prelinguistic infants reliably associated high-pitched tones with both high visuospatial elevation and thin visual forms, and correspondingly expected lower-pitched tones to be matched with low elevation and thicker forms.

This finding reinforces the possibility that these patterns are in place even prior to being entrained or shaped by metaphorical language.

Ubiquitous as crossmodal mappings are, the cognitive and neural mechanisms underlying such associations are not well understood. We currently lack a productive model for how crossmodal mappings are instantiated in our neurocognitive systems. How are the various incoming sensory signals integrated into a coherent cognitive experience, and why are certain sensory domains aligned as congruent or incongruent? What is it that makes certain sensory attributes 'go together'? A diverse array of possibilities has been advanced for the potential neural basis of crossmodal mappings, and different mechanisms likely underlie various instances of crossmodal associations.

Possible neural bases for cross-sensory mappings

The development of functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and transcranial magnetic stimulation (TMS) in recent decades has vastly expanded our understanding of sensory encoding and interactions in the brain. Across neurophysiological and behavioral studies, a popular paradigm for examining multisensory processing is to test the effect of presenting multisensory stimuli that are congruently- or incongruently coupled. Amedi et al. (2005) and others have observed that multisensory regions are not equally sensitive to all combinations of incoming stimuli in different modalities, but typically exhibit selectivity (or in some cases suppression effects) for particular crossmodal combinations of stimuli (Calvert, Campbell, & Brammer, 2000; Kadunce, Vaughan, Wallace, Benedek, & Stein, 1997). This shows that even at precortical stages of processing, certain multisensory combinations will be privileged, and allowed through for further integrative processing, while others will be gated or otherwise attenuated.

Identifying systems that are sensitive to a particular type of multisensory congruency is an important step in understanding how different crossmodal associations are represented in the brain. Recent research using similar paradigms has produced a rich set of fMRI studies examining multisensory integration and representation of meaningful perceptual events (e.g. a person speaking, a dog barking, or a hammer striking a nail). There is, however, a paucity of research on the neural underpinnings of more basic crossmodal correspondences such as the association of pitch and visuospatial elevation. It is clear that in some cases, sensory dimensions come to be associated over the course of experience; mapping between temperature and color, for instance, may not appear until late childhood (Marks, Ben-Artzi, & Lakatos, 2003). In other cases, mappings may emerge out of neural architecture; with certain aspects of perceptual experience associated either due to a an innate predisposition for particular neural wiring or connectivity, or by virtue of similarities in the neural encoding of stimuli in different sensory modalities leading to encoding in a common neural format. Over the past few decades, investigators have developed several major models for how information from the different senses is associated and integrated across cortical and other neural systems. Below, we outline several putative neurocognitive systems that have been hypothesized to underlie crossmodal mappings.

Intersensory connections or multisensory convergence. Certain crossmodal mappings may arise as a result of direct neural connections between sensory areas (intersensory connections) or systems farther along the processing stream where multiple sensory channels converge into multisensory regions (multisensory convergence). Converging evidence from several studies suggests that information from the different senses combines and influences processing in many regions previously considered to be unimodal (e.g. information from non-visual modalities may be encoded in visual cortex). This crossmodal influence could result from either direct inputs from the other senses ('sensory cross-activation'; Raij et al., 2010), or via feedback from multisensory systems or higher-order cortical regions (Deshpande et al., 2010, Lacey et al., 2010; Lacey et al. 2014). Regardless of the circuitry, it is evident that activity in sensory areas previously considered modality-specific can be modulated by incoming signals from the other senses (Sathian et al., 1997; Calvert, 2001; Calvert et al., 2000; Meyer et al., 2010; Meyer, Kaplan, Essex, Damasio, & Damasio, 2011; Raij et al., 2010; van Atteveldt, Formisano, Goebel, & Blomert, 2004) and that disruption of these areas can affect processing of other senses (Zangaladze, Epstein, Grafton, & Sathian, 1999). Using a combination of fMRI and MEG methods, Raij et al. (2010) found evidence for crossmodal activation at a very early stage in processing. Primary visual cortex responded to auditory stimuli just 10 milliseconds later than visual stimuli, and primary auditory cortex responded to visual stimuli only 52 milliseconds after auditory stimuli. The relatively short latencies between responses evoked by inputs in a canonical modality compared to input from a secondary modality was taken as support for direct cross-activation of early sensory areas (Raij et al., 2010). Zangaladze et al. (1999) found that transcranial magnetic stimulation (TMS) applied to visual cortex (likely V6) could disrupt discrimination of tactile stimuli known to activate that particular visual cortical region, indicating functional involvement of visual areas in processing tactile information. These studies provide compelling evidence that signals from the different senses interact as early in processing as the early sensory cortices, leaving open the possibility that crossmodal congruency could be encoded at these early stages of processing.

Recent advances in neuroimaging analysis provide additional tools for examining the processing and interaction of sensory information from different sensory modalities. Multi-voxel pattern analysis (MVPA) compares the blood-oxygen-level dependent (BOLD) activation or deactivation exhibited by each voxel in response to different classes of stimuli, and allows researchers to assess similarity of the patterns of neural activation in response to diverse stimuli and across many different systems. A common approach is to train a classifier algorithm on voxelwise activity for different categories of stimuli and then test whether the classifier can distinguish the same categories in a different dataset (Kamitani & Tong, 2005; Kriegeskorte et al., 2008; Meyer et al., 2010, 2011; Meyer & Kaplan, 2011; Norman, Polyn, Detre, & Haxby, 2006). Multi-voxel pattern analysis has been successfully used to compare activations in and between different sensory cortices. For example, Meyer and colleagues (2010) demonstrated that activity in auditory cortex could be reliably used to determine whether a participant had seen a silent video that showed an animal, instrument, or object. This showed that auditory cortex is involved in processing visual stimuli that in some way index or imply a sound. In another study, Meyer et al. (2011) trained a classifier algorithm to read out activity in somatosensory cortex and distinguish which of several objects were presented in a silent video implying touch/haptic interaction with an object. Thus, simply *seeing* a haptic interaction with an object produced activity in somatosensory regions that would be involved in directly experiencing a touch. These findings are particularly important because this research demonstrates that experience in one sensory modality can imply or index experience in other modalities, and accordingly, evoke activity in the corresponding sensory cortices. Together, these studies provide evidence for crossmodal influences in primary sensory cortices, and support the possibility that crossmodal mappings and representations of multisensory congruency could be based in intersensory connectivity.

In an effort to better understand crossmodal mappings found in typical individuals, some researchers look to extreme forms of crossmodal mappings such as synesthesia. Synesthesia is a condition in which stimulation in one sensory modality produces a concurrent sensation either in another modality (e.g. hearing musical notes producing visual experience of colors or hearing speech sounds producing flavors) or in an unstimulated domain within the same modality (e.g. reading a written letter producing a sensation of color). While the cross-domain mappings experienced by synesthetes are largely idiosyncratic, there do appear to be patterns in how sensory domains tend to be aligned (sounds that are higher pitch tend to be mapped to relatively lighter colors), and there is some evidence to suggest that crossmodal mappings of synesthetes match those of the general population (Bankieris & Simner, 2014; Sagiv & Ward, 2006). A recent study found that developmental synesthetes exhibited stronger crossmodal congruency effects for the sound-symbolic association between pseudoword and shape compared to non-synesthetes, but found no difference in synesthetes' processing of lower-level sensory correspondences between auditory pitch and size or auditory pitch and spatial elevation (Lacey, Martinez, McCormick, & Sathian, 2016; see also Spector & Maurer, 2009). This finding suggests that synesthesia may represent an extreme form of intersensory mapping processes experienced by the general population to a lesser extent, and that it likely a high-level, post-perceptual phenomenon (Cohen Kadosh & Henik, 2007; Cohen Kadosh, Henik & Walsh, 2007; Sagiv & Ward, 2006). Similar to the cross-sensory activation found in the neuroimaging studies described above, there is evidence of cross-sensory activity in developmental synesthetes, with specific sounds evoking neural activity in primary gustatory (Ward & Simner, 2003) or early visual areas (Aleman & Rutten, 2001; Spector & Maurer, 2009), indicating a direct functional connectivity between these systems. Correspondingly, Rouw and Scholte (2007) found grapheme-color synesthetes showed greater structural connectivity between cortical regions representing graphemes and color compared to neurotypical individuals (Rouw & Scholte, 2007).

A number of studies have found that infants (P. Walker et al., 2010, 2014 but see Lewkowicz & Minar, 2014) and toddlers (Maurer, Pathman, & Mondloch, 2006; Mondloch & Maurer, 2004) exhibit some of the same crossmodal mappings found in adults, suggesting that some mappings may be based in either innate predispositions or very early experience. For example, Walker et al. (2010) tested infants using a preferential looking paradigm, and found that at age 3-4 months, infants already preferred the coupling of an auditory tone rising in pitch with a visual object rising in visuospatial elevation and falling pitch coupled with visual object moving lower in spatial elevation in contrast to the reverse couplings. In an extension of the classic kiki-bouba study, Maurer et al. (2006) found that preschoolers showed similar preferences to adults, selecting labels such as kiki and takete for pointed visual shapes and labels such as bouba or maluma for more blob-like shapes. These findings have led some to propose that synesthesia is normal early in development, and while many connections are pruned over the course of typical development into adulthood, some intersensory connections remain latent in the brains of non-synesthetes (Maurer & Mondloch, 2005; Maurer, Pathman, & Mondloch, 2006b; Spector & Maurer, 2009).

Multisensory integrative systems. As we interact with the environment, incoming sensory information is often correlated across the different senses (Bahrick, Lickliter, & Flom, 2004; Calvert, 2001; Macaluso & Driver, 2005). Perceptual events that derive from a common source in the external world are likely to share characteristics such as spatial location, temporal properties, and relative intensity (Calvert, Spence, & Stein, 2004; Vroomen & Keetels, 2010; Zampini, Guest, Shore, & Spence, 2005). Correspondences of attributes such as these serve as fundamental cues in integrative systems for sensory binding and object and spatial representation (Deroy, Chen, & Spence, 2014; Ernst & Bülthoff, 2004; Parise & Spence, 2009; Spence, 2013; Vroomen & Keetels, 2010; Wallace et al., 2004) and it may be these integrative systems that underlie cross-sensory mappings. The capacity to combine channels of sensory information confers behavioral benefits (e.g. lowering detection thresholds, enhancing discrimination, speeding reactions), and the ability to use information from the different senses flexibly and interchangeably facilitates object recognition, especially in noisy or perceptually impoverished conditions (Calvert, 2001; Mossbridge, Zweig, Grabowecky, & Suzuki, 2017). Perhaps some crossmodal correspondences reflect encoding of intersensory redundancy across multisensory systems. Crossmodally consistent or *congruent* multisensory stimuli can modulate multisensory integration. For example, pairing audiovisual stimuli congruently (e.g. a small visual shape with a high pitched sound) can increase 'perceptual unity' effects, making individuals more likely to perceive temporally or spatially discrepant signals as belonging together as part of a unified percept, being simultaneous, or originating from the same location in space (Brunel, Carvalho, & Goldstone, 2015; Parise & Spence, 2008; Vatakis & Spence, 2007). In terms of temporal binding, congruently pairing audiovisual stimuli has been shown to produce 'temporal ventriloquism' effects, which can have the effect of either increasing or decreasing sensitivity to differences in stimulus onsets. Coupling stimuli congruently can increase the robustness of apparent synchronicity perception (the impression of synchrony even when there is a gap between onsets of the two stimuli), thereby expanding the temporal binding window. Compared to incongruent stimuli, crossmodally-congruent auditory and visual stimuli can be presented farther apart in time and still be integrated and experienced as a unified percept (Brunel et al., 2015; Parise & Spence, 2008, 2009).

Associations of complex environmental stimuli in different modalities are likely established through our multisensory experience of objects and events in the world (i.e., we often see a dog as we hear a dog's bark so this becomes a learned audiovisual coupling). However, the cognitive basis for mappings between more basic perceptual dimensions (e.g., pitch and visuospatial elevation or pitch and object size) is less apparent. A compelling possibility is that, much like associating a dog and its bark, crossmodal mappings between more basic perceptual dimensions may arise out of correlated perceptual experience of sensory signals in the physical environment (Shams & Seitz, 2008; Spence, 2011), and are based in multisensory integrative systems (Parise & Spence, 2009). For instance, small objects are more likely to make relatively high pitch sounds, and resonate at higher frequencies than large objects, which produce and resonate with relatively lower pitch sounds, phenomena which could explain why we associate higher pitch sounds with smaller objects, and lower pitch sounds with larger objects. If intersensory mappings arise as a result of sensitivity to crossmodal regularities in the environment, perceivers could become more likely to associate dimensional attributes that are coupled in experience.

There are a number of ways such mappings could be instantiated in the brain. An intersensory-connection account of cross-sensory associations would predict co-activation between early sensory cortices of the associated modalities. A percept in one modality could either activate a corresponding sensory region directly, or could activate multisensory representations in association areas, with feedback to produce activity in sensory regions representing the corresponding modality. For example, hearing high and low pitch tones could produce different patterns of activity, not only in auditory, but also in visual areas, either by means of direct connections between the two sensory cortices or via feedback from integrative systems. Von Kriegstein and Giraud (2006) trained individuals on novel cross-sensory correspondences and found that over the course of training, functional connectivity increased between the sensory and association regions that represent the newly associated domains. Other research has shown that even culturally-entrained mapping of written letters to sounds (reading) comes to modulate cross-sensory areas (van Atteveldt, Formisano, Goebel, & Blomert., 2004). We could thus expect cross-sensory associations to be reflected in early sensory areas, (typically considered sensory-specific) regardless of whether they are a product of innate wiring or acquired through experience.

Another possibility is that crossmodal mappings could be based in multisensory integrative systems. Research on the sensory-integrative basis of crossmodal correspondences has focused on three major types of correspondence: temporal co-occurrence, spatial location, and object identity. This research has implicated a number of areas involved in multisensory representations– including association areas where different streams of sensory information converge in the occipito-temporal, temporo-parietal, and occipito-parietal regions (Beauchamp, 2005a; Driver & Noesselt, 2008).

The superior temporal cortex is implicated in multisensory processing in several neuroimaging studies (for reviews, see Calvert, 2001; Amedi et al., 2005, also; Beauchamp, Argall, Bodurka, Duyn, &

Martin, 2004; Calvert et al., 2000; Kassuba, Menz, Röder, & Siebner, 2012; van Atteveldt et al., 2004), and has been shown to respond to audiovisual synchrony and object identity (Kassuba et al., 2012) in particular (van Atteveldt, Formisano, Blomert, & Goebel, 2006; Beauchamp, 2005a, 2005b). An early study examining audiovisual integration of speech found that the superior temporal sulcus (STS) showed a stronger response for *synchronous* presentation of auditory and visual recordings in contrast to asynchronous presentation of the two, indicating that the STS may have a role in binding or otherwise matching crossmodal stimuli that are temporally coupled (Calvert et al., 2000). Using high-resolution fMRI to compare BOLD activity profiles to various stimuli revealed heterogeneous receptive patches in human superior temporal sulcus (STS), with some patches responding selectively for either meaningful environmental auditory or visual input, and other patches responding preferentially to the conjunction of the two into a multisensory (auditory + visual) input (Beauchamp, Argall, et al., 2004). This patterning in activity corroborates findings based on single cell recordings in macaques, which have demonstrated mixed (audio, visual, and audio+visual) response properties of neurons in the STS (Dahl, Logothetis, & Kayser, 2009; Hikosaka, Iwai, Saito, & Tanaka, 1988). Another study found that bilateral segments of Heschl's sulcus (and in some individuals, planum temporale) are selective for temporally-congruent presentation of written letters and audio recordings of associated speech sounds (congruent letter + speech sound > incongruent letter + speech sound; van Atteveldt et al., 2004). Multisensory stimuli can also be crossmodally matched in terms of their spatial location or source in the environment. Sestieri et al (2006) employed a crossmodal matching task to examine the neural basis of spatial congruency processing. Participants in the task were presented with images and sounds of familiar objects and animals and responded as to whether the image and sound were presented on the same side (in another task, they responded as to whether image and sound were matched in identity). Relative to the object recognition task, the spatial location task modulated activity mainly in parietal regions including the intraparietal sulcus (IPS) and inferior parietal lobule (IPL) when audiovisual spatial congruency was manipulated.

Another way in which stimuli in different modalities can be matched is in terms of identity (Erickson, Heeg, Rauschecker, & Turkeltaub, 2014; Sestieri et al., 2006). One region widely shown to have multisensory response properties for object identity is the middle temporal gyrus (MTG). Beauchamp, Lee, Argall, and Martin (2004) conducted a series of fMRI experiments on multisensory integration of familiar environmental stimuli. They presented subjects with unimodal and multisensory stimuli showing tool use or animals, and contrasted neural responses to these meaningful familiar stimuli with responses for unfamiliar scrambled images and sounds. Unimodal stimuli consisted of static images or silent video or audio recordings (e.g., the sound or image of a hammer striking a nail, or a dog barking), and multisensory stimuli combined the images and audio recordings. They identified a portion of posterior MTG that responded to both meaningful auditory and visual stimuli, but showed greatest modulation of activity for the combined (multisensory) audio-visual stimuli. This indicates that the MTG has a role in encoding multisensory information related to object identity and could support crossmodal alignment of meaningful stimuli in different modalities.

Systems that encode multisensory information about object identity sometimes exhibit 'modality invariance', a response property wherein a neural system responds to the identity of the stimulus, regardless of the modality of sensory input (Lambon Ralph, Sage, Jones, & Mayberry, 2010; Man, Damasio, Meyer, & Kaplan, 2015). A modality invariant system exhibits the same response for corresponding stimuli in different modalities, essentially treating them as interchangeable or part of a unified representation, and with stimulation in either modality sufficient to trigger a response. For example, the lateral occipital cortex (LO) has been found to respond to both object shape in both visual and haptic modalities, and in some cases, even auditory modality (Amedi, Jacobson, Hendler, Malach, & Zohary, 2002; Beauchamp, 2005a; James, Stevenson, Kim, VanDerKlok, James, 2011). Response properties such as modality invariance as well as spatial and temporal congruency effects could reflect intersensory redundancy in encoding for object representation. A neural system that equates percepts deriving from a common source could do so on the basis of statistics of perceptual experience (Evans & Treisman, 2010; Kassuba et al., 2012). Perhaps cross-sensory mappings are supported by multisensory regions, and represent a sort of unified, modality-invariant processing of perceptual signals that are likely to derive from a common source in the environment (Evans & Treisman, 2010; L. Walker, Walker, & Francis, 2012).

By a grounded account of cognition, conceptual representations are based in sensory-motor systems (Barsalou, 1999, 2008; Goldstone & Barsalou, 1998; Lakoff & Johnson, 1980; Pulvermüller, 2013), and conceptual knowledge is brought online by re-activating or simulating the experience across the same sensory systems initially used to perceive and encode it. Thus, a percept in one sensory modality could serve to simulate or bring online a multisensory representation by co-activating semantically related sensory representations in the other modalities (Barsalou, Simmons, Barbey, & Wilson, 2003; Bergen, 2007; Meyer et al., 2010, 2011).

Semantically related perceptual experiences in different sensory modalities could be associated because they co-activate one another and form a common sensory-based conceptual representation. For example, Meyer and colleagues find compelling evidence for sensory cross-activation. In separate studies, they used multivariate pattern analysis (MVPA) and found that voxelwise BOLD activity in early auditory cortices (Meyer et al., 2010) and primary somatosensory cortex (Meyer et al., 2011) could be used to determine which silent video (implying a touch or a sound, respectively) had been played. These findings clearly demonstrated that crossmodal representations can be triggered by semantic stimuli. The finding that a percept in one modality can bring online multisensory representations sufficiently rich to allow researchers to 'read out' what the original stimulus was could be interpreted as evidence in support of sensory grounding of conceptual representations. Recent studies have shown that even simply *imagining* perceptual experiences (e.g., a particular color) can produce activity in sensory regions, so the idea that a meaningful unimodal stimulus could trigger a multisensory representation is not far-fetched.

In some ways, neural evidence for the *grounded semantic* account of cross-sensory mappings could be difficult to distinguish from the *multisensory integrative systems* account described in the previous section, as the two accounts make many of the same predictions in terms of neural regions involved. One point on which the *grounded semantic* and *multisensory integrative systems* accounts could differ is that the grounded account necessarily involves conceptual representation, whereas the multisensory account could involve nothing more than statistically correlated experience of low-level perceptual events (e.g., temporal co-occurrence of events in different modalities), without a unified conceptual source.

In addition to the multisensory congruency effects discussed above, some regions exhibit effects of crossmodal attention and priming (Adam & Noppeney, 2010; Kassuba et al., 2012). Adam and Noppeney (2010) found that auditory stimuli could crossmodally prime object category information, facilitating a performance on a task where subjects categorized visual stimuli as faces, landmarks, or animals. An effective connectivity analysis of the imaging data revealed that, relative to incongruent pairings, the crossmodally congruent auditory primes sharpened category-selective responses to visual stimuli in the ventral occipito-temporal cortex. Kassuba et al. (2012) found evidence for crossmodal matching effects for stimuli that encode information about object shape. They found that an auditory prime modulated subsequent processing of haptic targets in the lateral occipital cortex (LO). The LO has been found to respond to both visual and tactile stimuli (Amedi, Jacobson, Hendler, Malach, & Zohary, 2002; Beauchamp, 2005a), and is thought to encode sensory and conceptual aspects of object information (Kassuba et al., 2012; Kriegeskorte et al., 2008). In contrast to the response for *haptic* targets in LO, Kassuba et al. found that the pSTS showed crossmodal matching effects when *auditory* targets were matched to a haptic prime. The experiments discussed above implicate several regions in which object-based sensory knowledge can have a crossmodal influence on stimulus processing. Finding recruitment of these regions for any of the cross-sensory mappings in the dissertation experiments could be preliminary evidence for involvement of multisensory representations in these mappings.

Magnitude system. Some crossmodal correspondences may arise due to a common neural format for encoding dimensional information related to stimulus magnitude (Bueti & Walsh, 2009; Spence, 2011; Walsh, 2003; Ward, 2013). Theories of magnitude processing posit a domain-general system, which represents information about approximate number, quantity, extent, and intensity of perceptual experiences in a common neural format (Dehaene, Piazza, Pinel, & Cohen, 2003; Lourenco & Longo, 2010; Walsh, 2003). By a magnitude account of cross-sensory mappings, perceptual dimensions in various modalities could be input into a common magnitude system where they could then be aligned and associated by virtue of their shared more-less polarity or by matching the relative intensity of the defining attributes (Walsh, 2003). For example, dimensions of object size and auditory loudness could be aligned by virtue of 'large' and 'loud' both being encoded as being 'more' stimulating in their respective systems. By such an account, crossmodal correspondences arise as a by-product of the cognitive architecture of systems that support supramodal representation of information about stimulus magnitude.

A number of parietal regions are believed to be involved in magnitude representations and may facilitate cross-sensory comparison and association. Recent neuroimaging and TMS studies examining magnitude-related processing have consistently implicated specific regions including the intraparietal sulcus (IPS) (Cantlon, Brannon, Carter, & Pelphrey, 2006; Cohen Kadosh et al., 2005; Dehaene et al., 2003; Eger, Sterzer, Russ, Giraud, & Kleinschmidt, 2003; Holloway & Ansari, 2010; Hubbard, Piazza, Pinel, & Dehaene, 2005; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Sathian, Simon, Peterson, Patel, Hoffman, & Grafton, 1999), superior parietal regions (Dehaene et al., 2003; Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999; Holloway, Price, & Ansari, 2010; Kaufmann et al., 2008), and the angular gyrus (Cattaneo, Silvanto, Pascual-Leone, & Battelli, 2009; Dehaene et al., 2003, 1999; Göbel, Walsh, & Rushworth, 2001). There is general consensus among researchers that the IPS is a hub for magnitude processing. Segments of the IPS have been found to respond to approximate numerical magnitude of stimuli in a format-independent manner, suggesting that diverse inputs may be mapped onto a common abstract or supramodal representation of magnitude (Dehaene et al., 2003, 1999; Holloway & Ansari, 2010; Piazza et al., 2004). While there is general agreement that the IPS is important for magnitude processing, proposed models differ in the extent to which the system is lateralized, and in the nature of the contributions of the two hemispheres. In his foundational Theory of Magnitude (ATOM), and follow-up studies, Walsh (2003) emphasizes a predominantly right-lateralized inferior parietal system with a focused region of overlap between representations of space, time and number (Bueti & Walsh, 2009). Dehaene and colleagues (2003,1999) describe a bilateral system for magnitude representation, and posit different contributions of the two hemispheres. They found that the horizontal segment of the IPS (hIPS) was bilaterally responsive to numerical magnitude as represented by both symbolic (written Arabic numerals) and non-symbolic (visual arrays of basic shapes) stimuli, and activity in these regions was modulated more by tasks requiring estimation of approximate number than calculation of precise number. A number of magnitude studies have used BOLD adaptation paradigms- habituating participants to a series of similar stimuli and measuring the rebound effect (increase in BOLD activation) when a deviant stimulus is presented. Studies by Piazza et al. (2004) and Cantlon et al. (2006) identified bilateral sites in the IPS that were more responsive (showed a greater rebound from adaptation) to deviations in number of items in a visual array, as compared to changes in other visuospatial attributes such as changes in object shape or array density. Cohen Kadosh et al. (2005) found widespread bilateral involvement of the IPS in tasks comparing stimulus brightness, stimulus size, and symbolic number magnitude, with a portion of the left IPS responding specifically to symbolic number magnitude. In another study, Eger et al. (2003) presented participants with unimodal symbolic stimuli including written numerals ‘5’ and auditory recordings of number words ‘five’, as well as control conditions including colors, spoken color words, letters, and spoken letter words. They tested for regions that showed a supramodal response specifically for the number stimuli and found extensive bilateral involvement of the IPS in processing both visual and auditory number stimuli. These portions of the IPS responded to symbolic representations of number in both auditory and visual modalities. Together these studies indicate a role for the IPS that is format-general (responding to a variety of stimulus formats and modalities) yet domain-specific (responding to stimulus magnitude but not other visuospatial features). The finding that this system encodes incoming stimulus information from different modalities into a common supramodal format supports its potential role in cross-sensory mappings and multisensory congruency.

In addition to the IPS, several studies have identified a role for superior parietal areas, including posterior superior parietal lobule (pSPL) in the right hemisphere (Cantlon et al., 2006; Holloway et al., 2010; Kaufmann et al., 2008; Riemer, Diersch, Bublatzky, & Wolbers, 2016), or bilaterally in pSPL

(Dehaene et al., 1999), in representing magnitude. In contrast to the format-general representations reported for the IPS, a majority of the magnitude-related findings in the pSPL indicate that it responds primarily to non-symbolic (compared to symbolic) representations of approximate magnitude. For example, Dehaene et al. (1999) found that the bilateral pSPL responds more strongly for a magnitude approximation task compared to a similar task requiring precise computations. Holloway et al. (2010) reported that right posterior superior parietal region responded to non-symbolic stimulus magnitude (number of items in a visual array), but not to symbolic representations (e.g., written numerals). In line with these findings, Santens, Roggeman, Fias & Verguts (2010) conducted a functional connectivity analysis and found evidence that the posterior superior parietal cortex is part of a pathway involved in processing non-symbolic (and not symbolic) number stimuli, and which has outputs to the IPS. In addition to non-symbolic representations of magnitude, these superior parietal regions are involved in multisensory integrative processing of visuospatial information (Mulvenna & Walsh, 2006), indicating a possible link between the visuospatial system and representations of approximate magnitude. Several teams have advanced the theory that the magnitude system may have its origins in a neural system originally evolved for integrating various channels of sensory information into a unified spatial representation for action, and that this system was then co-opted and repurposed (exapted) for novel, more generalized functions (Bueti & Walsh, 2009; Dehaene & Cohen, 2007; Hubbard et al., 2005; Simon, Mangin, Cohen, Le Bihan, & Dehaene, 2002). Supporting this model, Kaufmann et al. (2008) identified a region of the right PSPL that responded to both a non-symbolic number task and a spatial task relative to a rest condition, as well as a more anterior region of the left SPL that responded for a non-symbolic number task (but not a spatial task using identical stimuli). Additional evidence that magnitude representations may be based in spatial systems comes from cross-cultural research. Researchers have observed a cross-cultural tendency for using space to represent amount/number (e.g., a number line) as well as auditory pitch and intensity (Dehaene & Cohen, 2007; Parkinson et al., 2012). The fact that we see relatively consistent patterns in the way people spatialize magnitude information across cultures suggests that the neurocognitive systems for representing space and magnitude may be functionally linked. Thus, the region(s) involved in magnitude processing could be a basis for cross-sensory mappings, and mappings involving spatial domains (e.g., the pitch-elevation mapping), in particular.

Another region consistently implicated in studies linking spatial cognition and representation of number and magnitude is the angular gyrus (AG). Situated at the confluence of visual, auditory, and somatosensory streams from the occipital, temporal, and parietal lobes, the AG is a multisensory integrative hub supporting capabilities such as crossmodal comparison and attentional reorienting that could underlie the processing of magnitude and the ability to mentally manipulate number (Seghier, 2012). Although there are mixed results and the AG may not exhibit all of the response properties one would expect of a primary hub in general representation of magnitude, some have proposed that the left AG may specifically underlie spatial aspects of number representation, such as use of a mental number line (Cattaneo et al., 2009; Göbel et al., 2001), which many consider to be a core component of the magnitude system. For example, Cattaneo et al. (2008) found that TMS to both left and right AG disrupted the spatial cueing effects elicited by presenting relatively large or small magnitude numbers.

Whereas the IPS and superior parietal cortex have generally been found to be involved in representing *approximate* number, there is some evidence to suggest that the AG is involved in more precise representations of number (Dehaene et al., 1999). Dehaene et al (1999, 2003) have suggested that rather than being part of the more general representation of magnitude, the left AG is recruited for linguistic aspects of numerical representations and verbally based arithmetic. Others have found the left AG to be more responsive to symbolic number than to non-symbolic arrays (Holloway et al., 2010). Together, these findings suggest the AG may have a specific role in precise, symbolic number representations, verbal manipulation of these representations, and the spatial representation of the mental number line. Finding involvement of the AG, IPS, and SPL in crossmodal mappings could indicate a basis in a form of magnitude processing.

Semantic systems. In this broad family of theories, perceptual experiences in different modalities are linked by a common conceptual meaning (Hein et al., 2007; Martino & Marks, 1999; Spence, 2011; L.

Walker et al., 2012). For example, an image of a lion, the sound of a lion's roar, and the word 'lion' all refer to the same concept, and could be associated through this shared conceptual representation. It may be that some more basic crossmodal associations are also conceptually based. Such a mapping could be neurally instantiated in a number of ways.

Another possibility is that information from different senses is associated by means of a more abstract conceptual representation. In such a system, different channels of sensory information would feed into a supramodal system, mapping onto a common meaning as represented in an abstract format. In a system responding to abstract semantic properties, we could expect to find responses modulated by the identity or meaning of crossmodal stimuli (this in contrast to low-level multisensory binding, which responds to temporal co-occurrence). Research on the semantic basis of cross-sensory associations often examines the effect of presenting multisensory stimuli that are semantically coupled – congruently- or incongruently (an image of a cat coupled with a cat's meow or a dog's bark). Identifying systems that respond differently for congruent and incongruent semantic stimuli is an important step in understanding the systems underlying semantically mediated associations. Hein et al. (2007) used novel shapes and sounds as well as animal pictures and animal sounds to examine the role of semantic congruency and familiarity of AV pairings. They found the inferior frontal cortex (IFC) was sensitive to semantic congruency, but not familiarity- responding for both novel/unfamiliar AV stimulus pairings (novel shape + novel sound) as well as familiar but incongruent pairings (image of a dog + a cat's meow). The posterior STS responded for familiar stimuli regardless of whether they were congruently- or incongruently paired. In contrast, the superior temporal gyrus (STG) responded only for the familiar AV congruent condition. This study elegantly demonstrates dissociations between regional involvement in stimulus *familiarity* and *multisensory congruency* in semantic stimuli, and indicates a specific role for the STG in representing multisensory congruency of semantic stimuli.

A related line of neuroimaging research identifies neural systems that exhibit a modality-invariant response for semantically matched stimuli in different modalities; in a sense these systems treat stimuli in different modalities as interchangeable, suggesting that the channels map onto a common representation.

In some cases these ‘format-general’ representations may even be brought online by linguistic stimuli including written letters or words. For example, van Atteveldt et al. (2004) found that unimodally presented written letters and auditory recordings of letters being named produced a similar BOLD response profile bilaterally in the STS and STG and that a portion of Heschl’s sulcus responded most strongly for congruent couplings of audiovisual stimuli. The finding that these systems exhibit ‘modality-invariant’ responses for semantic stimuli as well as selectivity for semantically congruent stimulus couplings suggests they may be components of a semantic system. Lambon Ralph, Sage, Jones, & Mayberry (2010) take a similar approach to identify supramodal conceptual/semantic hubs. Similar to the principle of modality invariance, they seek systems that respond similarly for conceptually matched stimuli. Across several studies, they identify a portion of the anterior temporal lobe that exhibits a concept-specific response, regardless of how perceptually similar the surface features of visual objects are in the stimuli. For example, in one study, they found similar patterns of activity for conceptually similar stimuli (e.g., teapots with a range of different perceptual surface characteristics) (Lambon Ralph et al., 2010). They interpret these response patterns as evidence for encoding of relatively abstract conceptual representations, as they may be brought online by any number of diverse sensory stimuli. These researchers also examined the various means of activating these semantic representations. Functional neuroimaging studies have shown that both the anterior temporal lobe (ATL) and the MTG respond to a variety of semantic stimuli including visual, auditory, and verbal stimuli (Visser & Lambon Ralph, 2011; Visser et al., 2012). Lambon Ralph et al. (2010) suggest the ATL is a hub for semantic-conceptual representations (Lambon Ralph, 2014; Lambon Ralph et al., 2010). A tractography study by Binney, Parker, & Lambon Ralph (2012) showed that the ATL receives inputs from visual and auditory regions and the team posited a role for the region in encoding modality- and context-invariant semantic representations. In addition, several studies have now implicated the inferior temporal cortex (IT) in encoding category-level information about visual objects, activity which some interpret as an early stage of abstracting sensory information into more abstract representations of semantic categories (Kriegeskorte et al., 2008; Leopold, Bondar, & Giese, 2006). Although we have a limited understanding of meaningful representations in the ATL, MTG, and IT, it is clear that these regions have a role in encoding relatively abstract supramodal semantic information.

Smith and Sera (1992) suggested that learning certain concepts (e.g., *less*/*more*) leads children to dimensionalize perceptual experience, draw parallels between dimensional attributes, and align sensory dimensions in a consistent manner. If this is the case, language could play an important role in establishing and reinforcing our associative mappings. Proponents of the Semantic Coding Hypothesis suggest that certain crossmodal mappings are linguistically mediated (Martino & Marks, 1999; Melara & Marks, 1990b). Using a common label to describe multiple perceptual domains could lead us to associate the domains. For example, it could be through the use of terms such as 'low' and ‘high’ that we come to associate perceptual dimensions of pitch, visuospatial elevation, and loudness (Martino & Marks, 1999; Melara & Marks, 1990b). By this account, cross-sensory mappings are mediated by post-perceptual linguistic processes in an abstract format and crossmodal interactions evidenced by behavioral tasks arise at this abstract level of encoding (Ben-Artzi & Marks, 1999; Martino & Marks, 1999; Melara & Marks, 1990b). Several teams have found that linguistic stimuli (words *low*/*high* or *big*/*little*) can produce behavioral congruency effects in speeded classification tasks similar to those found for simple perceptual stimuli (Gallace & Spence, 2006; Walker & Smith, 1985). They interpret these findings as generally supporting the idea that crossmodal interactions may be based in post-perceptual linguistic processing. For cross-sensory mappings that are linguistically mediated, we could expect to find involvement of the extended language system spanning perisylvian regions on the left lateral surface of the temporal, parietal and frontal lobes, and Broca’s area (Brodmann’s area 44/45).

The various semantic accounts make differing predictions about the neural regions involved, and functional neuroimaging research has provided evidence supporting both cross-sensory and more abstract semantic representations. The neural regions implicated by this family of theories vary widely depending on researchers’ operationalization of fundamental concepts such as ‘abstract’ and ‘semantic’ (Hein et al., 2007; Martin, 2007), but previous research can help constrain the accounts and interpret the potential patterns that arise. For crossmodal mappings that are semantically mediated, we could expect to find early sensory cortical regions reflecting conceptual grounding in the senses, with activity in regions such as STS/STG, and MTG, and activity reflecting more abstract semantic representations in regions such as the IT, ATL, and inferior frontal gyrus (IFG). In addition, we could expect to find involvement of linguistic systems. Many of these systems are likely to be functionally interconnected. For example, Binney et al. (2012) proposed that conceptual representations encoded in the ATL may interact with the language system and feed back to affect processing in sensory associative regions. For the purposes of the present study, patterns of activity across these regions can help to inform and constrain our models for crossmodal mappings, but we must be careful not to jump to conclusions about the directional flow of information.

The review above represents major families of current theories on the origins and neural bases of crossmodal mappings. These possibilities are by no means mutually exclusive, and for many of the crossmodal mappings people exhibit, a hybrid account appears likely. The neuroimaging and behavioral experiments conducted for this dissertation were designed to help disambiguate among the possibilities discussed above and constrain the set of potential mechanisms involved in crossmodal mappings.

Overview and Rationale of the Dissertation

In the present set of dissertation studies, we examine three crossmodal mappings that have been found to be largely consistent across individuals (Bonetti & Costa, 2017; Gallace & Spence, 2006; Parise & Spence, 2012; Rusconi, Kwan, Giordano, Umiltà, & Butterworth, 2006; Thompson & Estes, 2011). The mappings tested include 1) auditory pitch and visual-spatial elevation (high pitch tones associated with objects high in spatial elevation and low pitch tones associated with objects low in spatial elevation) 2) auditory pitch and object size (high pitch sounds associated with small objects and low pitch sounds associated with large objects) 3) auditory word form and object shape (pseudoword *keekay* associated with pointed shape and pseudoword *lohmoh* associated with rounded shape). Using fMRI, we compare the blood oxygenation level-dependent (BOLD) activity profiles for these three mappings ranging from simple perceptual stimuli to sound-symbolic linguistic stimuli. we also employ three functional localizers in an attempt to isolate systems hypothesized to play a role in crossmodal mappings and to evaluate their contributions in these three different crossmodal mappings.

In conjunction with neuroimaging and behavioral testing inside the scanner, we conducted behavioral testing outside the scanner. For each crossmodal association we tested in-scanner (pitch-elevation, pitch-size, pseudoword-shape), we also conducted behavioral testing outside the scanner to determine whether individual participants exhibited the same patterns found in previous research and reliably associated the auditory and visual stimuli used in our experiments.

The sensory dimensions examined in this study were chosen for several reasons. First, previous research had demonstrated that the crossmodal correspondences examined in the present studies are relatively robust and widespread across the population, so it was reasonable to expect that a random sample would likely include many individuals who exhibit the mappings of interest (Parise & Spence, 2009; Rusconi et al., 2006). Second, while extensive behavioral research has shown that these dimensions interact at perceptual and behavioral levels of processing (e.g. causing lagged stimulus presentation to appear simultaneous when stimuli are crossmodally matched, or slowing down response times when stimuli are crossmodally mismatched), the neural mechanisms underlying these types of perceptual mappings are poorly understood. Third, we considered it likely that the associations targeted by our different sets of stimuli would be based in different cognitive mechanisms, with pitch and size (and perhaps pitch and elevation) being correlated in perceptual experience, and domains of pseudoword and object shape being a somewhat more abstract mapping between sound structure and object form. By examining how activity differentiated for the three experiments, and by correlating performance on behavioral tasks (both in and out of scanner) with observed neural activity, we should be able to gain new insights into the contributions of various systems in cross-sensory mapping.

The experiments described in this dissertation are a first step in charting and comparing the systems involved in three different crossmodal mappings that could potentially be based in different cognitive mechanisms. Using stimuli that range from low-level perceptual to linguistic in nature, we examined the potential involvement of different neural systems (e.g. unisensory, multisensory, magnitude, semantic) in the representation of cross-sensory mappings. We hypothesized that the different crossmodal associations would produce distinct patterns of BOLD activity, differing in the extent to which unisensory, multisensory, and higher-level magnitude and semantic systems are recruited. Employing three functional localizers, we assessed involvement of multisensory integration, magnitude, and semantic functional systems hypothesized to play a role in cross-sensory mappings. By examining the extent to which the three mappings engage a common, core network, and to what extent profiles in activity differ, this study provides new insight into the contributions of these regions to our cognitive processes of crossmodal mapping and association, and begins to disambiguate among these contrasting possibilities. A review of neuroimaging literature on multisensory integration, magnitude, and semantic representations provides several a priori predictions about neural regions that may support cross-sensory mappings. These predictions offer a framework for interpreting different activity profiles that may be found, with different patterns of BOLD activity supporting different accounts of intersensory processing in the brain.

Elucidating these processes will provide insight into poorly understood aspects of sensory interaction and multisensory processing upon which complex cognitive functioning relies. Previous research has found that perceiving a meaningful stimulus in one modality (e.g., a silent video of a dog barking) can modulate activity in low-level cortices of other sensory modalities (Meyer et al., 2010; Meyer & Kaplan, 2011). We examine whether similar effects can be seen for cross-domain pairings of basic perceptual attributes that are demonstrated to be behaviorally associated (e.g., a high pitch tone and a small visual object). If we find that crossmodal mappings are based in correlated activity in early sensory cortices, this would support an intersensory-connection account by which modal experiences are directly associated, either by means of statistical learning though experience, or through pre-existing connections (Ramachandran & Hubbard, 2001). Alternatively, we may find that for cases in which we associate percepts in different senses, processing a stimulus in a given sensory modality could modulate activity in higher associative or supramodal cortical areas (Melara & Marks, 1990b, 1990a; Spence, 2011) but not in early sensory areas of the other modalities. Such a finding would be inconsistent with a strong low-level intersensory-connection account of cross-sensory mappings, and provide support for a model in which relations between the sensory modalities are established in higher-level cortical areas, at least for the mappings we examine in the present study.

Elucidating sound-meaning mappings in language. In addition to understanding crossmodal mappings more generally, one aim of the present studies is to examine how sounds in language may serve to bring online meaningful representations (Elman, 2004, 2009; Gallese & Lakoff, 2005; Lupyan, 2012; Lupyan & Bergen, 2015; Lupyan & Clark, 2015; Rumelhart, 1979). In Chapter 3, we explore the sound-symbolic mapping between pseudoword and object shape, a variation of the *takete/maluma* or *kiki/bouba* mapping that has been found across speakers of diverse languages (Bremner et al., 2013; Köhler, 1929). In the mapping, language users demonstrate preferences for mapping certain types of speech sounds such as sonorant/tonal voiced consonants to more rounded shapes, and other types of sounds including more punctate or disrupted sounds such as unvoiced stops, affricates and fricatives to more pointed objects (Blasi, Wichmann, Hammarström, Stadler, Christiansen, 2016; Köhler, 1929; Ković, Plunkett, & Westermann, 2010; Maurer et al., 2006; McCormick, Kim, List, & Nygaard, 2015; Nielsen & Rendall, 2011; Peiffer-Smadja & Cohen, 2010, 2019). It is not clear whether sound-symbolic mappings are processed in a manner similar to other language, or whether processing is qualitatively different. The experiment in Chapter 3 allowed me to assess the contributions of non-linguistic systems to the sound-symbolic pseudoword-shape mapping. By contrasting activity generated for our three localizers and pseudoword-shape pairing, we investigated the extent to which sound-symbolic language recruits systems outside canonical language areas, with particular focus on systems involved in representing non-linguistic perceptual stimuli. One possibility is that sound-to-meaning mappings in language are supported by very different systems from non-linguistic crossmodal mappings. However, it could be that sound-symbolic language activates meaningful representations by directly engaging these sensory systems outside the classic language system involved in more arbitrary language. For example, a system could be sensitive to the jarring, disrupted quality of the sounds that comprise the word keekay or the relatively gradual, smooth transitions in the word *lohmoh.* If this is the case, hearing the sound-symbolic words *keekay* and *lohmoh* could engage neural systems that are more sensory than linguistic in nature. Finding overlap in activity for sound-symbolic language and nonlinguistic stimuli could support the theory that sound-symbolic processing is largely distinct from the language system, and may be exploiting more general crossmodal correspondences to direct a listener's attention to an intended referent (Arata, Imai, Okuda, Okada, & Matsuda, 2010; Ković et al., 2010; Namy & Nygaard, 2008; Peiffer-Smadia & Cohen, 2019, Ramachandran & Hubbard, 2001; Revill, Namy, DeFife, & Nygaard, 2014). Such a finding would demonstrate functional utility of sound-symbolic mappings, and support the proposal that certain sounds are better than others for conveying certain information. Functional neuroimaging could reveal both common and distinct systems involved in representing these associations.

We designed this series of studies to chart the neural underpinnings of crossmodal mappings. By examining and comparing a range of audio-visual associations, we seek to establish which neurocognitive systems underlie each mapping, and determine whether there is a common core system supporting these mappings or alternatively, each type of mapping recruits distinct neural systems. In discovering the systems and cognitive processes involved in crossmodal mappings, we will also gain insight into myriad cognitive phenomena that are built upon on such mappings (e.g. symbol use, language, abstraction, metaphor use).

Chapter 2. Crossmodal association of auditory pitch and visuospatial elevation

An extensive body of research has documented a consistent mapping between auditory pitch and visuospatial elevation. Across these studies, individuals reliably associate higher pitch sounds with a relatively high vertical position in space, and lower pitch sounds with a relatively lower position in space (Ben-Artzi & Marks, 1995; Bonetti & Costa, 2017; Melara & O'Brien, 1987; Patching & Quinlan, 2002; Proctor & Cho, 2006; Rusconi et al., 2006). This pattern has been demonstrated across speakers of diverse languages, many (but by no means all) of which use common terms (e.g. *high* and *low*) to describe both acoustic pitch and visuospatial elevation (Eitan & Timmers, 2010; Melara & Marks, 1990b; Melara & O'Brien, 1987; Parkinson et al., 2012). Developmental research has shown that infants as young as 3-4 months display this mapping (Dolscheid et al., 2014; Walker et al., 2010, but see Lewkowicz & Minar, 2014). The finding that pitch and elevation are reliably associated across diverse cultures and from very early in development suggests that the mapping may have its basis in a system or systems common across humans. However, there is much debate as to the neural underpinnings of the phenomenon and researchers have yet to converge on an account. The present study sought to map the neural basis of this pitch-elevation correspondence, and examine the role of three systems that have been theorized to play a role in the phenomenon (multisensory, magnitude, and semantic systems).

Investigators studying the pitch-elevation mapping have employed a range of behavioral paradigms to characterize the phenomenon. The most explicit approach involves direct questioning, asking individuals about their preferences as to how stimuli in different modalities should be matched or which seem most similar (Eitan & Timmers, 2010). More implicit paradigms such as speeded classification and selective attention tasks have allowed researchers to test for cross-sensory interactions and the degree of automaticity of such processing (Ben-Artzi & Marks, 1995; Bonetti & Costa, 2017; Evans & Treisman, 2010; Melara & Marks, 1990b; Melara & O'Brien, 1987; Miller, 1991; Patching & Quinlan, 2002; Rusconi et al., 2006). Several researchers employing speeded classification tasks have found that individuals are faster to discriminate the vertical position of a visual stimulus (high or low) when it is accompanied by a tone of a crossmodally congruent pitch (relatively high or low; Ben-Artzi & Marks, 1995; Melara & O'Brien, 1987; Patching & Quinlan, 2002). In one such study, Melara and O'Brien (1987) used a speeded response paradigm to investigate a range of couplings of pitch and visuospatial elevation stimuli. In one condition, participants made responses about one dimension (e.g. high or low pitch) while values on the other, task-irrelevant dimension (e.g. vertical positioning on the computer screen) were held constant. In another condition participants made responses about one dimension while values on the other dimension varied orthogonally (some trials were crossmodally congruent and some were incongruent). They found that participants were faster to respond to the congruently paired stimuli than to stimuli in the constant condition. Interestingly, they also found that incongruent pairings did not detrimentally impact performance; participants responded just as quickly for incongruent pairings as for the constant stimulus baseline condition. So while the crossmodal congruence conferred a processing benefit, incongruence didn't interfere with, or hinder processing any more than performance in the baseline condition when stimuli in one dimension were held constant (i.e., there was no 'Garner interference'). In another study, Ben-Artzi and Marks (1995) asked participants to make speeded judgments about spatial elevation and auditory pitch and found that the two dimensions interact such that more extreme stimulus values of a congruent auditory pitch stimulus (the task-irrelevant dimension) improved accuracy in classification of a visual elevation stimulus. In another speeded classification study, Bonetti and Costa (2017) tested the effect of vertical positioning on classification of tones of different pitches. They found a spatial congruency effect; participants were faster and more accurate in classifying high-pitched tones when they were emitted from a loudspeaker positioned high in space, and low-pitched tones presented low in space than in the opposite conditions.

That individuals' processing of one perceptual dimension is affected by attribute values on another dimension has been described as a 'redundancy gain' when it leads to some improvement in behavioral performance (Melara & O'Brien, 1987) or a 'cross-sensory intrusion' when it negatively impacts performance (Martino & Marks, 1999, 2000, 2001). Such dimensional interaction has been interpreted as evidence for cross-sensory permeability of selective attention, and these crossmodal effects can distort or bias perception (Maeda, Kanai, & Shimojo, 2004; Mudd, 1963; Parise, Knorre, & Ernst, 2014; Pratt, 1930; Trimble, 1934). For example Pratt (1930), and subsequently several others (Mudd, 1963; Parise et al., 2014; Trimble, 1934) asked participants to estimate the spatial elevation of a sound source and found that the auditory pitch of the sound biased localization of the sound source. Across this cohort of studies, higher pitch sounds are estimated to originate from a higher spatial position relative to lower pitch sounds. Parise et al. (2014) found that the effect of auditory frequency of stimuli biased sound localization so strongly that participants' localization judgments were nearly independent of the true sound source. The influence of pitch on visuospatial perception also extends to dynamic stimuli. Maeda and colleagues (2004) found that playing a tone frequency sweep (rising or falling frequency) induced a visual illusion of vertical motion in a random dot display, with a rising tone inducing an illusion of upward motion and a falling tone inducing an illusion of downward motion (Maeda et al., 2004). Shintel, Nusbaum, and Okrent (2006) found a corresponding pattern in spoken language. They reported that individuals spoke with a higher fundamental frequency when describing upward motion of visual stimuli compared to when they were describing stimuli moving downward.

Developmental research has found that even prelinguistic infants (age 3-4 months) are sensitive to crossmodal congruency of pitch and elevation. Both Dolscheid et al. (2014) and Walker et al. (2010) found that infants looked longer at crossmodally congruent pitch-elevation stimuli than incongruent stimuli in a preferential looking paradigm (but see Lewkowicz & Minar, 2014). This preferential looking finding suggests that the pitch-elevation mapping is present even in young infants and demonstrates that the mapping can be present very early in life, and can be established prior to substantial language experience.

Research examining audiovisual the pitch-elevation correspondence indicates that the mapping is widely shared across the population (Parkinson, 2012) and has an array of perceptual and cognitive consequences. It is clear that a stimulus in one modality can directly influence perceptual discrimination, source localization, and attention in other modalities (Ben-Artzi & Marks, 1995; Chiou & Rich, 2012;

Eitan & Timmers, 2010; Martino & Marks, 2000; Parise & Spence, 2008; Parise, Spence, & Ernst, 2012). Yet despite the pervasiveness and far-reaching perceptual and cognitive consequences of the pitch-elevation mapping, its neural basis remains unknown. Thus the aim of the present experiment was to identify neural systems involved in the crossmodal mapping of auditory pitch and visuospatial elevation. Participants engaged in tasks that coupled auditory and visual stimuli in congruent and incongruent pairings both inside and outside the MRI scanner. On the basis of previous research on other forms of audiovisual congruency (Hein et al., 2007; van Atteveldt et al., 2004), we reasoned that neural systems supporting the pitch-elevation mapping would likely be sensitive to audiovisual congruency of our stimuli, and differences in BOLD activity (in-scanner) and that corresponding differences in reaction times and accuracy (for testing both in- and outside of the scanner) would be found for the congruent and incongruent stimulus conditions.

In addition to the main experiment, three independent localizer tasks were employed to functionally define neural systems involved in audiovisual integration, magnitude, and semantic processing, all of which have been theorized to play a role in crossmodal mappings (these systems are detailed in Chapter 1, with specific implementation of localizers is described in the Methods section). To test the hypotheses that these three systems could be involved in the crossmodal correspondence of pitch and elevation, analyses compared the locations of activity produced for congruent and incongruent couplings of pitch-elevation stimuli, and identified overlaps between this activity and the regions highlighted by the localizer tasks (conducted in the same individuals).

Previous research has provided theoretical rationale for how each of these systems could underlie the pitch-elevation correspondence. One account for the pitch-elevation correspondence is that it reflects the statistics of perceptual experience in the environment (Cabrera & Morimoto, 2007; Jamal et al., 2017; Parise et al., 2014). A recent study by Parise, Knorre, and Ernst (2014) indicated that pitch and elevation are, in fact, directly correlated in natural scene statistics, with higher-pitched sounds tending to originate from sources that are relatively higher in elevation compared to lower-pitched sounds. There is also research indicating that this correlated perceptual experience is further biased by sound filtering properties of the head and outer ear, which somewhat attenuate higher frequency sounds emanating from low in space. Thus, in our cumulative experience, the higher pitched sounds we hear tend to have sources higher in space than the lower pitched sounds (Cabrera & Morimoto, 2007; Parise et al., 2014) and this crossmodally correlated experience biases our expectations about the types of sounds that should come from visual sources that are relatively high and low in space. If the pitch-elevation correspondence is based in this type of multisensory integrative processing, we could expect BOLD activity for our pitch-elevation experiment and the multisensory integration localizer to be co-localized, as it selects for regions that respond to concurrent auditory and visual stimuli, a response profile we would expect in system integrating signals and localizing their sources in the environment.

An alternative account is that the association of pitch and elevation is based in magnitude representations (Lourenco & Longo, 2011; Piazza et al., 2004, 2007; Pinel, Piazza, Le Bihan, & Dehaene, 2004; Walsh, 2003). By this theory, both auditory pitch and visual elevation could be represented in terms of amount or extent along a scaled or 'prothetic' dimension with one end 'less' and one end 'more'. Contrasting stimulus values in different modalities could thus be aligned and associated by virtue of this 'more-less' relationship along their respective dimensions. So if a high-pitched sound and a visual object located high in space are both encoded as 'more than' a low pitched-sound and a visual object low in space, this could be a basis for associating these two types of stimuli. If the general magnitude system is a basis for the pitch-elevation mapping, we expect activity related to crossmodal congruency from the experiment will be co-localized with activity for the magnitude localizer (see Chapter 1 for review of neuroanatomical basis of magnitude system).

A third possibility is that the association of high- and low-pitched sounds with high and low spatial elevations, respectively, may be linguistically or semantically mediated (Spence, 2011; Walker et al., 2012; P. Walker, 2012; Walker & Smith, 1984). It has been suggested that the use of a common label, or *polysemy*, across two otherwise unrelated domains could lead us associate them (Walker & Smith, 1984; Spence, 2011). For example, using the terms '*high*' and '*low*' to refer to both dimensions of pitch and visuospatial elevation could lead us to think of these two domains as similar (Martino & Marks, 1999; Melara & Marks, 1990b; see Chapter 1 for expanded discussion of semantic coding hypothesis). A few studies have found behavioral congruency effects or interference effects when participants are classifying either word based stimuli 'high' and 'low' (words presented either auditorily or written) or that are high or low in space (Melara & Marks, 1990b, Ben Artzi & Marks, 1995; Shor, 1975). The authors of these studies note the similarity in behavioral responses obtained for linguistic and perceptual stimuli, and have interpreted these findings as supporting the possibility that language could provide a basis for the mapping. We examine this possibility using a language-based semantic localizer.

These three candidate systems provided a set of a priori predictions about the neural basis of pitch-elevation mapping. By comparing the distribution of congruency-related activity from the main pitch-elevation experiment with the regions highlighted by these localizers, we test the contributions of these functional systems to the pitch-elevation mapping (McCormick, Lacey, Stilla, Nygaard, Sathian, 2018a).

Method

Participants

Twenty healthy adults were recruited from the Emory University community. Two participants were later excluded from analyses due to excessive movement during scanning (>1.5mm), leaving a total of 18 participants (nine male, nine female) in our dataset (mean age 24.8 years, range 19-33 years). All participants were native speakers of English and were right-handed as determined by a validated subset of the Edinburgh handedness inventory known to have particularly high validity in determining hand use (Raczkowski, Kalat, & Nebes, 1974). Participants reported normal hearing and normal or corrected-to-normal vision, and none reported or showed signs of neurological disorders. All participants gave informed written consent and were compensated for their time. The study protocol was approved by the Emory University Institutional Review Board.

Stimuli

For the main experiment we employed two sets of stimuli (one auditory, one visual), which contrasted along dimensions of auditory pitch and visuospatial elevation respectively. Auditory stimuli consisted of two tones, one low-pitched (180 Hz), and one high-pitched (1440 Hz), both with durations of 200 ms (including a 20 ms ramp at the onset and offset). The visual stimulus consisted of a gray circle (RGB values: 240, 240, 240, diameter 70 pixels/subtending approximately 1° of visual angle) with its center positioned either high on the screen (300 pixels/approximately 4.2° of visual angle above fixation cross) or low on the screen (300 pixels/approximately 4.2° of visual angle below fixation cross). These simple stimuli were combined to create audiovisual triplets, comprising three repetitions of identical stimuli (200ms on, 200 ms off) presented over the course of 1000 milliseconds. Multisensory stimulus pairings were either crossmodally congruent (high pitch+high elevation or low pitch+low elevation) or incongruent (high pitch+low elevation, low pitch+high elevation) with respect to the crossmodal pitch-elevation correspondence (see Fig. 1). Visual stimuli were projected onto a screen at the back of the magnet bore and viewed through a mirror angled over the head coil. Auditory stimuli were presented via scanner-compatible headphones.

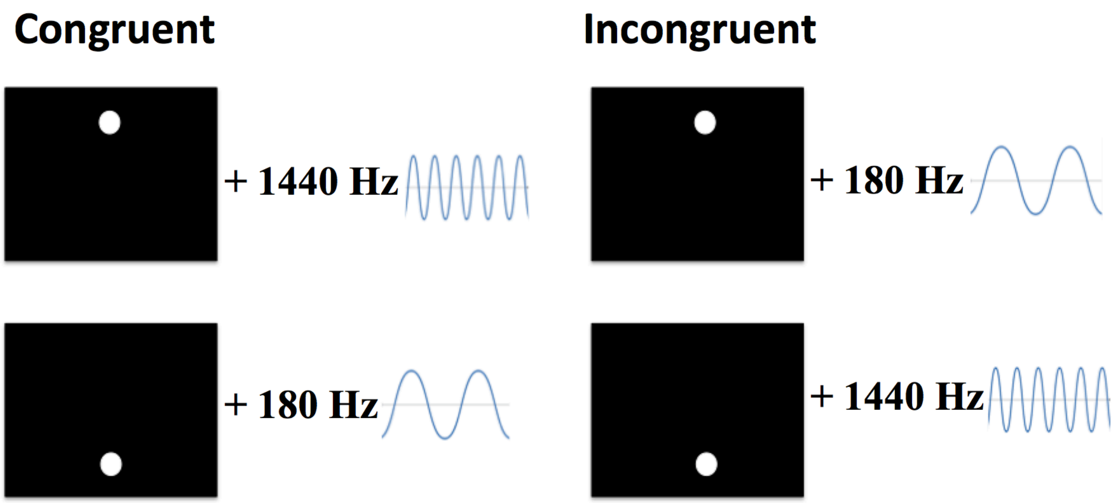


Figure 1. Visual and auditory stimuli for Pitch-Elevation experiment. Stimulus pairings were either congruent or incongruent. Each of the four pairings was presented in 40 trials across four functional runs.

Procedure

General. Scans for this study were conducted in 1-2 sessions depending on the number of dimensional pairings being tested. In addition to the pitch-elevation experiment, a subset of the participants (n=9) were tested on pitch-size dimensions (findings are reported in Chapter 4); for these participants, pitch-elevation scans and localizer scans were performed in separate sessions approximately 1-2 days apart (experimental scans always preceded localizers). The other nine participants completed the pitch-elevation experiment and localizer scans in a single session. Participants first performed the four runs of the pitch-elevation task, followed by three functional localizer scans (described in the localizer section below). This fixed task order was followed in order to avoid possible priming effects of the localizer and behavioral tasks on the pitch-elevation scans. The order of localizer tasks was also fixed, so that the tasks progressed from most difficult to least difficult: participants completed the magnitude localizer, followed by the temporal synchrony localizer, and finally the semantic localizer task. Each localizer comprised two runs; each run had fixed stimulus order and run order was counterbalanced across participants. Additional behavioral testing was conducted following scan sessions in order to confirm that participants exhibited the expected crossmodal mappings. All experiments were presented using

Presentation software (Neurobehavioral Systems Inc., Albany, CA) administered on a laptop computer, which synchronized stimulus presentation with fMRI scanning, and recorded button-press responses and response latency for responses made using a scanner compatible hand-held button box.

Pitch-elevation fMRI task. Prior to testing, participants were fitted with earplugs and scanner-safe headphones (Beyerdynamic DT 100). Once inside the scanner, participants were played the high-pitched tone stimulus at a range of amplitudes and asked to select one that was sufficiently loud but not uncomfortable and which would be clearly audible over scanner noise. They were then asked to select a low-pitched tone from an array matching the apparent loudness to the 1440 Hz tone. This step was necessary because tones at different frequencies differ in perceived loudness and we wanted to avoid experimentally confounding loudness and pitch for each participant (Moore, 2012; Suzuki & Takeshima, 2004). The selected tones were then used as the high- and low-pitched sound stimuli in the test phase. Participants selected high-pitched tones that ranged from approximately 95 to 102 dB SPL and low-pitched tones that ranged from approximately 85 to 92 dB SPL. On average, tones selected for the high-pitched stimuli were 10 dB SPL greater intensity than the tones selected for the low-pitched stimuli.

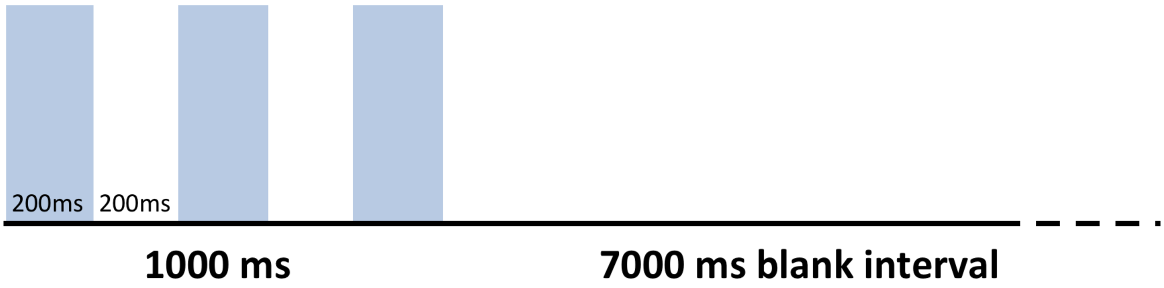


Figure 2. Trial structure in the Pitch-Elevation experiment. Audio-visual stimuli were presented in a triplet pulse over the course of 1000 ms.

A high-resolution anatomical volume was collected prior to functional testing. Functional data for the pitch-elevation experiment were collected over the course of four runs (run duration 370 s) in a slow event-related design. Each run consisted of 40 multisensory trials with 5 rest periods (duration 10 s) interleaved. An auditory cue (‘rest’) played at the beginning of each rest period, and another cue (‘ready’) indicated when a rest period was almost over and the task was about to resume. Stimuli within a run were presented in pseudo-random order, with immediate repetition of a given token (in back-to-back trials) occurring on 25% of the trials in each run. The one-second stimulus was followed by a blank interval of 7 seconds (8 seconds from one stimulus onset to the next; see Fig. 2). In each run, 20 of the trials consisted of congruently paired audiovisual stimuli (high pitch+high vertical position or low pitch+low vertical position object) and 20 of the trials consisted of incongruently paired stimuli (high pitch+low vertical position object or low pitch+high vertical position object), in a pseudorandom order. Each of the four unique stimulus pairings was presented in 10 trials per run (totaling 40 trials across experimental runs). The order of runs was counterbalanced across subjects.

Participants engaged in a one-back same/different detection task, responding ‘same’ by button-press when a given pairing of audiovisual stimuli was presented in two consecutive trials (25% of trials), and responding ‘different’ when stimuli were not identical across trials (when auditory, visual, or both auditory and visual were different from the previous trial; 75% of trials). Trial transitions for each of the four unique stimuli were equiprobable to ensure that the stimulus presented in a given trial was not predictive of the stimulus in the following trial. Performing this task accurately required that participants attend to the perceptual dimensions of interest while keeping extraneous task demands to a minimum.

Functional localizer tasks. To generate data-driven predictions about where crossmodal associations would be likely to be represented, three types of functional localizer tasks were run. For each of the three functional localizers, we contrasted the two task conditions within the given localizer, isolating voxels that showed a selective BOLD response for one of the conditions in contrast to the other. The activity observed for each localizer task established the neural regions sensitive to temporal synchrony of audiovisual stimuli, magnitude-related processing, or semantic processing, and which we would expect to be active under the various competing hypotheses of crossmodal association.

Multisensory integration localizer. While there are many possible approaches to test for multisensory integration, we reasoned that mappings originating out of the statistical regularities of perceptual experience in the environment could potentially have a basis in the temporal synchrony system, which plays an important role in aligning and binding co-occurring sensory signals from the environment and producing a unified percept. Therefore, in designing this localizer we focused on isolating areas sensitive to audiovisual synchrony. Similar to many previous studies on multisensory integration (van Atteveldt et al., 2007; Beauchamp, 2005a, 2005b; Erickson et al., 2014; Marchant, Ruff, & Driver, 2012; Noesselt et al., 2012), our multisensory integration localizer was designed to identify areas where activity for paired audio-visual stimuli presented simultaneously was greater than activity when audio and visual stimuli were temporally-offset. This localizer consisted of simple auditory and visual stimuli presented in temporally synchronous and asynchronous conditions, and employed an oddball detection task to ensure that participants attended to the stimuli. Stimulus attributes were defined by intermediate values relative to those used in the experimental runs. The auditory stimulus was a tone of 810 Hz (intermediate between the180 Hz and 1440 Hz used in the main experiment) that was 800ms in duration, including 20ms ramp at the onset/offset. The visual stimulus was a gray circle (70 pixels in diameter) subtending approximately 1° of visual angle and presented centered on the screen. Trials had a duration of 4000 ms (4 trials per 16 s block). In the synchronous condition, auditory and visual stimuli were presented simultaneously for 800 ms followed by a 3200 ms ITI. In the asynchronous condition, auditory and visual stimuli were presented for 800 ms each with an intervening blank interval of 200 milliseconds followed by a 2200 ms ITI (Beauchamp, 2005b; Wallace, Meredith, & Stein, 1992, 1998). Half of the asynchronous trials presented auditory then visual and half of the trials presented visual then auditory stimulus (see Fig. 3). The localizer had a block design (synchronous, asynchronous, and rest conditions) and was collected in two runs (run duration 374s, each containing 12 active blocks (6 synchronous blocks x 16 seconds) + (6 asynchronous blocks x 16 seconds), with 13 rest blocks (14 seconds) between the active blocks.

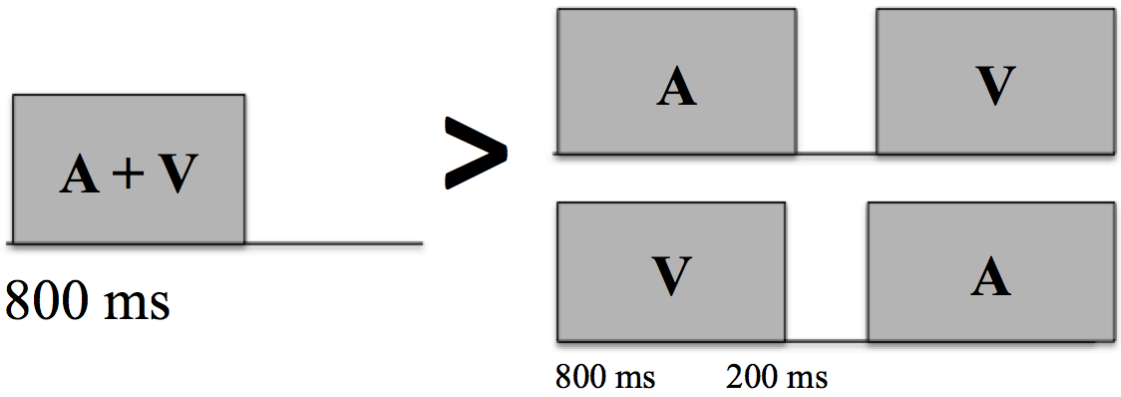


Figure 3. Multisensory Temporal Synchrony localizer- design schema for synchronous and asynchronous stimulus conditions.

Participants were asked to press a button when an oddball, either a square or a burst of white noise, was presented in either modality (Crottaz-Herbette & Menon, 2006). In each run, two visual oddball trials and two auditory oddball trials were presented, one in a synchronous block and one in an asynchronous block.

To isolate voxels that were selective for simultaneous auditory and visual stimulus presentation, we contrasted Synchronous>Asynchronous trials. We expected that the synchronous>asynchronous contrast would show activation in the STS and adjacent regions of superior temporal gyrus (STG) as has been reported in a number of previous studies on multisensory temporal synchrony (van Atteveldt et al., 2007; Beauchamp, 2005a, 2005b; Erickson et al., 2014; Marchant, Ruff, & Driver, 2012; Noesselt et al., 2012).

Magnitude localizer. To identify brain regions sensitive to magnitude, we adapted an estimation task developed by Lourenco, Bonny, Fernandez, and Rao (2012). On each trial, participants judged whether there were more black or white elements in a visual array of small rectangle shapes (see Figure 5A). We designed a control task using a modified set of these stimuli, with a single triangle element appearing within the array of objects. For the control task, subjects indicated whether the triangle in each array was black or white (see Figure 5B), thus the response, black or white, was the same for both the magnitude and control conditions. Data for the Magnitude localizer were collected over the course of two runs (run duration: 374 s). Each run consisted of 12 active blocks per run (6 magnitude x 16 seconds) + (6 control x 16 seconds) with 13 rest blocks (14 seconds) between the active blocks (see Fig. 4). For both tasks, each test block was comprised of four 4-second trials with stimuli presented for 1 second and a 3 second ISI. The contrast of magnitude estimation>control identified regions sensitive to magnitude. We expected that magnitude-sensitive regions would be appear in the posterior parietal cortex, including the IPS (Eger et al., 2003; Lourenco & Longo, 2011; Piazza et al., 2004, 2007; Pinel, Piazza, Le Bihan, & Dehaene, 2004; Sathian et al., 1999; Walsh, 2003) .

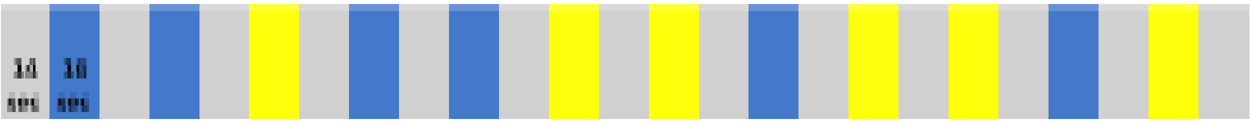
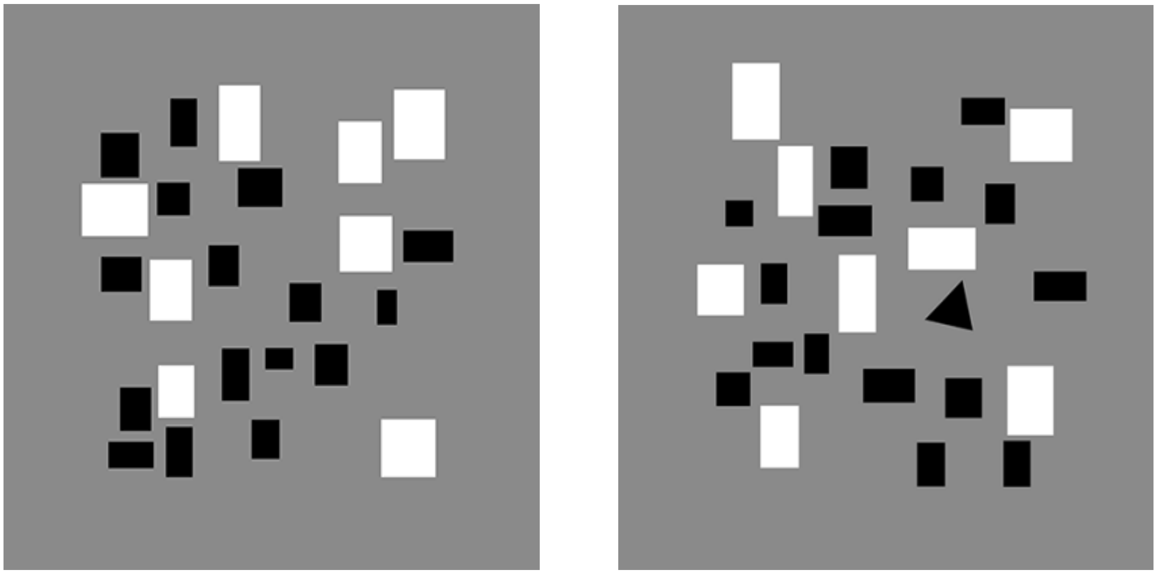


Figure 4. Block design of a magnitude localizer run showing blocks of magnitude task (yellow) and shape-finding task (blue).



A. Are there more black or white shapes? B. Is the triangle black or white?

Figure 5. The magnitude localizer contrasted BOLD activity for magnitude (A) and shape-finding (B) tasks.

Semantic localizer. A semantic localizer was adapted from Fedorenko et al. (2010). In this task, participants read complete sentences and strings of pronounceable non-words; both presented one word at a time. Following each string, a cue appeared indicating participants should press a button. By contrasting activity produced when reading sensible (semantically and syntactically intact) sentences with activity produced when reading strings of non-words, this localizer identifies brain regions sensitive to word- and sentence-level meaning. The localizer was collected in two runs. Run duration was 492 seconds, and each run contained 16 active blocks (8 semantic blocks x 18 seconds) + (8 non-word blocks x 18 seconds), with 17 rest blocks (12 seconds) between the active blocks. Sentences consisted of 12-word sequences (each written word presented for 450 ms, for a total of 5.4 seconds) and were followed by a 600 ms screen indicating that subjects should press a button. This task required that participants remain attentive, although it did not explicitly require language processing or output in order to successfully complete the task (one could merely press the button on the cue). However, previous studies (Fedorenko, Behr, & Kanwisher, 2011) have found this task has good convergent validity with other approaches to isolating the extended language system. Blocks for this Semantic condition were comprised of three such sentence strings (totaling 18 seconds; Fig 6). For the contrast condition, non-word blocks with same trial structure presented strings of pronounceable non-words. Rest blocks were 12 seconds long. Contrasting these conditions, we identified regions which responded preferentially to sensible sentences as compared to non-word strings We expected the sentences>non-word strings contrast to show activity in regions involved in both semantic and syntactic processing, including a widespread network in the left hemisphere including the inferior frontal gyrus (IFG), inferior parietal cortex, along with swaths of the temporal lobe including the superior temporal sulcus (STS) (Bedny, Pascual-Leone, Dodell-Feder, Fedorenko, & Saxe, 2011; Fedorenko, Behr, & Kanwisher, 2011).

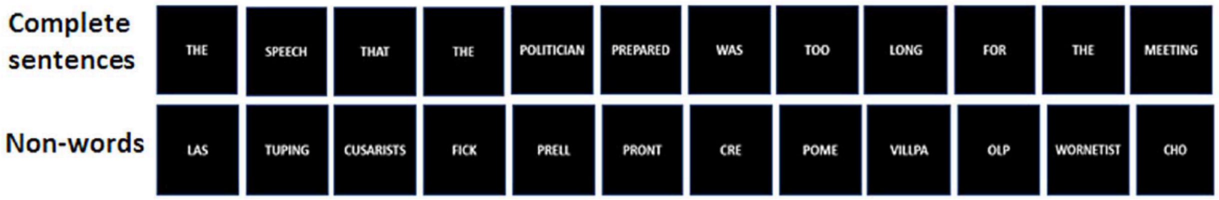


Figure 6. The semantic localizer contrasted complete sentences with strings of non-words.

Post-scan behavioral testing: Implicit association of Pitch and Elevation. In conjunction with the fMRI testing described above, we conducted behavioral testing outside the scanner to establish that participants reliably demonstrated the predicted mappings between the specific audiovisual stimulus pairings in our study. As discussed above, previous research has employed a variety of speeded response tasks demonstrating interactions between dimensions or modalities (Bernstein & Edelstein, 1971; Evans & Treisman, 2010; Lacey et al., 2016; Marks, 1987; Melara & O'Brien, 1987; Parise & Spence, 2012; Rusconi et al., 2006). If the stimuli assigned to a particular response key are congruent, the participant is faster to respond than if they are incongruent. We used an implicit association paradigm (IAT) for a few reasons: first, presenting each unimodal stimulus in isolation eliminates possible confounds of selective attention that could arise from multisensory stimuli; second, the IAT provides an objective measure of preferred crossmodal mappings (using reaction time as an index of cognitive preference) without overt questioning about preferences (Greenwald, Mcghee, & Schwartz, 1998); and third, several studies have successfully used an IAT to demonstrate congruency effects between pitch and elevation stimuli very similar to ours (Jamal et al., 2017; Parise & Spence, 2012).

The same basic unimodal stimuli from the neuroimaging experiment were used except that stimuli were presented for 1000ms and the auditory stimuli had 100ms ramp at tone onset/offset. Following the neuroimaging study described above, participants were seated at a desktop computer and fitted with headphones. As described for the neuroimaging experiment, participants selected two tones (1440 Hz and 180 Hz) that were clearly audible and sounded equally loud. The selected tones were then used as the high- and low-pitched sound stimuli in the test phase. The experiment was administered using

Presentation software (Neurobehavioral Systems Inc., Albany, CA), which also recorded responses and RTs. Stimuli were presented for up to 1000 ms (stimulus presentation was cut short if the participant made a response faster than 1000 ms.

Testing was conducted in eight blocks of 48 trials per block. Participants went through two blocks (96 trials) in each of the four response key mappings. Preceding each test block, participants were presented with an instruction screen assigning particular response keys to press when each of the four unique stimuli were presented. Participants were instructed to associate pairs of stimuli (one auditory stimulus, one visual stimulus) with one of two response keys (the 'left' and 'right' arrow keys on a standard US QWERTY keyboard. Participants then engaged in a 12-trial practice block in which they received feedback on their accuracy following each response. Critically, within a given block the same response key was used to respond for both an auditory and a visual stimulus (see Table 1). This allowed for either congruent or incongruent stimulus-response key couplings. For instance, in two of the eight testing blocks, participants were asked to press one button when they saw an object presented high on the screen or heard a high-pitch tone, and press the other button when they saw an object low on the screen or heard a low-pitch tone. These response mappings were coupled in configurations we expected to be experienced as 'congruent' for four of the test blocks and 'incongruent' during the other four blocks. Block order was counterbalanced across participants, such that half of the participants started the testing session with a congruent key mapping, and half started with an incongruent mapping. Additionally, each subject used both the left and right buttons to respond about high and low stimuli. This was important because previous research has shown that spatial configuration of response keys can interact with stimuli. For example, when using left and right response keys, individuals are faster to respond with a left key-press for visual stimuli lower in vertical space and with a right key-press for stimuli higher in vertical space (Lu & Proctor, 1995), and are faster to respond with the left for low-pitched tones and the right for high-pitched tones (Rusconi et al., 2006). Participants were given opportunities for rest between blocks.

Table 1. Response key combinations for Implicit Association Task (IAT). Each of the four key mappings was used for two test blocks (out of 8 blocks total).

	Response key mapping	
	Left key	Right key
Congruent blocks	Low Pitch/Low Elevation	High Pitch/High Elevation
	High Pitch/High Elevation	Low Pitch/Low Elevation
Incongruent blocks	Low Pitch/High Elevation	High Pitch/Low Elevation
	High Pitch/Low Elevation	Low Pitch/High Elevation

Subjects engaged in the speeded response task, pressing a button to indicate which of the four unique stimuli was presented. Testing was self-paced, with a trial ending as soon as a subject made a response. A 1000 ms inter-trial interval (ITI) ensured that the trials were perceptually distinct. In the event that a participant made no response, trials would terminate after 4.5 seconds and advance to the next trial. Presentation software was used to present stimuli and record button-press responses and response latency.

Image acquisition. MR scans were performed using a 3-Tesla Siemens Trio scanner (Siemens Medical Solutions, Malvern, PA), using a 12-channel head coil. T2*-weighted functional images were acquired using a single-shot, gradient-recalled, echoplanar imaging (EPI) sequence for BOLD contrast. For all functional scans, 34 axial slices of 3.1mm thickness were acquired per volume (whole brain) using the following parameters: repetition time (TR) 2000ms, echo time (TE) 30ms, field of view (FOV) 200mm, flip angle (FA) 90°, in-plane resolution 3.125×3.125mm, and in-plane matrix 64×64. High-resolution 3D anatomic images were acquired using an MPRAGE sequence (TR 2300ms, TE 3.9ms, inversion time 1100ms, FA 8°) comprising 176 sagittal slices of 1mm thickness (FOV 256mm, in-plane resolution 1×1mm, in-plane matrix 256×256). Once magnetic stabilization was achieved in each run, the scanner triggered the computer running Presentation software so that the sequence of experimental trials was synchronized with scan acquisition. Volume acquisition for the each run was as follows: 246 volumes in each of two semantic runs, 187 volumes in each of two multisensory temporal synchrony runs, 187 volumes in each of two magnitude runs, and 185 volumes for each of four multisensory pitch-elevation runs.

Image processing and analysis. Image processing and analysis was performed using BrainVoyager QX v2.8.4 (Brain Innovation, Maastricht, Netherlands). In individual analysis, retrospective processing was conducted on each participant's functional runs which had been real-time motion-corrected utilizing Siemens 3D-PACE (prospective acquisition motion correction). Functional images were preprocessed utilizing cubic spline interpolation for slice scan time correction, trilinear-sinc interpolation for intra-session alignment of functional volumes (all functional volumes motion-correction-aligned to first volume of functional run closest to anatomical scan), and high-pass temporal filtering to 2 cycles per run to remove slow drifts in the data, Anatomic 3D images were processed, co-registered with the functional data, and transformed into Talairach space (Talairach & Tournoux, 1988). Talairach-normalized anatomic data sets from multiple scan sessions (2-3 per participant) were averaged for each individual, to minimize noise and maximize spatial resolution.

For group analysis, the transformed data were spatially smoothed with an isotropic Gaussian kernel (full-width half-maximum 4mm). The 4mm filter is within the 3-6mm range recommended to reduce the possibility of blurring together activations that are in fact anatomically and/or functionally distinct (White et al., 2001). Runs were normalized using the "percent signal change" option in BrainVoyager (signal in a voxel at each timepoint is divided by the mean timecourse signal and multiplied by 100).

For group activation display and statistical analysis, we created a group average brain. A participant-specific Talairach template was created in BrainVoyager by first selecting a representative "target" Talairach-normalized brain from the 18-participant group. Each of the 17 remaining "non-target" Talairach-normalized brain was individually co-registered to the target brain's gyral/sulcal pattern and then transformed to 3d-space using since interpolation. These 17 individual aligned/transformed brains were then averaged using the “combine 3D data sets” option in BrainVoyager. This “average” brain was then combined with the single “target” brain, creating a group-specific Talairach template. The 18-subject Talairach template was then manually segmented to generate a group average cortical voxel mask file with 3mm spatial resolution, equivalent to the spatial resolution of the functional data files. This group average brain was then used to display group activation maps using the real-time volume rendering option in BrainVoyager QX. All subsequent analyses at the group-level were restricted to this cortical mask.

Statistical analysis of group data used random effects general linear models (GLM) treating participant as a random factor (so that the degrees of freedom equal n-1, i.e. 17), restricted by a mask of cortical voxels (see below), followed by pairwise contrasts. This analysis allows generalization to untested individuals. Correction for multiple comparisons within a cortical mask (corrected $p<0.05$) was achieved by imposing a threshold for the volume of clusters comprising contiguous voxels that passed a voxel-wise threshold of $p<0.001$, using a 3D extension (implemented in BrainVoyager QX) of the 2D Monte Carlo simulation procedure described by Forman et al. (1995). Following recommendations of Woo et al (2014), and Eklund et al (2016), the stringent correction threshold of $p<.001$ was applied to minimize potential false positive results and also permit more accurate spatial localization of activations than when the more liberal thresholds are used (Eklund, Nichols, & Knutsson, 2016; Woo, Krishnan, & Wager, 2014). Activations were localized with respect to 3D cortical anatomy with the help of an MRI atlas (Duvernoy, 1999).

Results

Behavioral

In –scanner tasks.

Localizer tasks. For the multisensory integration localizer, participants correctly identified a mean of 7.11 out of 8 oddball targets and had a mean false alarm rate of 3.72 false alarms over the two runs. In other words, participants correctly identified (mean ± sem) 88.9±4.2% of the oddball trials on average. For the magnitude localizer, accuracy for both the magnitude estimation and the control condition was near ceiling for most participants, with an insignificant trend for more accurate responses in triangle task (96.9% correct) versus magnitude (94.7% correct) task ($t_{17} = -1.02$, $p = .3$). One participant showed markedly poorer performance on the magnitude task (65.7%) as compared to the triangle search (100%). Response times were significantly faster for the control task (980±60ms compared to 1109±64ms; $t_{17} = 4.53$, $p < .001$, d=.5). For the semantic localizer, most participants had accuracy near ceiling, correctly responding to the visual cue at the end of each sentence or non-word list in 98.5±0.8% of trials on average. There was a mean of 1.06 false alarms per subject over the course of the two runs.

Pitch-elevation task. To prepare accuracy and reaction time data for analysis, we excluded trials for which there was no response (n=326, 11.3% of all data). The remaining dataset (all trials for which there was a response) was used to calculate overall accuracy (correct/correct+incorrect). Additional analyses decomposed the dataset to compare accuracy for i) task conditions: same versus different trials and ii) stimulus conditions: congruent versus incongruent trials. To prepare reaction time data for analysis, we then filtered incorrect responses (n=136, 5.3 % of responses). We then trimmed outliers from the remaining dataset (comprising 94.7% of responses) by calculating subjects' mean response times, then trimming responses with latencies in excess of ±2.5 standard deviations from each subjects' mean. This resulted in the exclusion of 76 responses or 3.14% of the correct trials (mean 4.2 responses trimmed per subject, range of 3-6 trials per subject). With the trimmed dataset, we calculated mean RTs for congruent and incongruent conditions for each subject.

We then interrogated the dataset to determine whether audiovisual congruency of trials affected task performance in the scanner (either in terms of overall accuracy or response time latency (RT)). For the in-scanner task, participants judged whether audiovisual stimuli in each trial were the same or different from the preceding trial. There was a difference in the relative frequency of 'same' trials (25%) and 'different' trials (75%) in the experiment, potentially leading to response bias. However, overall accuracy was comparable for *same* (mean ± sem) (93.4±2.3%,) and *different* (95.0±1.9%,) task conditions (paired samples *t* test; t_{17} =-0.559, *p*=0.583, two-tailed; Fig. 7) and false alarm rates were low (6.6% and 5.0% respectively), so we did not go on to correct for response bias by calculating d'. Mean accuracy did not differ significantly for congruent trials (94.5±1.6%,) versus incongruent trials (94.8±1.5%) (paired samples t test; t_{17} =-.453, *p*=0.657, two-tailed). Although there was not a significant effect of congruency on accuracy at the group level, some individuals exhibited differences in patterns of responses across the two conditions. Six subjects were more accurate in the congruent condition, while ten were more accurate in the incongruent condition, and two were equally accurate in the two conditions (behavioral data for the task are summarized in Fig 8). Subjects XM10 and XM12 showed pronounced differences in their performance for same and different trials (Fig 7.), raising the question as to whether they failed to *detect* the attributes that made a given trial 'same' or 'different', or simply misunderstood the instructions and failed to recognize that particular attributes of the stimuli they detected qualified the trial as being 'same' or 'different' and make the appropriate response. A within-subject paired-samples t-test revealed no reliable differences in RTs for congruent (1197±75ms) and incongruent task conditions (1199±75ms) (paired samples *t* test; t_{17} = -0.194, *p*=0.848; Fig. 9). Nine of the subjects had faster RTs for the congruent condition and nine were faster for the incongruent condition. Further, a within-subject paired-samples t-test compared only the congruent trials that had been immediately preceded by congruent trials (CC trials) and incongruent trials preceded by incongruent trials (II trials). This comparison revealed no significant difference in RTs for the two types of trials (CC RTs 1178±78ms, II RTs 1185±74ms t_{17} =-0.293, *p*=0.773). Accuracy was not significantly different for the CC (95.9±1.3%) and II (96.5±1.1%) conditions (t_{17} =-.91, *p*=.4). These results were somewhat surprising in light of previous research reporting faster responses for congruent audiovisual stimulus pairings than incongruent pairings in an array of speeded response paradigms. It is worth noting that a majority of these experiments involved a task in which subjects made responses about attributes of the immediate stimuli (e.g., classifying whether a tone was high or low in pitch). In our task, subjects were asked to compare the stimuli in the immediate trial with a prior trial. These task demands may have had the effect of making multisensory stimulus congruency less salient than in the classification tasks. Van Wanrooij, Bremen, and Van Opstal (2010) found that multisensory congruency effects can be sensitive to statistical likelihood of audiovisual stimulus couplings, noting that congruency benefits (faster response time, higher accuracy) occurred when auditory and visual stimuli were incongruently aligned in 50% or more of trials (Van Wanrooij et al., 2010). Perhaps the lack of a behavioral congruency effect in our data is due to the equal statistical likelihood of stimuli being incongruent or congruently paired.

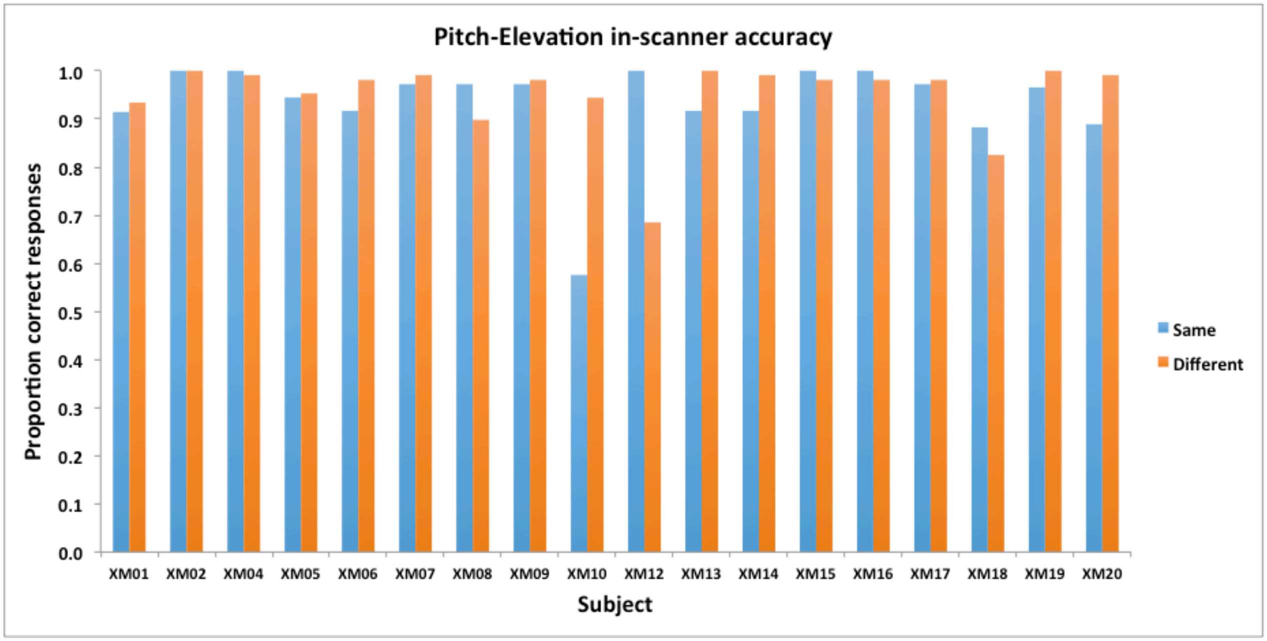


Figure 7. Accuracy for 'same' and 'different' trials of the in-scanner task.

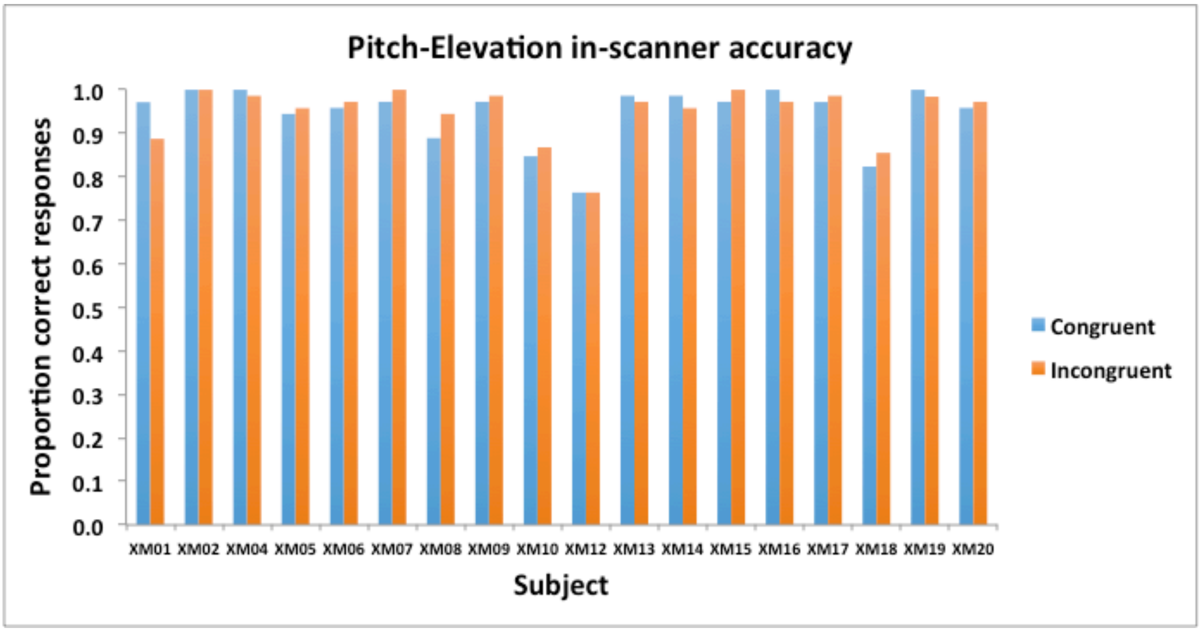


Figure 8. Accuracy for crossmodally congruent and incongruent trials of the in-scanner task.

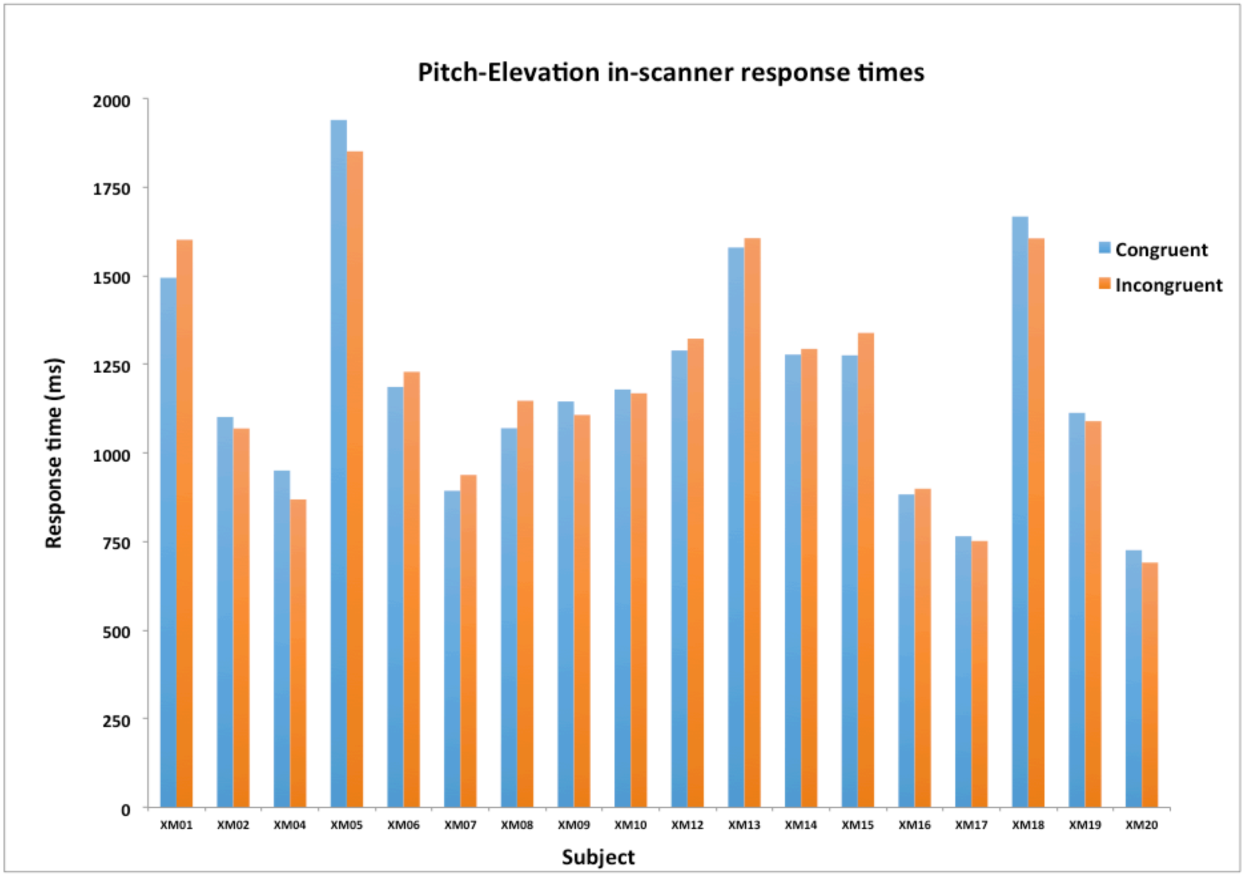


Figure 9. Response times for crossmodally congruent and incongruent trials of the in-scanner task.

Post-scan pitch-elevation IAT. We investigated whether audiovisual congruency of trials affected task performance on the IAT (either in response time latency (RT) or overall accuracy). A logging error resulted in no data logged for 33 trials (across 6 subjects). This omission represents 0.48% of the trials. A 2x2 (congruency x modality) repeated-measures ANOVA compared accuracy for the congruent and incongruent conditions in auditory and visual modalities (with both factors within-subjects). There was a significant main effect of congruency. Subjects were significantly more accurate in the congruent key-mapping conditions (96.1± 0.4%) compared to incongruent conditions (91.5± 1.5%; F1,17 = 10.19, p = .005, d=1.5; see Fig. 10). There was no main effect of modality on accuracy (auditory 94.0±0.8%, visual 93.6±1.1%; F1,17 = 0.13, p = .72, d=.18).

To prepare reaction time data for analysis, we excluded incorrect responses and trimmed outliers. we calculated subjects mean response times for auditory and visual conditions separately, trimming responses with latencies in excess of ±2.5 standard deviations from each subjects' mean for correct trials. For the auditory condition, this resulted in the exclusion of 97 responses, or 3.0% of the data being trimmed (mean 5.39 responses trimmed per subject, range of 3-8). For the visual condition, 89 responses, or 2.8% of the data, were trimmed (mean 4.94 responses trimmed per subject, range of 2-7). With the remaining dataset, we calculated mean response times for congruent and incongruent trial conditions in auditory and visual stimulus modalities. Response times (across trials) were submitted to a 2x2 repeated-measures ANOVA using congruency (congruent vs. incongruent) and modality (Auditory vs. Visual) as the within-subject factors. There was a significant main effect of Congruency ($F_{1,17} = 28.5$, $p < .001$, d=2.6). Participants were faster to respond when response key mappings were congruent (553±23ms) than when they were incongruent (695±37ms). As illustrated in Figures 11 and 12, this pattern is stable across individuals in the sample; seventeen out of eighteen participants exhibited the expected response pattern, being slower to respond in the incongruent condition. There was also a main effect of Modality on response times ($F_{1,17} = 105.8$, $p < .001$, d=4.9), with participants responding more quickly for visual (537±26ms) compared to auditory (711±32ms) stimuli. The interaction of congruency and modality was not significant ($F_{1,17} = 3.8$, $p=0.07$; Fig. 13). The main effect of modality on response times indicates that stimulus values in the two modalities were not equally discriminable, and leaves open the question as to whether we would find comparable effects of crossmodal congruency if our auditory and visual stimuli were equally discriminable.

Participants' performance on post-scan speeded-classification task showed robust congruency mappings. Every participant except one showed a congruency effect (faster RTs for congruent trials). The remaining participant showed a congruency effect for the visual, but not the auditory, stimuli. Although participants were responding to unimodal stimuli (either auditory or visual in any given trial), inter-modal interactions were evident from participants' faster/slower response times depending on response key mappings and corresponding perceptual dimensions. These findings accord with previous research indicating pitch-elevation correspondences and demonstrate the psychological reality of the crossmodal correspondences examined in this dissertation (Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002)

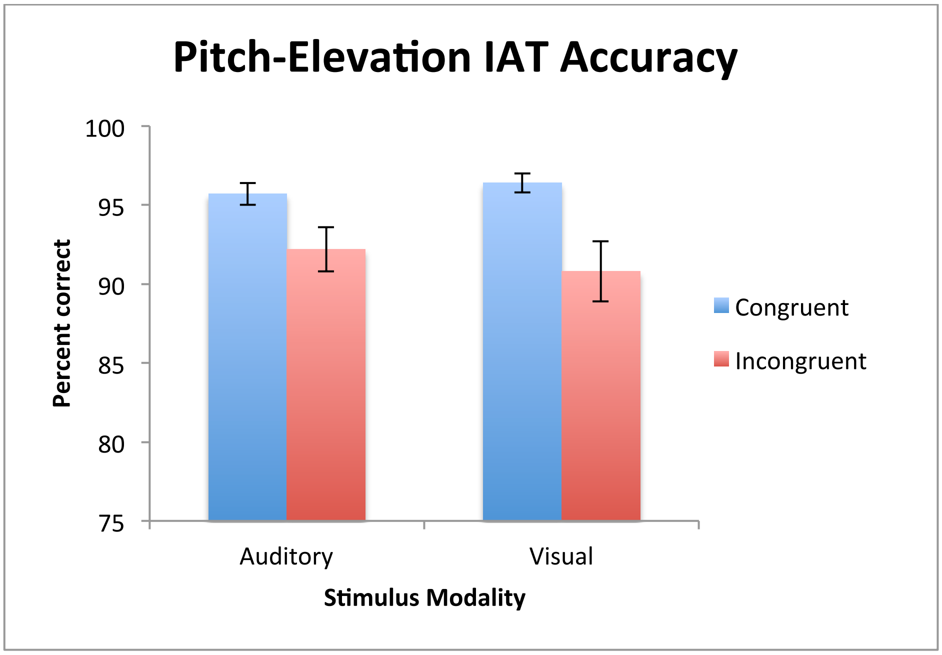


Figure 10. Mean accuracy on pitch-elevation IAT by stimulus modality and congruency of key mapping.

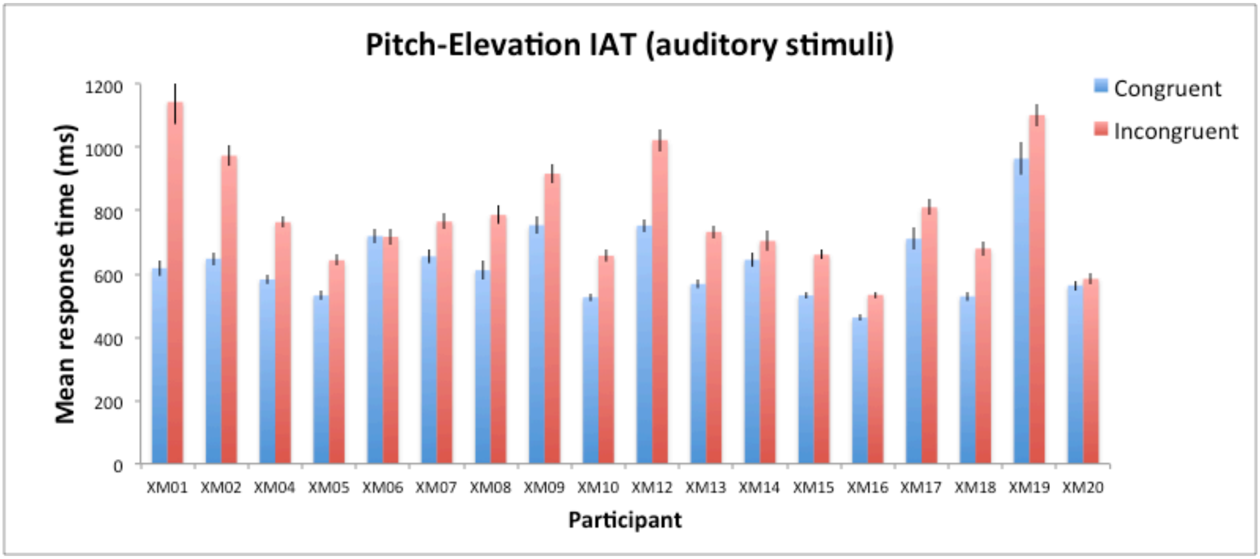


Figure 11. Mean response time to auditory stimuli for each participant on pitch-elevation IAT by congruency of key mapping. Seventeen of the eighteen subjects responded more quickly in the congruent conditions.

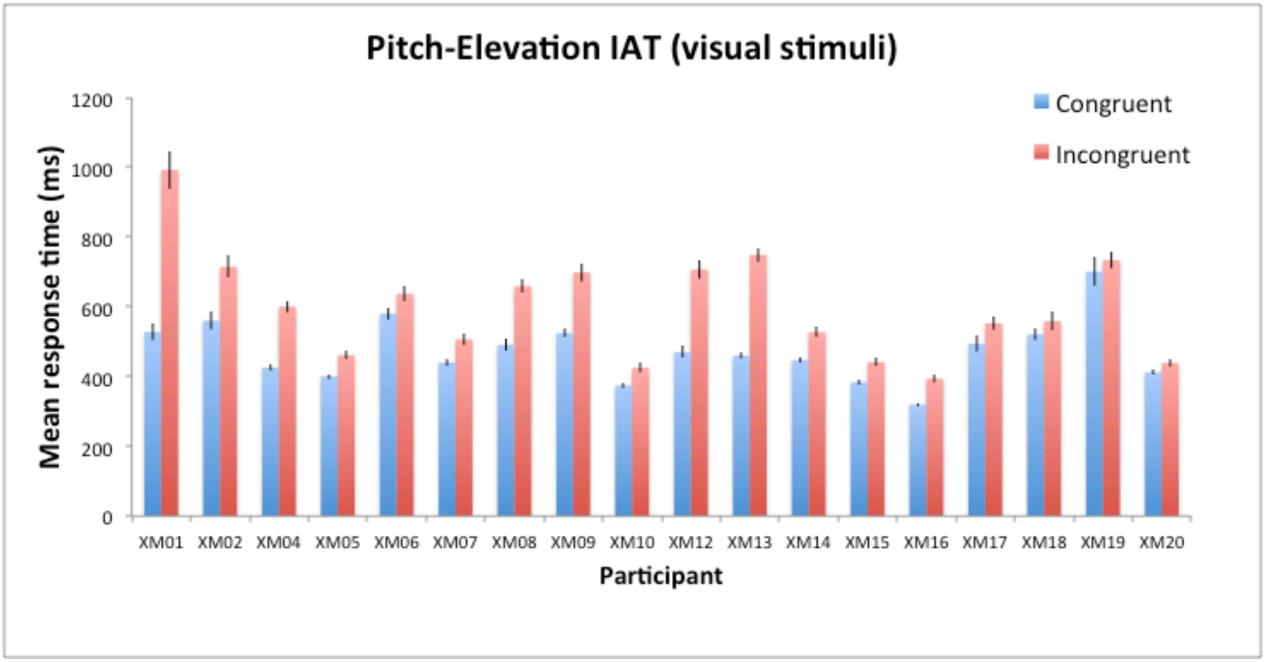


Figure 12. Mean response time to visual stimuli for each participant on pitch-elevation IAT by congruency of key mapping. All eighteen subjects responded more quickly in the congruent conditions.

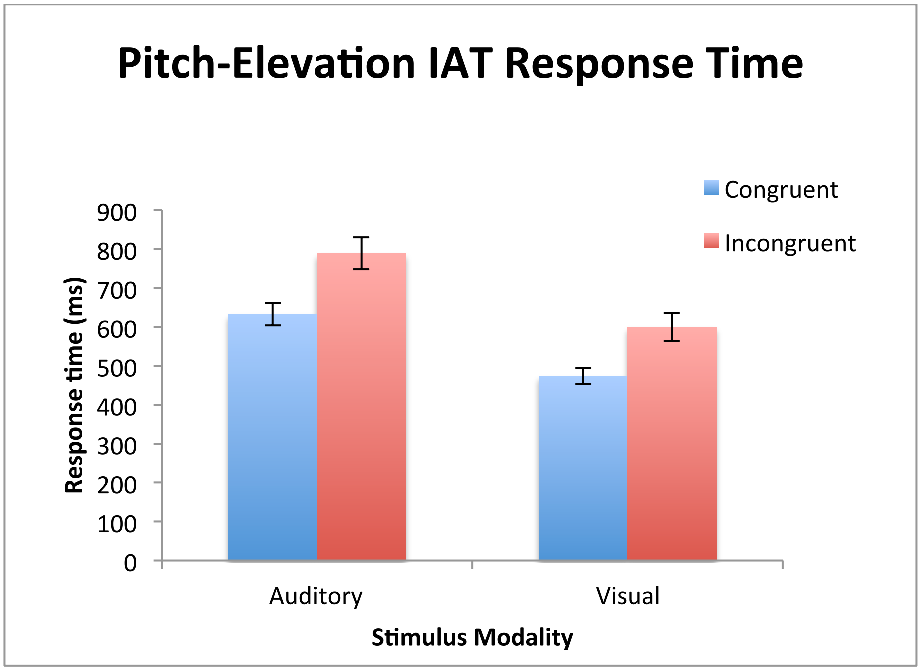


Figure 13. Mean RTs on the pitch-elevation IAT by stimulus modality and congruency of key mapping

Imaging

Localizer tasks.

Multisensory integration. The synchronous>asynchronous contrast within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels) showed active clusters in the right superior occipital gyrus (SOG) and the left parieto-occipital fissure (POF) extending into the posterior cingulate gyrus. We also examined the reverse contrast (asynchronous>synchronous), which revealed three active clusters, one extending from the right AG to the anterior IPS, one extending from the right middle frontal gyrus (MFG) into the inferior frontal gyrus (IFG) and inferior frontal sulcus (IFS), and a left hemisphere activation in the IPS.

The synchronous>asynchronous contrast did not show synchrony-selective activity in many areas implicated by previous research on sensory integration. Notably absent from the regions highlighted by our synchrony localizer were classic sensory integration areas: STS, STG, MTG, and IT (Beauchamp, 2005a; Beauchamp, Lee, et al., 2004; Calvert et al., 2000). This could be due to differences in stimuli and/or task demands of the present experiment (see discussion section for further). An exception to this was the site in the right SOG which showed a similar selectivity for synchrony>asynchrony as had been found nearby in previous study by Stevenson et al. (2010). In addition, of the foci that showed a stronger response for the asynchronous condition of the localizer (asynchronous>synchronous), the right frontal cluster was near a site implicated in processing both audiovisual asynchrony and incongruency in a meta-analysis by Erickson et al. (2014).

Magnitude. The contrast of magnitude > control within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 7 voxels) produced several foci of activity, all in the right hemisphere. Several foci were located along the right IPS and adjacent segments of the supramarginal gyrus (SMG). Previous research implicates the IPS as a major hub of the magnitude system (Eger et al., 2003; Mock et al. 2018; Pinel et al., 2004; Piazza et al., 2004, 2007). Another focus of activity appeared in the right middle occipital gyrus (MOG) near sites that have been found to respond to stimulus magnitude, showing adaptation effects (Piazza et al 2007) and involvement in subitizing visual objects (Sathian et al. 1999) (see discussion).

Semantic. The contrast of complete sentences>non-words within the cortical mask (voxel-wise threshold p < .001, cluster corrected p < .05, cluster threshold 8 voxels) produced large clusters throughout the canonical language system, including the pars triangularis of the IFG in the left hemisphere, and bilateral activations along the STS extending to the STG. This system largely matches that reported by Fedorenko et al. (2010, 2011), whose localizer was adapted for this study. Previous research by Hein et al. (2007) indicates that the STS and STG are recruited during semantically-mediated integration of auditory and visual stimuli, but not for arbitrary couplings of unfamiliar audiovisual stimuli (see discussion). The reverse contrast of non-words>complete sentences within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels) produced activations bilaterally at several sites including the cingulate gyrus and cingulate sulcus, the superior frontal gyrus (SFG), and the middle frontal gyrus (MFG) extending to the inferior frontal sulcus (IFS). In the right hemisphere there were additional activations in MFG and SFG, and the angular gyrus (AG) as well as an extensive activation from the inferior frontal gyrus (IFG) through the IFS to the MFG. In the left hemisphere, there were additional activations in the IFG extending to the lateral orbital gyrus, in the parieto-occipital fissure (POF) extending to the superior parietal gyrus (SPG), the supramarginal gyrus (SMG) extending to the AG, and a large region extending from the posterior through the mid-insula into the pars opercularis of the IFG. Although the semantic localizer has been employed in several studies, to my knowledge, findings for the reverse contrast (non-words>complete sentences) have not been previously reported.

The data produced from these three localizers were used to generate BOLD activity maps for use in further analyses and inspection. We went on to compare activity generated by these functional localizer tasks to the multisensory congruency effects produced by the main experimental task (described below) to assess overlap in BOLD responses in these regions of interest (ROIs) for the different sets of stimuli used in the experiment.

Pitch-elevation task. To identify voxels sensitive to the congruency of our audiovisual stimuli, we first conducted a univariate analysis of the blood-oxygen-level dependent (BOLD) signal to compare patterns of activity for congruent and incongruent pairings of the multisensory pitch-elevation stimuli. We first performed a contrast analysis to identify a basic congruency effect, voxels showing a stronger BOLD response for congruent trials compared to incongruent trials (C>I). To mitigate the influence of changing conditions from trial to trial, a second contrast examined congruent trials preceded by congruent (CC) versus activity for incongruent trials preceded by incongruent (II). We went on to compare the distributions of congruency-related BOLD activity with activity for three functional localizers (described above).

Multisensory Congruency effects.

Basic congruency effect. A contrast of congruent trials versus incongruent trials (voxel-wise threshold of p < .001, correcting for multiple comparisons) did not reveal any areas with significant differences in BOLD activity. This finding raised the question as to whether activity related to stimulus congruency during a given trial could be obfuscated by task demands. Successful performance on the one-back same/different task required that the participant maintain the previous trial in working memory for comparison with the present trial in order to correctly determine whether it was the same or different. When back-to-back trials are of different congruency conditions, either congruent preceded by incongruent (IC) or incongruent preceded by congruent (CI), both congruent and incongruent representations must be maintained. Perhaps maintaining and comparing trials of different congruency conditions resulted in contaminated BOLD activity profiles for the two trial conditions. To address these concerns we conducted a series of follow-up analyses.*Congruency effect for trials preceded by trials of like congruency condition (CC>II effect)*.

In an effort to reduce the effect of extra-trial factors, isolate congruency-related activity, and focus the analysis on the conditions of interest, a second analysis examined only trials that were preceded by a trial of the same congruency condition. Specifically, this analysis contrasted mean BOLD activation for congruent trials that were preceded by a congruent trial (CC) versus incongruent trials that were preceded by an incongruent trial (II). This CC>II contrast within the cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$, cluster threshold 7 voxels) revealed six distinct clusters of significant activity (Table 2; Fig 14): 1) bilateral foci in the anterior insula; 2) three foci in the inferior frontal gyrus (IFG)-bilaterally in the frontal operculum and anteriorly in the right hemisphere, and in the pars opercularis in the left hemisphere; and 3) a focus of activity in the right mid-intraparietal sulcus (mid-IPS)/angular gyrus (AG). Thus, the CC>II contrast shows that when back-to-back trials were of the same condition, there was a congruency effect. This analysis does not rule out a modulatory effect of the preceding trial, it simply eliminates the possible confounding effects of heterogeneous back-to-back trial types on the congruency effect. While there was a CC>II congruency effect in the imaging data, there was no corresponding congruency effect in the behavioral data for the same subset of trials (see in-scanner behavioral results).

I also examined the contrast of trials preceded by a different congruency condition (IC>CI), testing whether there was a congruency effect within this subset of trials. This contrast did not produce any significant activity. This difference between the CC>II and IC>CI contrasts suggests that any extant congruency effect may be washed out by changing conditions and/or task demands in the case of mixed back-to-back trial types.

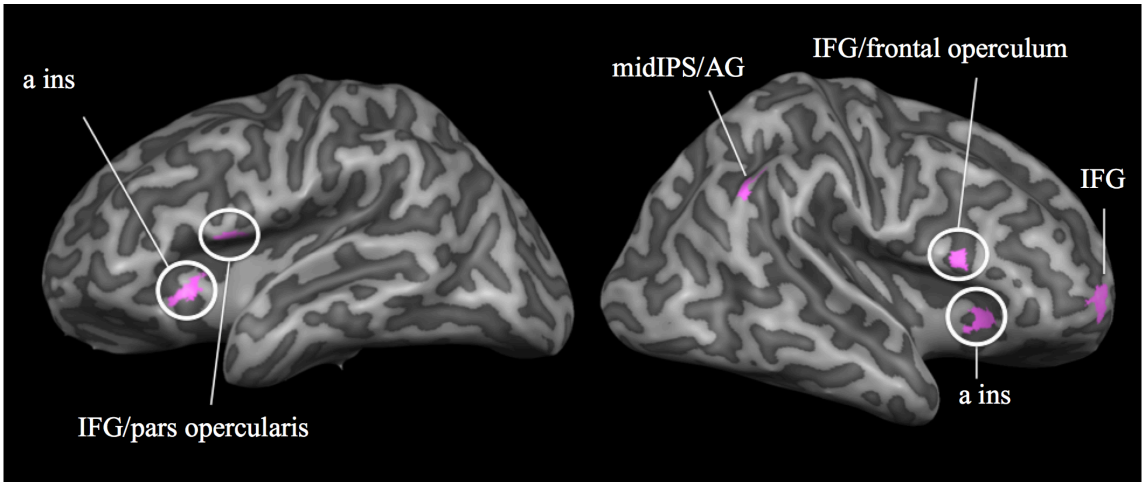


Figure 14. Multisensory Pitch-Elevation Congruency effects. CC>II shown in magenta. Circled regions indicate areas where congruency activity did not interact with congruency of previous trial.

Interaction of immediate trial congruency with previous trial congruency. While the CC>II contrast unambiguously reveals voxels that are sensitive to congruency of the pitch-elevation correspondence, it does not rule out effects of the previous trial. To further probe the relationship of trial congruency with the congruency of the preceding trial, and test for interaction of congruency conditions of the preceding trial and immediate trial, we examined the interaction contrast (CC>II)>(IC>CI). This interaction contrast identified voxels that showed different response profiles for a given trial type depending on congruency of the preceding trial. We reasoned that any voxels showing greater activity for back-to-back trials of different congruency conditions (e.g., an incongruent trial followed by congruent trial) compared to back-to-back trials of *like* conditions (e.g., a congruent trial followed by another congruent trial) might reflect modulatory effects of the previous trial. In contrast, voxels showing the CC>II effect, but not the interaction effect, would be responding more to congruency of the immediate trial and are not significantly modulated by previous trial condition (see analysis below).

There was a significant interaction effect in several regions: In the right hemisphere, there was an effect in the anterior inferior frontal gyrus (IFG) and middle frontal gyrus (MFG). The Angular gyrus

(AG) showed an interaction effect bilaterally, extending to mid-IPS in the right hemisphere. Additional foci in the left hemisphere included activations in the lateral orbital gyrus and the Supramarginal gyrus (SMG)(Fig. 14). Of the areas exhibiting the interaction effect, the right AG/midIPS and the right anterior IFG had areas of overlap with sites showing the basic (CC>II) congruency effect (Table 2; regions of overlap are uncircled in the activity map shown in Fig. 14). The finding that activity in these regions is modulated by the congruency of both the immediate trial and the previous trial makes it difficult to interpret the contributions of these regions in congruency processing. However, previous research can provide insight into possible contributions of these regions to both congruency processing and task-related processing (see Discussion).

Several foci showed the interaction effect but not the CC>II congruency effect, indicating regions modulated by previous trial, but not showing a basic congruency effect. This included parts of the left AG, SMG, and orbital gyrus, and the right MFG. Given the present contrast, we believe activity in these loci is likely to reflect some aspect of changing task demands, including maintenance of information and comparison to previous trials of different conditions (and possibly differing degrees of perceptual salience of the two trial types). We discuss foci where the interaction effect overlapped with the congruency effect in the Discussion section.

Regions modulated by congruency of immediate trial but not previous trial (i.e. 'pure congruency effect'). To address the concern that the one-back task may have introduced confounding effects of extra-trial context (i.e., the congruency of the preceding trial), we conducted a more restricted analysis (CC>II) and not (CC>II)>(IC>CI). This analysis isolated voxels that were sensitive to congruency condition of the immediate trial (show a CC>II effect) but were *not* modulated by condition of the previous trial (did not show an (CC>II)>(IC>CI) effect). We interpreted these remaining regions as showing a true pitch-elevation congruency effect, with activity reliably modulated by the multisensory congruency of the immediate trial and were not influenced by the context of the previous trial condition or other cross-trial demands of the task. Areas that showed the CC>II congruency effect, but which *did not overlap* with the interaction effect were observed bilaterally in the opercular inferior frontal gyrus (IFG), in the anterior insula (AI) on the left, and mid insula on the right, and the right frontal eye fields (FEF) (these surviving 'pure congruency' regions circled in Figs. 14 & 15).

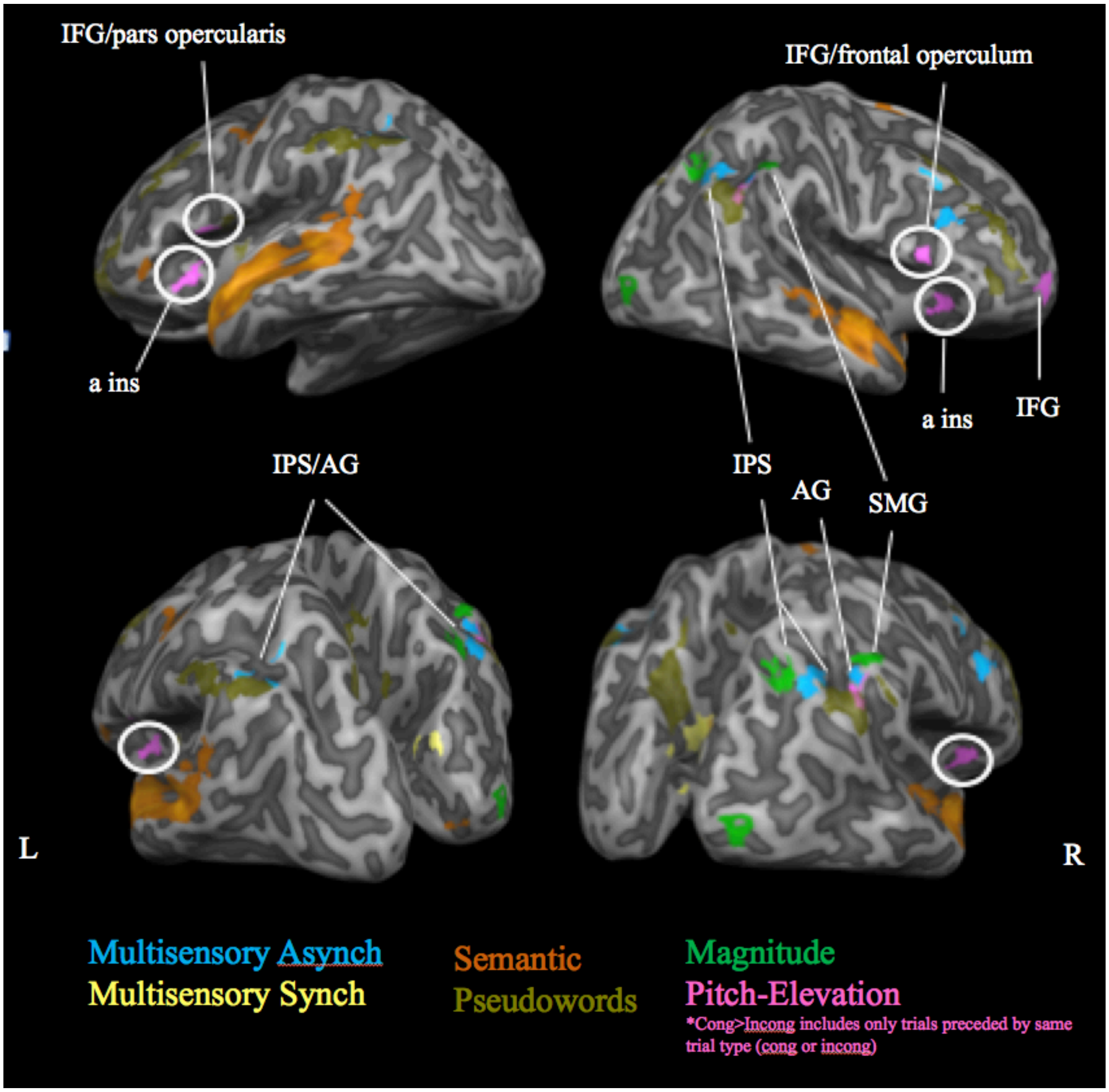


Figure 15. Overlaps of Pitch-Elevation congruency effect (pink) and functional localizers. Abbreviations a anterior; **AG** angular gyrus; **IFG** inferior frontal gyrus; **ins** insula; **IPS** intraparietal sulcus; **SMG** supramarginal gyrus. Multisensory localizer: asynchronous>synchronous (blue), synchronous>asynchronous (yellow). Semantic localizer: Semantic>Non-words (orange), non-words>semantic (olive). Magnitude localizer: magnitude>control (green).

Table 2. Sites of activity for CC-II effect. The CC-II contrast produced six distinct clusters. Plotted are hotspots for each cluster and notes about overlap with the functional localizers and the interaction effect. *Two clusters overlapped with the BOLD activity for the interaction effect. The remaining four clusters are considered to reflect a 'pure congruency' effect of immediate trial.

Significant Clusters for CC-II Effect					Overlap of CC-II effect with functional localizers					
REGION	x	y	z	tmax	MS(Async)	MS(Sync)	Sem	Sem ctrl	Mag	Interaxn
R antr IFG*	33	56	4	6.042	no	no	no	no	no	yes*
R IFG/frontal operclm	45	11	13	5.55	no	no	no	no	no	no
R mid insula	39	-1	7	5.791	no	no	no	no	no	no
R midIPS/AG*	39	-52	37	5.456	near (5mm)	no	no	yes	no	yes*
L IFG/pars opercularis	-45	8	4	5.059	no	no	no	yes	no	no
L antr insula	-30	17	1	6.357	no	no	no	no	no	no

Overlap of congruency effect with localizers. To test a priori hypotheses about possible mechanisms underlying the pitch-elevation mapping, we compared the distribution of the multisensory congruency effects to the neural systems functionally defined by our three localizer tasks. At a stringent voxel-wise threshold of *p*< .001 for all maps, we examined the extent to which BOLD activity for congruent pairings overlapped with magnitude-, semantic-, and multisensory systems respectively. This strict thresholding would reduce the risk of false positives and thus any overlaps would be compelling support for involvement of that mechanism in the pitch-elevation congruency mapping. However, absence of overlaps does not allow us to definitively rule out any mechanism.

Multisensory integration. The clusters of activity produced by the Synchronous>Asynchronous contrast did not overlap with the CC>II effect. For the reverse contrast, (Asynchronous>Synchronous) an asynchrony-selective region in the right IPS/AG was near (5 mm) activity for the CC>II congruency effect and was overlapping with the previous trial interaction effect.

Magnitude. The magnitude map did not overlap with the pitch-elevation congruency effects.

Semantic. Activity for the complete sentences > non-word contrast did not overlap with any of the congruency effects. However, the reverse contrast of non-words>complete sentences did overlap with the activation for the CC>II contrast. There were areas of overlap with the non-word condition of the semantic localizer (non-words>sentences) in a portion of the right AG/mid-IPS and in the left IFG/pars opercularis) and mid-insula (see discussion for further).

Discussion

In the present study, we sought to establish the neural underpinnings of the crossmodal mapping of pitch and visuospatial elevation. In the discussion below, we first discuss and interpret the various contrasts of congruency activity, considering known functions of the regions revealed in these different analyses. We then go on to evaluate the extent to which results from the present study provide evidence for the mechanisms hypothesized a priori to be a basis for the pitch-elevation mapping.

'Pure congruency' regions modulated by congruency of immediate trial but not previous trial: (CC>II) not overlapping with (CC>II)>(IC>CI)

This analysis revealed regions that were modulated by the congruency of the immediate trial but not the previous trial, and thus may reflect processing of stimulus attributes (e.g. congruency) rather than task-related effects influenced by previous trials (as in the interaction effect).

We found bilateral activity in the IFG, a region with well-documented multisensory response properties (Calvert et al., 2000; Downar, Crawley, Mikulis, & Davis, 2002; Hein et al., 2007; Venezia et al., 2016), and which has been found to be engaged by a range of audiovisual stimuli, including speech (Calvert et al., 2000), familiar environmental stimuli, and arbitrary, untrained parings of novel stimuli (Hein et al., 2007). These studies indicate the region is sensitive to temporal coincidence of audiovisual stimuli and may support integration of incoming sensory signals. Some regions of the IFG exhibit more selective multisensory responses, indicating it is not simply responding to simultaneous auditory + visual stimulation). For example, Calvert et al. (2000) found regions showing depressed activity (a sub-additive BOLD effect) for incongruently aligned/mismatched auditory and visual speech compared to when either stimulus was presented unimodally. Correspondingly, the CC>II activations we find in proximal areas showed reduced BOLD activity for incongruent couplings and relatively greater activity for congruent couplings of pitch and elevation stimuli. The finding that BOLD response in these proximal portions of the IFG is suppressed or reduced by misaligned auditory and visual speech stimuli as well as the incongruently paired pitch-elevation stimuli in the present study is evidence for a possible role in multisensory binding systems of these regions in the pitch-elevation mapping- a possibility that warrants further study.

In contrast to these suppression effects, others have found incongruent couplings to increase the BOLD signal in parts of the IFG. Hein et al. (2007) examined the role of familiarity in processing audiovisual couplings of semantic stimuli. They found a portion of right inferior frontal gyrus (IFG) more active for incongruent audiovisual couplings of familiar stimuli (images and sounds of animals), as compared to AV couplings of artificial stimuli. The bold response for these incongruent pairings of animal stimuli was also greater than the responses for either the auditory or visual stimuli presented unimodally, supporting a possible role in sensory integration for this region. This region sensitive to incongruency of these familiar environmental stimuli was near (8 mm) one of our right IFG foci of the pure congruency effect. Moreover, Venezia et al (2016) found that the left IFG responded preferentially to visually or audiovisually presented speech (spoken syllables), although it was modulated by audio-only speech to a lesser extent. This site was within 3mm of our CC>II effect (given that the site listed by Venezia et al. was the center of mass in a cluster of 1028 voxels, the two effects almost certainly overlap). Taken together, these findings are consistent with a role for the IFG in aligning/matching incoming sensory signals based on temporal and other cues and representing multisensory congruency of a variety of stimuli including complex environmental stimuli, speech, and the simpler stimuli such as tones and shapes from our study.

Substantial previous research implicates both the right IFG and the anterior insula as major hubs in the ventral attentional network and several studies now suggest the network (originally conceptualized as a mainly *visual* attention system) involves auditory and other sensory information (Corbetta, Patel, & Shulman, 2008; Downar et al., 2002; Eckert et al., 2009; Macaluso et al., 2002). Activity across the ventral attention network has been shown to be driven by behaviorally relevant environmental stimuli, with irrelevant stimuli suppressing the network, even when these stimuli are relatively salient. Researchers have interpreted this response profile as reflecting attentional control processes such as gating/filtering of task-irrelevant sensory information (Corbetta et al., 2008; Downar et al., 2002; Indovina & Macaluso, 2007). Research on the ventral attentional network reports evidence for a system that is largely right-lateralized (Corbetta & Shulman, 2002; Eckert et al., 2009; Lenartowicz, Verbruggen, Logan, & Poldrack, 2011), however several teams have reported recruitment of regions in the left hemisphere (in addition to the right) during the types of tasks employed to study the ventral attention system (Indovina & Macaluso, 2007; Seeley et al., 2007). Indovina and Macaluso (2007) find evidence for bilateral involvement of the AG, IFG, and insula in task-relevant processing and gating of irrelevant (but salient) sensory information. The bilateral insular activity they report overlapped with our pure congruency effect in the right hemisphere and was 3mm from the activation site in the left hemisphere. Another active site they identified was located within 10mm of our congruency activity in right IFG/frontal operculum. The correspondences between foci for our congruency effect and the active sites reported by Indovina and Macaluso, suggest our pitch-elevation task could involve a common system. Across several studies, the IFG and AI are implicated as part of a distributed system for flexibly negotiating complex task demands (Cai, Ryali, Chen, Li, & Menon, 2014; Ghahremani, Rastogi, & Lam, 2015). A slew of neuroimaging studies have shown strong functional connectivity between the IFG and anterior insula (AI), and point to their potential role in mediating salience and visual attention and response inhibition to achieve task goals. While several teams have posited a role for the IFG in a network involved in attentional detection of salient sensory signals in the environment (Chen et al., 2015; Corbetta et al., 2008; Corbetta & Shulman, 2002), its specific role within this network has been a subject of much debate. Some have suggested that rather than reflecting salience processing per se, the activity in the IFG may actually reflect inhibitory control and response-inhibition that must be exerted in order to *override* salient (but task-irrelevant) signals and ensure adequate attentional resources for less salient (but task-relevant) sensory information (Cai, Ryali, Chen, Li, & Menon, 2014). To disambiguate the functional contributions of the IFG and AI, Cai et al. (2014) analyzed functional connectivity between regions expected to have a role in mediating salience and task demands while individuals performed a stop task. Their results suggest distinct networks, with the rAI more involved in detection of behaviorally salient events, and the rIFC playing more of a role in exerting inhibitory control (Cai et al., 2014; Ghahremani, Rastogi, & Lam, 2015; Lenartowicz et al., 2011). Supporting this model, Aron and colleagues (2004, 2014) have argued that role of the rIFC in many of these tasks is primarily one of inhibitory control or overriding prepotent responses when task demands require it as in go-no go (GNG) and Stop-signal task (SST) paradigms (Aron, Fletcher, Bullmore, Sahakian, & Robbins, 2003; Hampshire, Chamberlain, Monti, Duncan, & Owen, 2010; Menon, Adleman, White, Glover, & Reiss, 2001). This interpretation could make sense with respect to our findings if the CC trials are more perceptually salient than II trials. If CC trials are more salient than II trials, they may trigger a stronger inhibitory control response from the IFC (Aron, Robbins, & Poldrack, 2014). Critically, the audiovisual congruence of our stimuli is orthogonal to the one-back task. By this model, activity in the AI could reflect the greater relative salience, whereas activity in the IFG could reflect the greater degree of inhibitory control required to override stimulus salience and to perform the task accurately.

We found the pitch-elevation congruency (CC>II) effect in bilateral foci of the anterior insulae near sites implicated in studies on multisensory processing (Calvert et al., 1997; Sestieri et al., 2006; Willems et al., 2009) and as important hubs in the ventral attention system (Seeley et al., 2007), and task-related control (Hampshire et al., 2010). A number of studies have sought to isolate the contributions of the anterior insula in the ventral attention network and other cognitive systems. Seeley et al. (2007) employed both event-related fMRI and resting state functional connectivity analyses, and identified systems involved in salience and executive control processes. Although the systems they isolated were largely distinct, they identified the anterior insula as being a node in both salience and executive control networks, with the most extensive area of overlap focused approximately 4mm from the left AI site of our CC>II BOLD effect. In another study, Hampshire et al. (2010) reported bilateral foci near the insular CC>II activity (4 mm from our left insular site, and 12mm from our right site) to be involved in a response-inhibition task. This overlap of the two systems suggests that the anterior insula functionally links systems for salience processing and attentional and control, consistent with its proposed role in mediating stimulus salience relevant for task performance (Eckert et al., 2009; Seeley et al., 2007). Several teams have posited the AI as a critical hub for multisensory attention and salience processing (Cai et al., 2014; Chen et al., 2015; Downar et al., 2002; Ghahremani et al., 2015) and have noted its specific involvement in audiovisual correspondences of meaningful environmental stimuli (Corbetta et al., 2008; Sestieri et al., 2006). Further, the rAI appears to represent multisensory salience in an integrated manner (rather than encoding salience of the different senses separately), generating a common control signal for novel audiovisual stimuli in a multisensory attention task (Chen et al., 2015). Among the foci with these response properties, Chen and colleagues identified a site 6mm from our pure CC>II effect in the right AI, and another area 13 mm from our activation in the rIFG/FO. While the precise role of the anterior insulae in the pitch-elevation mapping remains speculative, the fact that the region is sensitive to the crossmodal congruency of our stimuli supports the idea that pitch and elevation may be aligned and associated in much the same way as meaningful environmental stimuli known to engage the AI. The finding that multisensory salience and multisensory congruency of our pitch-elevation stimuli recruit nearby regions of the ventral attention network corroborates an account in which pitch and visual elevation are associated through systems for binding environmental stimuli. Several studies implicate the AI as an important center for aligning or matching information from different sensory channels. Calvert et al. (1997) identify regions more responsive to lipreading unfamiliar pseudospeech in contrast to lipreading familiar words (participants' task was simply counting number of mouth movements) at bilateral sites very close to our own insular activations (4mm left, 9mm right). The authors interpret this finding as reflecting an increased demand on phonological processing as they interpreted the unfamiliar facial movements. Although the stimuli in this experiment were strictly visual, the contrast offers insight into the potential contributions of the insulae in multisensory systems. This finding indicates that the anterior insula is bilaterally sensitive to complex spatiotemporal information, and this representation is sufficiently rich that it distinguishes between familiar (lipreading actual words) and novel (lipreading pseudowords) stimuli. Extending on this work, Fernandez et al. (2015) find that the rAI is recruited during a task presenting incongruently coupled audiovisual speech stimuli. Thus within the same region or at least very close proximity, we find fine-grained spatiotemporal encoding and clear sensitivity to the statistics of experience-based visuospatial information. The finding that nearby foci are sensitive to the pitch-elevation correspondence in our study suggests a common system is recruited in processing these different kinds of audiovisual correspondences. Perhaps multisensory congruency and response-inhibition tap into a common system involved in analyzing converging sensory inputs and weighing conflicting evidence to guide behavioral responses.

Interestingly, Willems et al. (2009) report multisensory congruency effects in bilateral foci of the anterior insula very close to those we find for the pure CC>II effect (within 6mm on the left, 10mm on the right; Willems, Özyürek, & Hagoort, 2009). In this study, Willems et al. tested several couplings of audiovisual stimuli. Auditory stimuli were meaningful speech, which was coupled congruently or incongruently with videos of a) co-speech gesture and b) pantomime. They found that the anterior insula was sensitive to incongruent versus congruent couplings of speech and *pantomime but not* for speech and *gesture.* They argued that speech and pantomime are semantically redundant – with position and movements is directly iconic representing a referent, whereas gesture is not always redundant and often involves body movement/position that convey meaning in a more abstract, relational and context-dependent manner. Their interpretation of this difference is a compelling one, which could have strong implications for our own findings. They postulate that speech and pantomime are both independently readily interpretable, and represent different channels for conveying the same meaning, and which are mapped onto a common semantic representation (Willems et al., 2009). This means that combining pantomime and speech redundantly encodes information and that mismatching the two channels can be truly semantically incongruent. In contrast, they argue, speech and gesture interact in a semantically non-redundant manner – a particular gesture interacts with the speech (and vice versa) to arrive at a new meaning (which would be unavailable from either stream independently). For this reason, Willems et al. and others suggest that these channels should not be considered to redundantly contribute to meaning. Given this interpretation, it may not make sense to consider a particular co-speech gesture to be congruent or incongruent with speech, although certainly there are relatively more and less comprehensible couplings. Thus, Willems et al. demonstrated an interesting case in which co-occurrence of meaningful AV stimuli was not sufficient for engaging the AI, whereas co-occurrence of *semantically redundant* AV stimuli did engage the AI. Another study sought to tease apart the neural systems involved in different types of audiovisual matching. Sestieri et al. (2006) used localization and recognition tasks to compare the systems involved in spatial and semantic matching across auditory and visual modalities. In their experiment auditory and visual stimuli (static images and sounds of animals, musical instruments etc.) were presented concurrently, and could be presented on either the right or left side. In this way an AV coupling could be either semantically congruent (if the image depicted something that would emit the sound played), or spatially congruent (if the sound originated from a source on the same side that the image is presented). By varying these two types of crossmodal congruency they identified systems that were differentially sensitive to one or the other of these correspondences. Among the areas they identified were two foci in very close proximity to our own bilateral anterior insula hotspots (within 2mm on the left, 4mm on the right). These foci showed greater BOLD activation for the semantically congruent AV stimulus pairings as compared to the semantically incongruent couplings. They also identified a region of the right STS, which was modulated by spatial congruency of the stimulus couplings, showing greater BOLD signal for congruent versus incongruent AV couplings, this activity was not close to any of our congruency effects. They interpret these findings as evidence that the systems underlying spatial and semantic congruency are at least partially distinct, and the distribution of these systems is more or less consistent with the dorsal and ventral processing streams originally posited for visual attention. By pitting semantic congruency and spatiotemporal congruency against one another, the present study adds to the findings of Sestieri et al. and provides an opportunity to compare two systems that could support the pitch-elevation mapping. We find strong bilateral co-localization of our congruency effect with the semantic congruency effect reported by Sestieri et al., but did not observe active sites near their rSTS hotspot that was responsive to spatial congruency. It is worth noting that the insular activity reported by Sestieri et al was evoked in response to environmental sounds coupled with static images. This means the congruency response in this region reflects *semantic* congruency of the AV stimuli rather than responding to temporal frequency properties of the incoming signals (Sestieri et al., 2006). There is also evidence to suggest that the AI responds more strongly to audiovisual couplings that are semantically incongruent compared to AV couplings of stimuli that are novel and unfamiliar. For example, a study by Hein et al. (2007) identified several sites that were activated more strongly by incongruent couplings of familiar AV stimuli (animal images and sounds) compared to couplings of novel, untrained stimulus pairings. One of these sites was approximately 14mm from our congruency effect in the left AI. Couplings that were not semantically related did not evoke similar responses in AI, so in a sense the AI appears to respond more strongly for semantic *mismatch* than a *non*-match (e.g. a coupling of meaningless stimuli). This is in contrast to patterns in the IFG, which responds to both novel- and familiar pairings (Downar et al., 2002;

Hein et al., 2007; Lewis, 2010). Together, these findings implicate the AI in learned, experience-based semantic associations across visual and auditory modalities. The colocalization of the pure congruency effect from our study with sites from this previous research leads us to conjecture that the pitch-elevation mapping may indeed reflect the correlated coupling of pitch and elevation in our perceptual experience in the environment (Parise et al., 2014).

Regions exhibiting both a congruency effect and interaction effect : (CC>II) overlapping with (CC>II)>(IC>CI)

Of the areas exhibiting the congruency effect (CC>II), the right AG/midIPS and the right anterior IFG overlapped with sites showing the interaction effect (CC>II)>(IC>CI). These overlaps reveal regions that are modulated by the congruency of previous trial and thus do not solely reflect basic processing of stimulus attributes (e.g. congruency) of the immediate trial. The finding that these regions exhibit both a congruency effect and an interaction effect makes it difficult to ascribe particular functional contributions to these regions. However, reviewing previous research can provide insight into possible contributions of these regions in both pitch-elevation congruency processing and demands related to our task.

Several studies have implicated the right AG in task-related attentional control such as response inhibition (Wager, Sylvester, Lacey, Nee, Franklin, & Jonides, 2005) and conflict resolution (Nee, Wager, & Jonides, 2007) during go-no go tasks, which could involve similar task monitoring and conflict processing to our same/different 1-back task. The right AG appears to be important for modulating stimulus processing according to task demands (Indovina & Macaluso, 2007). In addition to being sensitive to immediate task demands, the angular gyrus has been shown to be sensitive to recent task history (Taylor, Muggleton, Kalla, Walsh, & Eimer, 2011). Taylor et al. (2011) identified a portion of the right AG close to the congruency activation from the present study that was important for reorienting attention or attentional control to stimuli depending on task history and stimulus salience. The right AG also appears to be involved visuospatial attention (Cattaneo et al., 2009). Previous research using transcranial magnetic stimulation (TMS) has found that stimulation to the right angular gyrus can modulate visual pop-out, binding and related multisensory processes (Kamke, Vieth, Cottrell, & Mattingley, 2012; Taylor et al., 2011). In one such study, Taylor et al. (2011) report that TMS to right AG modulated perceptual pop-out in a visual search task. The CC>II effect we observed in the rAG was confirmed to be in close proximity (7mm) to the site that was targeted with their TMS. Another TMS study targeting adjacent areas of the right temporo-parietal region found that stimulation to the right AG knocked out a somatosensory cueing effect (for somatosensory targets) while stimulating an adjacent portion of the right SMG decreased crossmodal cueing, making somatosensory cues less effective in cueing visual targets (Chambers, Payne, & Mattingley, 2007). Kamke et al. (2012) found that stimulation to right AG reduced individuals' susceptibility to a sound-induced flash illusion (where hearing multiple beeps gives the illusion that the visual stimulus flashes with the beeps), leading participants to more accurately report that the visual stimulus was uninterrupted. At least in this case, TMS stimulation to the right AG reduced cross-sensory interactions and led to a more veridical representation of multisensory events (Kamke et al., 2012). Although the regions stimulated in these studies were spatially distinct, taken together they indicate a role for AG/IPS in crossmodal mappings and congruency. The involvement of the right AG in overriding salient sensory signals and selectivity for task-relevant information could explain why it is recruited across trials in our task. Performing our task accurately requires comparison of the stimuli in an immediate trial with a previous trial. If multisensory congruency affects salience or the relative fluency of processing these stimuli, this system may underlie the ability to ignore or override salience of the stimuli in order to focus on the same/different task at hand. Supporting this interpretation, several attentional cueing studies have reported greater activity in frontoparietal networks (including the AG) for invalidly cued trials compared to validly cued trials (Arrington, Carr, Mayer, & Rao, 2000; Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000; Indovina & Macaluso, 2007; Macaluso, Frith, & Driver, 2002). The preceding trial in our one-back task could serve much the same function as the cues in these other studies, such that when participants make a same/different judgment about a trial, they are engaging in similar processing to when they engage with a target that was either validly- or invalidly cued. Together, these previous findings indicate that the AG is functionally involved in crossmodal cueing of spatial attention, binding, salience, and maintenance of recent task history, all of which could contribute to the interaction effects (influence of the preceding trial) in the AG, as well as differences in processing at transitions between congruent and incongruent trials (producing the observed differences between trials of different congruency conditions).

The other region where we found both CC>II and interaction effects was a focus in the right anterior IFG (BA 10). Although the functions of this most rostral portion of the prefrontal cortex (PFC) are not thoroughly understood, the region is broadly implicated in studies on memory recall, monitoring, and executive function and is believed to be the part of PFC most heavily connected to supramodal regions of cortex (Ramnani & Owen, 2004). Ramnani and Owen (2004) reviewed an array of human and animal research and posited a specific role in coordination of top-down cognitive operations (Ramnani & Owen, 2004). The region appears to be involved in maintaining information when responses are delayed. For example, Farooqui, Mitchell, Thompson, & Duncan (2012) identified a region (in the vicinity of our right anterior inferior frontal activation) that showed strongest BOLD response when an expected target finally appears (in contrast to intervening foils), signaling the completion of an ongoing task episode spanning several smaller tasks (Farooqui et al., 2012). If these interpretations are correct, the cross-trial modulation we find in the anterior IFG could reflect this monitoring and task-fulfillment effect as subjects maintain and coordinate representations of multisensory stimuli in order to make a same/different response after a sustained delay of 7 seconds from the previous trial. These known functions can help explain our finding that the area is recruited by both the CC>II contrast and the interaction effect, perhaps possibly related to attentional episodes and task fulfillment effects.

Limitations related to task and congruency effects. Our original rationale for using a 1-back task was to ensure that subjects would attend to relevant stimulus properties without posing an undue task burden. One concern with this task is that maintaining a representation of a previous stimulus in working memory for comparison to present stimulus, as is necessary to successfully perform the task, could obfuscate a simple congruency effect. The nature of the one-back ‘same/different’ task could also make transitions between congruent and incongruent trials differentially difficult. To account for possible cross-trial contamination effects, we examined the CC>II effect, which although still potentially subject to modulation from previous trial, we took as a relatively more homogeneous sample since the preceding trial was of the same experimental condition as the immediate trial. However, this contrast raises concerns about repetition suppression and adaptation. One concern is that presenting back-to-back of trials of the same condition could suppress the magnitude of the BOLD response, and so the CC>II contrast could fail to detect areas that are sensitive to multisensory congruency but which show an adaptation effect. A region could be sensitive to stimulus congruency but if it exhibits a strong effect of adaptation/repetition suppression, it may not be detected in a CC or II analysis, which focuses only on trials that were preceded by a trial of the same condition, making them subject to suppression. Our ISI of 7 seconds is fairly long as event-related fMRI designs go, but even substantially longer intervals can produce a suppressed hemodynamic response resulting in underpowered contrasts (Dale, 1999; Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000). Another concern is that performing the same/different task may rely on a common or interconnected system to that involved in processing intersensory congruency. It is possible that these systems are sensitive to congruence over time (same/different) and across modality (crossmodally-congruent/incongruent). Indeed, the finding that the interaction effect overlaps with the congruency effect could support the possibility that making a judgment as to whether stimuli are *the same across a temporal gap* (as in the 1-back same-different task) may involve some of the same systems as analyzing whether multisensory stimuli are *congruent across sensory modalities*. As previously discussed, the greater fronto-parietal ventral attention network comprises overlapping systems for multisensory attention, salience, and task-related control, all of which could figure into the signature BOLD activity observed for our task. Several teams have proposed that systems for multisensory attention and task-related control are functionally interconnected. In a 2003 paper, and in parallel to the contemporaneous research on the ventral attention system, Melara and Algom posited a selective-attention network for balancing attentional constraints and striking a balance between avoiding distraction, and focusing on conspicuous elements and patterns in the environment. They argued that

Stroop-like effects that arise when processing conflicting sensory stimuli are a product of this system, which modulates attentional focus and mediates stimulus salience in the context of task demands (Melara & Algom, 2003). This theory meshes well with the functional neuroimaging literature on the interactions between networks for salience and task-level control in the ventral attention system and could account for the overlap we find between congruency and interaction effects.

Localizers

We had several a priori predictions about systems that might support the pitch-elevation correspondence. In particular, we hypothesized possible roles for systems supporting multisensory integration, magnitude, or semantic processing. We tested the involvement of these three putative functional systems by first isolating them using functional localizer tasks, then assessing overlap with congruency-related activity from our primary task. Comparing the distribution of congruency-related effects with the activity profiles produced by three functional localizers allowed us to evaluate the involvement of these systems in the crossmodal mapping of these dimensions. Although the congruency-related activity was not found to overlap with the three main systems we had predicted to play a role, there were areas of overlap with the control conditions of the semantic and multisensory localizers. Considering the known functions of these loci of overlap can help in building an account for the functional basis of the observed congruency effect and functional basis for overlaps.

Potential basis in semantic processing. The semantic localizer in the present study identified an extensive system that was more active in processing the complete sentences compared to the nonword strings, broadly replicating findings reported by Fedorenko et al. (2013). There were no overlaps between the semantic system and the pitch-elevation congruency effect, so at least on the basis of this language-based localizer, the present study fails to find evidence for semantic mediation of the pitch-elevation correspondence. However, there are a number of ways semantic representations could underlie crossmodal mappings, so the lack of overlap with this particular localizer does not conclusively rule out a semantic basis for the pitch-elevation mapping. Another possible semantic basis for crossmodal mappings is on the basis of object-level or event-level representations or other encoding of information related to multisensory signals originating from, or corresponding to, a common source in the environment (e.g. an image of a dog and the sound of a dog's bark). This possibility somewhat blurs the boundaries with the multisensory account detailed in the next section, and the two accounts are not mutually exclusive. For example, a semantic object representation could be directly grounded in the senses, with conceptual information encoded across the same sensory systems used to perceive an object or event (Barsalou, 1999; Barsalou et al., 2005; Feldman and Narayanan, 2004; Glenberg and Kaschak, 2002; Rizzolatti and Craighero, 2004). In this case, we could expect fairly basic integrative systems involved in aligning incoming sensory signals to play a role in encoding semantic congruency. For the purpose of this discussion, we elaborate on this multisensory/semantic account in the following 'multisensory' section, although it could reasonably be considered a hybrid of the semantic *and* multisensory accounts as delineated here. In contrast to this most sensory-based semantic account, several researchers have suggested that semantic object and action representations could also be encoded in a more supramodal or abstract format (Damasio and Tranel, 1993; Lambon Ralph et al., 2010; Binder, 2016). Interestingly, several of the sites exhibiting a congruency effect in the present study are in or near regions that integrate diverse inputs from sensory and motor systems to form supramodal representations related to objects and actions in space. Extensive previous research has implicated the AG and IPS as heteromodal convergence zones involved in integrating sensory information into more abstract spatial and conceptual representations (Damasio, 1989; Binder and Desai, 2011; Fernandino et al., 2015). Over the course of several studies in monkeys, Rizzolatti and colleagues (Luppino & Rizzolatti, 2000; Rizzolatti, Fogassi, & Gallese, 1997; Rizzolatti, Luppino, & Matelli, 1998).found evidence that these parietal regions are functionally linked to frontal regions, forming a parieto-frontal network for representing and executing actions and interacting with objects in space. The frontal component of the network is comprised of several premotor regions near the pure congruency focus from the present study. These regions integrate information from visual, somatosensory, and motor systems into a supramodal system for representing and controlling goal-directed actions in space. Building on this earlier work, Graziano et al. (1999) found that the region encodes both auditory and visual information about object locations in peripersonal space. Although our current understanding of peripersonal space and action-related representations in these regions is based largely on research in monkeys, it is widely accepted that homologous systems exist in humans (Rizzolatti & Sinigaglia, 2010). In light of these previous findings, and given our ever-expanding understanding of parieto-frontal network, it appears plausible that the pattern of congruency-related activity in the present study could reflect a supramodal type of semantic representation related to the location of objects in space.

An independent line of research on semantic representation provides additional evidence that the AG is an important hub for processing a wide range of meaningful stimuli (Binder et al., 2009; Binder & Desai, 2011; Damasio et al., 2004; Turkeltaub et al., 2002). Binder et al. (2009) conducted a meta-analysis of 17 neuroimaging studies on semantic processing found that the angular gyri exhibit a stronger response (bilaterally) when tasks required concrete semantic knowledge compared to abstract semantic knowledge. The authors argued that this could be due to a difference in the degree to which these different kinds of semantic knowledge are perceptually encoded, with concrete semantic knowledge learned through direct perceptual experience and encoded in the senses, and abstract semantic knowledge tending to be verbally learned and encoded (Binder et al., 2009). Thus, the congruency activity we observe in the rAG could reflect perceptually-based semantic knowledge related to pitch and elevation.

Given these independent veins of research implicating the rAG in semantic processing, and the similarity in profiles of activity in the putative parietal-frontal action network, it appears plausible that these systems could be a basis for the congruency effect observed in our study. However, it is worth noting that the right AG was one of the foci where the congruency effect overlapped with the interaction effect. While this does not negate a potential role for the site in congruency processing, it does indicate a response profile that is more complicated than we would expect of a region simply encoding or responding to stimulus congruency (see above for discussion of task-related activity and the interaction effect).

Overall, our evidence appears more consistent with a supramodal multisensory attentional basis of the pitch-elevation mapping, with congruency effects appearing in relatively supramodal or abstract systems, rather than sensory areas. The distribution of the pitch-elevation congruency effect appears more consistent with a supramodal semantic account, than one based in primary sensory or basic multisensory integrative regions.

Potential basis in phonological processing. The activity produced by the nonword contrast (non-words>sentences) of the semantic localizer could reflect the greater demand on the phonological system involved in reading the unfamiliar non-word stimuli compared to the complete sentences condition of the localizer, which also involves syntactic and semantic processing (Fedorenko et al., 2010). This activity could also reflect the greater effort required in reading the unfamiliar non-words compared to complete sentences (Price et al., 1996). This is in line with Fedorenko et al. (2013) and Duncan (2013), who have interpreted the non-word contrast in a variation of this localizer task as reflecting greater processing difficulty. Consistent with this interpretation, there was substantial overlap between the activations on this (non-word>complete sentences) contrast (including in the IFG, IFS, and MFG bilaterally) and the frontoparietal multiple-demand system described in Duncan (2013). This possibility is examined in greater detail below.

The congruency effect overlapped with the non-word contrast of the semantic localizer (non-words>sentences) in two clusters: one in the left IFG/pars opercularis extending into the mid-insula and the other in a portion of the right AG/mid-IPS. Previous studies have reported on proximal regions exhibiting stronger responses during processing of non-words compared to real words or sentences in various formats (during reading, silent lipreading, and listening to spoken language), which could reflect phonological processing (Fedorenko et al., 2010). For example, Binder et al. (2005) identified a site in the left IFG (proximal to the site of overlap in the present study), which exhibited a stronger response when individuals were presented with written non-words relative to concrete words. In another study, Calvert et al. (1997) examined neural activity as they presented subjects silent videos of a person pronouncing unfamiliar pseudowords and familiar words. They found bilateral insular activity for silent lipreading of a person pronouncing unfamiliar pseudowords compared to familiar words including a site near our left IFG/pars opercularis activation, and suggest this activation of the insula during pseudospeech could reflect increased demand on phonological processing. Given these previous findings, the activity in the present study could be consistent with a phonological basis for activation at this site. Even more generally, the common activity found for both non-word localizer and congruency effect could reflect the mapping between visual and auditory modalities involved in both reading unfamiliar non-words and in processing the coupled pitch-elevation stimuli..

The cluster in the right AG/mid-IPS also exhibited the inter-trial interaction effect (sensitive to congruency type of back-to-back trials) making it difficult to attribute specific functions to the site. The finding that this site was sensitive to multisensory congruency previous trial congruency, and the nonword>word contrast supports the proposal that the area is recruited during demanding task conditions. As previously noted, the non-word contrast of the semantic localizer co-localized extensively with the domain-general frontoparietal multiple-demand system (Duncan, 2013; Fedorenko et. al, 2013) including the right AG/mid-IPS site. The involvement of these systems makes sense given Duncan’s proposal that the multiple demand system is important for regulating and maintaining attentional episodes during complex cognitive tasks as would be required in our one-back task (see above for discussion of interaction and task effects).

Potential basis in multisensory integration. The audiovisual synchrony localizer was administered as a means of isolating a system that is selective for temporal synchrony of multisensory signals. We reasoned that such a system would have the essential components necessary for registering audio-visual contingencies and could potentially represent multisensory scene statistics of the environment. This is because synchronous signals very often originate from a common source in the external world. However, the audiovisual temporal synchrony localizer did not reveal synchrony-selective activity in areas most consistently implicated by previous studies manipulating audiovisual synchrony (e.g. STS, STG, and MTG; Beauchamp, 2005a; Beauchamp, Lee, et al., 2004; Calvert et al., 2000). One likely reason for these differences in activation is that previous neuroimaging studies examining audiovisual synchrony have typically employed meaningful 'environmental' stimuli such as images and sounds of animals, tool use, and human speech. In contrast, our localizer employed basic stimuli (circles, squares, and tones) in effort to capture synchrony-selectivity in a most basic form, and to avoid involving semantic representations to the extent possible (as those were meant to be captured by a separate localizer). So while the localizer in this study identified a system sensitive to AV synchrony of our particular stimuli, it failed to capture some of the systems known to respond to audiovisual temporal synchrony for other types of stimuli. It is also worth noting that concurrent experience of sensory signals is just one of many ways we learn about the multisensory contingencies of world. In the future, additional localizers could further examine the role of multisensory learning and attention in the pitch-elevation mapping.

Ultimately, it was the reverse contrast (Asynchronous>Synchronous) of the multisensory localizer that revealed more activity adjacent to foci for other activations of interest. An asynchrony-selective region in the right IPS/AG was near (5 mm) the right parietal activity for the CC-II congruency effect and was overlapping with the previous trial interaction effect. It is possible that the asynchrony-sensitive foci are part of a multisensory integration system as we had originally targeted with the localizer contrast, and that the asynchronous presentation of auditory and visual stimuli poses a greater processing burden on this system than the synchronous condition. This is in line with recent multisensory integration research indicating increased recruitment of IPS (and an extended frontoparietal network) during processing of degraded multisensory stimuli relative to clear multisensory stimuli (which engaged a distinct more ventrolateral network; Regenbogen et al, 2017). Within this putative frontoparietal network, the AG is thought to integrate sensory information into a supramodal spatial representation, and the presence of both congruency and asynchrony effects in close proximity reinforces the possibility that the region is sensitive to multiple forms of congruency (i.e. spatial, temporal, and multisensory). Thus, the congruency activity in the right parietal site could reflect multisensory integrative processing (including segregation of asynchronous inputs). However, another explanation is that the asynchronous condition of the oddball task used in the localizer may have required more cognitive effort than the synchronous condition (independent of sensory integrative processes). This could account for the proximity between the asynchrony-preferring site and the interaction effect, with both contrasts representing the condition posing a relatively greater cognitive load across a temporal gap in the respective tasks.

Potential basis in magnitude processing. The Magnitude localizer contrast (magnitude>control) replicated previous research and identified classic magnitude regions in and around the rIPS (Eger et al., 2003; Pinel et al., 2004; Piazza et al., 2004, 2007; Sathian et al., 1999; Sokolowski et al., 2017) but these foci did not overlap with the pitch-elevation congruency effect. Although it did not overlap with our magnitude localizer, a site showing our pitch-elevation congruency effect was relatively near (17mm, Euclidean CoG distance) the IPS activity produced by the magnitude localizer. While pitch-elevation congruency activity did not co-localize with the magnitude localizer from the present study, it was close to areas implicated by previous research in magnitude processing. Previous research has revealed that spatial aspects of magnitude-related processing often engage the rAG in the vicinity of one of our CC-II congruency sites (as well as the interaction effect). For example, Cattaneo et al. (2009) found that TMS to the rAG disrupted the priming of spatial attention using small and large numbers, which they take as support for a spatial representation of the mental number line. Arsalidou & Taylor (2011) find additional evidence that the rAG links magnitude and spatial representations and is involved in visuospatial attention during arithmetic calculation tasks.

I had hypothesized that the magnitude system could underlie the cross-dimensional pitch-elevation mapping, with lower acoustic frequencies aligned with lower spatial elevation and relatively greater acoustic frequencies being aligned with greater spatial elevation. However, it is worth noting that researchers have noted inconsistencies in pitch-elevation mappings. While high pitch is mapped to greater values on some stimulus dimensions (brightness, height, intensity; Eitan & Timmers, 2010; Marks, 1987), it is mapped to ‘less’ when it comes to amount, size, or quantity of a stimulus (Eitan et al., 2014). Thus, one possibility is that the magnitude system does, in fact, encode cross-sensory mappings, but that the stimuliweconsidered congruent pairings are actually treated by the magnitude system as incongruent. Another possibility is that auditory pitch is not neurally encoded in terms of magnitude or a quantitative less-more values. Pitch is considered by many to be a *metathetic* dimension, with different stimulus values along the perceptual dimension producing percepts that differ in terms of qualitative category (e.g. different musical notes or colors) in contrast to a prothetic dimension wherein different stimulus values represent *quantitative* differences of a percept (e.g., loudness of a sound or brightness of an image). It may be that crossmodal mappings involving auditory pitch are not encoded by the magnitude system, as such a mapping would likely depend on percepts being prothetically encoded and mapped onto a dimension in terms of their more-less relationship.

Another consideration is thatwemay be capturing only a subsystem of the greater magnitude system. For example, recent research has found that judging ‘how many’ (a precise count of individuated objects) versus ‘how much’ (an approximate judgment of amount or extent) produces dissociable activity profiles in the parietal lobe, including the classic magnitude system (Lecce, Walsh, Didino, & Cappelletti, 2015), although other teams find considerable overlap between different types of Magnitude processing (e.g. luminance and numerosity; Pinel et al. 2004). My magnitude localizer was based on an estimation task in which participants were asked judge whether there were “more black or white shapes in the image”. The brief stimulus presentation and speeded pace of the task were intended to make it impossible to count the number of items in the array, so that participants would have to make their judgments based on a gist representation of approximate quantity rather than precise number. This notwithstanding, it is possible that our task engaged a representation more related to *individuated number*, and failed to capture the greater approximate number system (ANS) thought to be a component of the more general systems involved in representing amount or magnitude (Lourenco & Longo, 2010; Walsh, 2003). This possibility is supported to some extent by the finding that the magnitude localizer included a site in the MOG previously implicated in subitizing of visual objects (Sathian et al, 1999). Whileweacknowledge this as a limitation of the localizer,weare satisfied that it did, indeed, capture at least a subsystem of the classic magnitude system that has been robustly reproduced in neuroimaging literature. Future research could use additional localizers to distinguish systems for number versus amount or other related things such as individuated amount versus amount or extent of an unindividuated mass.

Limitations and future directions

The absence of a significant BOLD effect for the *all congruent* > *all incongruent* contrast requires further investigation. These null findings could indicate that we are simply not isolating the congruency phenomenon effectively. It has been shown that multisensory processing benefits are most robust when the stimuli are relatively weak or degraded (Calvert, 2001). The stimuli in our experiment may have been too strong and unambiguous to recruit some multisensory integrative systems. Another possibility is that the system(s) supporting the pitch-elevation mapping are similarly responsive for both congruent and incongruent stimulus conditions. If this is the case, our congruent>incongruent contrast would not show differences in BOLD activity so we would fail capture the involvement of the systems.

There was also a lack of behavioral differences on the different conditions of the in-scanner task. Subjects' behavioral performance on the in-scanner 1-back task did not differ in the two congruency conditions, with no difference in accuracy (both conditions 95%) and minimal differences in mean response times (2 milliseconds). Because participants did not have to engage in congruency processing to complete the task, and because we do not find differences in response times for the two conditions, we cannot be certain that they engaged in congruency-related processing at all. Extensive experimentation testing pitch and elevation stimuli suggests that processing of these perceptual dimensions is tightly linked (Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002). Garner characterized cross-dimensional correspondences as integral or automatic, meaning that attention cannot be focused on one of the dimensions without intrusion from the other (Garner, 1976; Garner & Felfoldy, 1970). For this reason, we expected that simply attending and responding to the stimuli (as is required to successfully performing the one-back task) would produce a behavioral congruency effect. Without substantive behavioral differences, any functional differences can be difficult to interpret and raises the concern that our task may not be isolating the phenomena of interest or could be engaging underlying systems to a *similar extent* in congruent and incongruent conditions.

In contrast to the in-scanner results, however, the implicit association task conducted outside the scanner produced robust congruency effects, with all subjects responding faster in the congruent condition. The difference in behavioral performance inside and outside of the scanner suggests that the tasks differ in the extent to which they involve congruency processing relative to other task demands. A limitation of both the in-scanner and out-of-scanner behavioral tasks is that they did not include a baseline condition (e.g., one stimulus modality held constant while the other varied; Martino & Marks, 2000; Melara & O'Brien, 1987; Patching & Quinlan, 2002). This limits our ability to interpret response time differences observed for the two congruency conditions. When we find faster response times for congruently paired stimuli than for incongruent, does this represent a *facilitation* effect for congruent condition, an *interference* effect for the incongruent condition, or a combination of the two effects? The design of the present experiment does not allow us to distinguish between these effects, but for the purposes of this experiment, we were satisfied to find any congruency-related effects, which helped us identify systems sensitive to multisensory congruency. Future neuroimaging studies could work to disentangle these related phenomena for the pitch-elevation mapping as well as other basic perceptual correspondences.

Task limitations. The task used in this study poses its own limitations. One concern with the one-back task is that it requires processing of information outside the window of the immediate trial introducing inter-trial heterogeneity beyond the intended modulation by the multisensory conditions of interest. To perform the task, individuals must maintain the previous trial in memory for comparison with the present trial in order to make the same/different required in the task. A related concern is whether the task is truly orthogonal to the congruency processing of interest. It is likely that maintaining one trial in memory while evaluating an immediate trial engages common neural systems (Barsalou 1999, Gallese & Lakoff, 2005). This being the case, remembering a trial in one condition while being presented with a trial of the other condition may confound the BOLD activity signatures (a subject may be simultaneously representing congruent and incongruent stimuli). To avoid task-related contamination of any extant congruency-related effects, future studies should employ multiple types of tasks to isolate the phenomena and systems of interest. For example, a forced-choice about identity of the immediate stimulus would focus individuals on the stimulus at hand, and eliminate the need to represent information about previous trials. Our congruency-related activity aligns closely with systems implicated in stimulus salience and task-related response inhibition, so in addition to parsing out the role of task-demands, future research should consider how and why congruency of pitch-elevation stimuli affects saliency.

Localizer limitations. We did not find overlaps between activity for the CC>II congruency effect and any of the functional systems we had posited on an a priori basis. While this could be due to a lack of involvement of these systems in the pitch-elevation correspondence, it does not *rule out* their involvement. It may also be that the localizers failed to capture important aspects of the systems of interest. For example, the audiovisual temporal synchrony localizer did not identify the major AV synchrony-selective temporal regions identified by previous research. So while the localizers in this study served to isolate functional systems sensitive to our particular stimuli and contrasts, they may have failed to capture components of the targeted system that does, in fact, underlie the pitch-elevation mapping. Future research could employ additional localizers to capture subsystems and to further elucidate the involvement of the various systems in the pitch-elevation correspondence. For example, to examine the role of statistical learning of cross-sensory correspondences an additional localizer could identify regions that exhibit modality invariant response properties, but do not respond to temporal synchrony of multisensory stimuli.

Analysis limitations. It is possible that the stringent statistical thresholds we applied in analysis and the group averaging wash out weak effects. There are several strategies we could employ to improve statistical power of our analyses. In the present study, our functional localizers were defined by grouped data (averaged from all participants). While this was an effective approach for constraining the search space for the present study (which we could readily probe for congruency effects), we are liable to miss interesting individual differences (Fedorenko & Kanwisher, 2009, 2011). Averaging functional activity across subjects washes out unique spatial distributions of the functional systems. In a subsequent analysis, the respective activity profiles for each subject could be examined. Examining the unique activity profiles produced by the stimuli in our main experiment could allow us to capture individual differences in cross-sensory mapping that are unavailable by interrogation of the group data (Fedorenko & Kanwisher, 2009, 2011). We could further compare the contributions of audiovisual binding, magnitude, and semantic systems by examining voxelwise response patterns in the ROIs identified using our localizer tasks. Methods such as Multi-voxel Pattern Analysis (MVPA) offer another approach for capturing and differentiating the nuanced patterns in the BOLD response to our different experimental conditions. Multivariate analyses could be applied to detect extant effects not captured by the univariate analyses. For example, if we included unimodal stimuli, we could train a classifier using the data from trials in one modality and then test its accuracy at classifying data from the other, corresponding modality. A classifier trained on activity for auditory pitch trials could be tested on activity for elevation trials and vice versa. We could then compare classification accuracy score in the different regions of interest, either places where we find a congruency effect, or the regions identified by our functional localizers for each subject. Comparing classifier performance in different ROIs would help gauge the extent to which the different regions encode corresponding auditory and visual signals similarly. While this type of voxelwise pattern analysis is capable of detecting heterogeneous responses over a large swath of voxels, it is still limited to the resolution of one functional voxel (here ~3mm^3). The BOLD response of a single voxel represents the averaged responses of millions of neurons, so with current fMRI methods, we are not able to resolve neuronal responses in more fine-grained spatial resolution. Because it is based an average of activity across a large population, fMRI cannot effectively capture heterogeneous response profiles of neurons within a voxel, so we are liable to miss any areas that are sensitive to stimulus congruency but respond in a heterogeneous manner (Calvert et al., 2000).

Theory limitations. The present study provides preliminary evidence for the neural instantiation of a cross-sensory mapping of pitch and elevation, a mapping that is widely shared in western populations. Recent research has produced conflicting reports as to whether pitch-elevation mappings are found in infants, and it remains an outstanding question whether we are prewired or predisposed to make these mappings, or whether they arise purely as a result of experience in the world. Although this question is outside the scope of the present study, a similar paradigm could be employed in developmental or cross-cultural studies. By 1) examining variability of mappings over the course of development, and 2) examining the extent to which a mapping is shared cross-culturally, we gain insight into the nature-nurture question.

Conclusion

The aim of the present study was to elucidate the neural underpinnings of the pitch-elevation correspondence, a well-studied but poorly understood phenomenon. Although it is common across diverse cultures and appears to be present early in child development, its neural basis has, until now, remained unclear. Using fMRI, we found several brain regions sensitive to the crossmodal congruency of pitch-elevation stimuli. We then used functional localizers to examine and contrast several possible hypotheses on a within-subject basis. The congruency effect did not overlap with our functional localizers (a priori hypotheses) except for the non-words > sentences localizer (in the right angular gyrus). Although congruency-related activity did not overlap with the three main systems we had predicted to play a role, the patterning of activity provides critical insight into the neural basis of the pitch-elevation mapping. Our results indicate involvement of inferior frontal gyrus and anterior insula in congruency-related processing of the pitch-elevation correspondence. Substantial previous research implicates these regions as hubs of a frontoparietal system involved in multisensory attention, control and salience processing. In addition, several teams have reported a role for a focused region of the anterior insulae in learned, environmentally-based audiovisual associations, and the results of the present study suggest the mapping of pitch and elevation may be supported in much the same way as semantic association of meaningful images and sounds. The finding that our CC>II effect is in close proximity to areas consistently reported in previous studies to be sensitive to multisensory semantic congruency (i.e., that found for corresponding environmental stimuli) generally supports a hybrid semantically-based multisensory attentional account for the pitch-elevation mapping. The findings from the present study expand on previous research in a few key respects. Although a number of functional neuroimaging studies have examined multisensory congruency, to our knowledge this is the first study to examine the cross-sensory mapping of pitch and elevation. By mapping sensitivity to the pitch-elevation correspondence in the human brain, we take an important step in bolstering our understanding and developing an account for the phenomenon.

Chapter 3. Crossmodal association of pseudowords and object shape

A central question in the study of language and cognition is how the sounds in language encode and convey information. The relationship between the word form and meaning is believed to be largely arbitrary, that is, the sounds that comprise words bear no inherent relationship to the objects, actions, and events that they represent (de Saussure, 1916; Gasser, 2004; Hockett, 1960). In an arbitrary system, words come to represent meanings by learned associations, and in principle, any combination of the finite inventory of sounds in a given language should be just as fit as any other for representing a meaning. Yet despite evidence that language is largely arbitrary, research continues to discover exceptions, and demonstrates that certain types of sounds in language are more likely than others to be associated with particular meanings (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016; Davis, 1961b; McCormick, Kim, List, & Nygaard, 2015; J.B. Nuckolls, 1999; Ohala, 1994). The reasons for these patterns and the neural mechanisms underlying them, for the most part, have yet to be explained.

Why are certain speech sounds associated with particular meanings and in what kinds of words do we find systematic sound-to-meaning mappings? In the case of onomatopoeia, when a word sounds like what it means, the relationship between sound and meaning is relatively clear. Such words are composed of linguistic segments that in some way mimic the sound described by the word (e.g. *crush*, *flop*, *bark*). But the motivation for systematic mappings found in other types of words is less obvious. For example, many languages have a distinctive class of words (termed *ideophones*, *mimetics,* or *expressives* in different lines of research) that uses sounds to depict a range of sensory imagery (Dingemanse, 2011a; Dingemanse & Majid, 2010; Kilian-Hatz, 2001; Kita, 1997). There is wide variation across languages in the richness of such vocabulary and the extent to which it is used (J.B. Nuckolls, 2003; Perniss, Thompson, & Vigliocco, 2010), and it is believed that in some languages such terms may number in the thousands (Dingemanse & Majid, 2010; Kakehi, Tamori, & Schourup, 1996). Similar to onomatopoeia, ideophones (also known as 'mimetics') exhibit systematic mappings between elements of the phonological sound structure of a word and properties of the intended meaning (Vigliocco & Kita, 2006). However, in contrast to onomatopoeic words, which use sound to represent sound-related meanings, ideophones often use structural attributes of a word to depict sensory meanings outside the auditory domain. Systematic sound-to-meaning mappings have been identified in a wide range of terms describing perceptual experiences such as texture, shape, taste, quality of light, manner of motion, and even proprioceptive and emotive states (Dingemanse, 2011b; Kita, 1997; Tufvesson, 2011; Vigliocco & Kita, 2006). The Japanese language is known for its particularly rich lexicon of mimetics, with terms such as *kirakira* (flickering light), *pikapika* (a bright metallic glint of light), and *nurunuru* (slimy but firm texture of fish skin). In the West African language Siwu, food with a bland taste can be described as *buàà*, the feeling of vertigo as *ɣììì*, something fluffy could be *wùrùfùù,* and a bubbling pot could be described as *gblogblogblo*. The meanings of these words can be quite nuanced and specific, and are often difficult to express using other words. This has led experts to suggest that ideophones may be a uniquely effective means for a perceiver to share how it felt to experience something. In some way or another, these words may cause listeners to simulate perceptual aspects of the intended meaning (Bergen & Chang, 2003; Dingemanse, 2011b; Stivers, 2008).

There is evidence to suggest that sound-symbolic words may trigger meaningful representations in listeners who lack prior experience with a given language. Empirical studies have found that naïve listeners often share intuitions about what unfamiliar words ought to mean. Language studies, notably Kunihira (1971) and Nygaard et al. (2009), have found that adults are able to correctly guess meanings for sound-symbolic words from unfamiliar foreign languages at rates better than predicted by chance. To systematically examine the contributions of sounds to attributed meanings, a related line of research has employed made-up words. These studies typically ask individuals to guess what an unfamiliar word means, rate how fitting it would be as a label for a particular concept (Maglio, Rabaglia, & Feder, 2014; McCormick et al., 2015), or even to invent a novel term to describe something for which there is not an existing term in their language (Magnus, 2001). Such studies have demonstrated systematic sound-meaning mappings in terms for brightness (Newman, 1933), texture (Magnus, 2001), size (Newman, 1933; Sapir, 1929; Thompson & Estes, 2011), and even abstract meanings related to precision (Maglio et al., 2014), or qualities of pleasantness, arousal, and potency (Osgood, Suci, & Tannenbaum, 1957).

The finding that individuals correctly infer meanings of unfamiliar words at rates above chance supports the idea that sound-to-meaning mappings in language are not entirely arbitrary. Although languages vary widely in terms of their respective phonetic inventories and the specific sounds they use to mark meanings, interesting cross-linguistic patterns emerge at different levels of analysis. Individuals appear to exploit these patterns when tasked with inventing novel words to represent particular meanings (Magnus, 2001). Interestingly, when inferring word meanings, individuals appear to rely on the elements of sound structure that have been found to be mapped to meanings relatively consistently across languages (Magnus, 2001; J.B. Nuckolls, 1999; Nygaard, Cook, & Namy, 2009; Ohala, 1994). Thus it appears that mappings for these terms are based not in language-specific conventions, but rather on properties of the sound to meaning mappings that are relatively stable across languages.

The finding that certain classes of sounds appear in words with particular meanings across diverse languages, suggests that particular sounds or certain aspects of sound structure are in some way favored or more fit for representing certain meanings (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016; Davis, 1961b; Kirby, 1996; McCormick, Kim, List, & Nygaard, 2015; Nuckolls, 1999; Ohala, 1994). Researchers have suggested that the mappings we see in sound-symbolic language may be rooted in more general crossmodal systems, and that there may be natural biases in how these systems interact (Maurer, Pathman, & Mondloch, 2006; Mondloch & Maurer, 2004; Ozturk, Krehm, & Vouloumanos, 2013). For example, in a study on 2.5-3-year-old children, Mondloch and Maurer (2004) found that children reliably associated a high-pitched sound with the lighter and smaller of two balls shown (Maurer et al., 2006; Mondloch & Maurer, 2004). In light of these findings, Mondloch and Maurer suggest that humans have a natural predisposition for pitch to be crossmodally linked with domains of size and lightness. This sort of correspondence could be exploited in the service of communication, for example, by systematically using the pitch of a word to represent meanings related to size or lightness. More generally, if there are naturally motivated associations between the sounds in language and their meanings, and listeners are sensitive to these correspondences, a word could evoke representations of meaning in a listener, even without any prior experience with the word (or even the language). Non-arbitrary sound-meaning mappings like these could be particularly helpful for language learners, who could utilize this structure as a cue to meaning when tasked with learning or guessing the meaning of unfamiliar words, effectively using crossmodal mappings to bootstrap word-learning (Asano et al., 2015; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Imai & Kita, 2014; Imai, Kita, Nagumo, & Okada, 2008; Kantartzis, Kita, & Imai, 2011; Nygaard, Cook, & Namy, 2008; Ozturk et al., 2013; Revill, Namy, & Nygaard, 2018; Tzeng, Nygaard, & Namy, 2017). Berlin and O'Neill (1981) speculated that non-arbitrary labels would likely be easier to remember than arbitrary ones, possibly reducing cognitive effort involved in leaning, and recalling terms. Supporting this theory, several studies have since found evidence that children are sensitive to sound symbolism in language, that iconic words represent a disproportionate amount of children's early vocabulary (Monaghan, Shillcock, Christiansen, & Kirby, 2014; Perry, Perlman, & Lupyan, 2015; Perry, Perlman, Winter, Massaro, & Lupyan, 2017), and that they use newly acquired terms more productively if they are sound-symbolic (Imai et al., 2008; Kantartzis et al., 2011). Not only are highly iconic words learned at an earlier age than less iconic words, but adults tend to favor them when speaking with young children (Perry et al., 2015, 2017). Studies in adults provide further support for the idea that non-arbitrary language confers a benefit for the language-learner, reporting that adults learning Japanese learned sound-symbolic vocabulary words more readily than words that were not sound-symbolic, or which were sound-symbolically mismatched (Lockwood, Hagoort, & Dingemanse, 2016; Nygaard et al., 2008, 2009). Together these studies indicate that sound symbolism facilitates word learning in both children and adults, possibly by drawing upon existing crossmodal associations that are independent of language.

Word-Shape mappings: The Bouba-Kiki phenomenon

In the present study, we examine a particularly well-documented case of sound symbolism- the crossmodal mapping between auditory pseudowords and novel visual object shapes. This phenomenon was described by Wolfgang Köhler in 1929 based on a sample of native Spanish speakers in the Canary Islands. In one version of his experiment, Köhler asked subjects to match nonsensical labels *takete* and *maluma* to two novel shapes, one pointed and the other a rounded blob shape. He found that individuals showed a consistent bias in their responses, favoring *takete* as a label for the pointed/angular shape, and *maluma* for the rounded/blob shape (Köhler; 1929, 1947). Variations of this study (some using different pseudowords) have since replicated the original finding across diverse subject populations with an array of languages and cultures (Bremner et al., 2013; Davis, 1961a; but see Rogers & Ross, 1975; and Styles & Gawne, 2017 for noteworthy exceptions). A 2001 paper from Ramachandran and Hubbard sparked widespread interest in the phenomenon, which they dubbed the 'bouba-kiki effect' (Ramachandran & Hubbard, 2001). Subsequent studies have examined the developmental trajectory of the phenomenon and have found that even 4-month-old infants (Ozturk et al., 2013) and toddlers (mean age 2.5 years) (Maurer et al., 2006) exhibit a matching bias similar to that found in adults.

Despite the robustness of the bouba-kiki phenomenon, the cognitive underpinnings of this crossmodal mapping remain unclear. What is it about these words that leads people to match them to particular shapes? In order to develop hypotheses about the possible basis for these correspondences, researchers have sought to characterize how particular components of the speech signal or sound structure of the word are correlated with or matched to aspects of visual shape (Aveyard, 2011; Fort, Martin, & Peperkamp, 2014; Hamano, 1994; Magnus, 2001; Nielsen & Rendall, 2011; Ramachandran & Hubbard, 2001).

Early research on sound symbolism for shape largely focused on the contributions of the vowel sounds in these words. A host of studies have reported that closed unrounded front vowels (e.g. /i/, /e/) are associated with more pointed or jagged forms, whereas open rounded back vowels (e.g. /o/, /u/) are associated with rounded or bloblike forms (Maurer et al., 2006; Nielsen, 2011; Tarte, 1974; Tarte & Barritt, 1971). Tarte and Barritt (1971) found that participants made reliable mappings between vowel sounds and object shapes, preferring to pair /u/ with ellipses and /i/ with triangles (Tarte, 1974; Tarte & Barritt, 1971). This may be due in part to the higher frequency of the second formant (F2) in front vowels compared to back vowels, which makes them perceptually higher in pitch. Analogous findings outside the domain of language demonstrate that relatively higher-pitch sounds tend to be associated with more angular visual shapes (Marks, 1987; Parise & Spence, 2012; Walker, 2012). Maglio and colleagues (2014) found that participants favored nonsense words containing front vowels for representing conceptual *precision and intensity* (e.g. shorter durations, more acute pain) compared to nonsense words containing back vowels (Maglio et al., 2014), suggesting that the mapping extends into more abstract conceptual domains.

More recent studies confirm that consonants also influence how spoken words are mapped to visual shape (Fort et al., 2014; McCormick et al., 2015; Nielsen, 2011; Nielsen & Rendall, 2011; Ozturk et al., 2013). Across several studies, when matching pseudowords to jagged or rounded objects, individuals prefer to use unvoiced plosive or strident consonant sounds that are made by constricting or disrupting airflow (as in *takete*) for jagged, angular objects, and prefer to match more unobstructed, voiced sonorant sounds (as in *maluma*) to objects with a more rounded form (Aveyard, 2011; Fort et al., 2014; Maurer et al., 2006; McCormick et al., 2015; Nielsen & Rendall, 2011; Westbury, 2005). Two empirical studies employing rich sets of consonants, have found that words containing the sonorants /l/ and /m/ sound especially rounded, being matched to rounded shapes at high rates (Fort et al., 2014) as well as being rated both high on a Likert scale of roundedness and low on a separate scale of pointedness (McCormick et al., 2015). Nielsen and Rendall (2011) point to major differences in spectral density and attack of the strident consonants compared to sonorant consonants, and propose that the more abrupt transitions in sound structure of strident consonants would be associated with harsh or jagged forms, whereas the more gradual increases in amplitude and slowly changing structure of the sonorants would be associated with a more smooth, rounded form (Nielsen & Rendall, 2011). In line with this hypothesis, researchers studying natural language have found correlations between particular consonants and meanings related to object form or other physical properties in a number of languages. For example, Hamano (1996) examined phonemic correlates to word meaning in Japanese, and found that certain semantic contrasts are marked by phonemic featural contrasts (Hamano, 1996, as described in Kita 1997).

Hamano found that continuant consonants (which can be uttered for a sustained duration and are characterized by relatively gradual articulatory transitions) appear in words describing continuous movement, amorphous forms, whereas non-continuants (which are characterized by relatively short duration and often more sudden articulatory transitions), tended to occur in words describing abrupt movement, or disrupted surfaces. Several studies have found that unvoiced consonants are more associated with small, fine, precise meanings, whereas voiced consonants are associated with large, coarse qualities (Magnus, 2001; Thompson & Estes, 2011; Westermann, 1927, 1937 as cited in Dingemanse, 2012). Additional research has specifically examined the link between consonant voicing and meanings related to object shape. For example, in a corpus analysis of the English lexicon, Monaghan, Mattock, and Walker (2012) found that words denoting angular meanings contained more unvoiced consonants whereas words with rounded meanings were more likely to be comprised of voiced consonants (Monaghan, Mattock, & Walker, 2012). In a related line of research, Magnus (2001) examined phonemic correlates to meaning, and reported that /b/ clusters in terms with rounded meanings (e.g. *bubble* and *bulge*), whereas /p/ appears in terms describing smaller, more precise meanings (e.g. *point*). In another experiment, Magnus asked people to make up a word to describe the texture of a hedgehog and found that they were most likely to use /k/, /r/, and /p/, all of which involve disrupted airflow in their articulation. These findings demonstrate that systematic sound-to-shape correspondences are found in natural language and that language users make use of this structure (e.g. in guessing word meaning or making up a novel label for something).

In some cases, the sound-symbolic mappings outlined above are consistent with research on crossmodal associations found outside the realm of language. For example, in an implicit association task, Parise and Spence (2012) found that tones composed of square waves (which have a disrupted, noisy quality) were associated with more pointed visual objects as compared to tones composed of sinusoidal waves (which have a smoother tonal quality), which were associated with more rounded shapes. These mappings are analogous to those reported by Nielsen and Rendall (2011) and others, wherein strident consonants (p, t, k) are mapped to pointed shapes and the more mellifluous sonorant consonants (l, m, n) are mapped to rounded shapes (Fort et al., 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011). In another experiment, Parise and Spence (2012) found that high-pitched tones were associated with more acute visual angles than lower pitch tones, a mapping paralleled by vowel pitch being mapped to more pointed visual forms. These parallels between linguistic sound structure and non-linguistic auditory domains, suggests that sound-symbolic mappings in natural language may arise from more general tendencies to associate experiences across perceptual domains (Maurer & Mondloch, 2005; Maurer et al., 2006; Mondloch & Maurer, 2004; Namy & Nygaard, 2008; Parise & Spence, 2012; Spector & Maurer, 2009).

The above discussion identifies several phonemic correlates to word meaning that could account for the perceived rounded or pointedness of the pseudowords in our experiment. Together, these findings provide insight into well-studied bouba-kiki phenomenon, and allow us to make specific predictions about what sorts of words should sounds especially rounded or pointed. The pseudowords *keekay* and *lohmoh* used in this experiment consist of classes of sounds that have been implicated as sounding rounded or pointed by previous research (the stop consonant /k/ versus the sonorants /l/ and /m/, and the unrounded /i/ or /e/ versus the rounded vowel /o/) (Bremner et al., 2013; Fort et al., 2014; Köhler; W, 1929; McCormick et al., 2015; Monaghan et al., 2012; Nielsen & Rendall, 2011; Ramachandran & Hubbard, 2001). These particular words were selected because they had been rated in a previous study as sounding extremely pointed and rounded, respectively, while also scoring low on the contrasting scale (McCormick et al., 2015). While empirical studies have accomplished much in the way of elucidating what kinds of sounds are associated with roundedness and pointedness, there is little consensus among researchers as to *why* individuals make such mappings.

Mechanism. In some cases sound-symbolic words may reflect statistical regularities of multisensory experience in the natural world (Berlin, 1994; Ković, Plunkett, & Westermann, 2010; Sidhu & Pexman, 2017). For example, Ohala (1984) reported that fundamental frequency in spoken language is systematically mapped to object size (Ohala, 1984). He proposes that higher pitch sounds could be sound-symbolically linked to size because smaller objects resonate at higher frequencies, and, at least within ethnozoological class, relatively smaller animals tend to emit relatively higher-pitched sounds than corresponding larger animals (Ohala, 1984, 1994). Berlin (1994) finds some support for this theory in a study on the Huambisa language of Peru, in which names for smaller birds tended to have more high vowels (which have a higher fundamental frequency) than names for larger birds (Berlin, 1994). Kovic et al. (2010) suggested that sound-symbolic mappings may be based in a more general multisensory feature integration process in which incoming signals from the different senses are linked. The word-shape correspondence could be based in the co-occurring multisensory properties of physical objects in the environment, for example, the tendency for harder objects to break into sharper pieces, or make higher frequency sounds in collisions or when resonating (Parise & Spence, 2012; Walker et al., 2010). The present study is based on similar rationale to that described for the pitch-elevation experiment in Chapter 2. That is, if sound symbolism has its basis in multisensory processing, neural activity related to sound-symbolic processing could co-localize with activity related to multisensory integration, such as audiovisual synchrony, spatial congruency, or audiovisual identity (as described in Chapter 1). If we find that a common neural system responds to word-shape congruency as well as audiovisual synchrony, this could be taken as evidence in support of a sensory integrative model for sound-meaning mappings in language. Such a finding could be consistent with simulation-based accounts of sound symbolism, which posit that the sound structure of some words serves to simulate sensory aspects of the meaning conveyed by the word (Berlin, 1994; Kita, 1997; Janis B Nuckolls, 1996). By this account, sensory simulation provides a direct link between the symbol and the symbolized, offering a straightforward model for how language sounds could trigger multisensory conceptual representations, and thus how sound could be systematically related to sensory experiences beyond the auditory domain. Hearing particular language sounds could evoke or bring online multisensory representations related to object shape (how does a particular shape look, feel, or sound as it is experienced by a perceiver in the environment?). For example, the Japanese mimetic term *gorogoro* is used to describe a heavy object rolling across a surface as well as the rumbling sound of thunder (Kita, 1997). While this term clearly depicts *auditory* aspects of a perceptual experience, it also seems to convey information about multisensory attributes such as object shape, mass, and motion. In this manner, the word-shape correspondence could be a linguistic form of representing or referring to multisensory properties of physical objects (Parise & Spence, 2012; Walker et al., 2010).

This type of word-meaning association is a step more direct than those conceptualized in perceptually-grounded theories of language comprehension. Rather than an arbitrary word serving as a handle or pointer to conceptual representations and bringing online relevant perceptuomotor simulations, in sound-symbolic language, the word itself *produces* the simulation that brings online the intended meaning (Barsalou, 1999, 2003; Bergen, 2007; Bergen & Chang, 2003; Bergen, Lindsay, Matlock, & Narayanan, 2007; Elman, 2009; Glenberg & Robertson, 2000; Pulvermüller & Fadiga, 2010; Zwaan, 2003; Zwaan & Kaschak, 2008). More specifically, the sound structure of the word itself may partially reactivate the same sensory traces as would have been engaged in experiencing the event described, directly interfacing with the sensory representations it represents. Supporting this sensory simulation theory of sound symbolism, several studies have now found evidence that listening to sound-symbolic words recruits relatively more sensory and multisensory regions of the brain compared to more arbitrary language (Arata, Imai, Okuda, Okada, & Matsuda, 2010; Hashimoto et al., 2006; Lockwood & Dingemanse, 2015; Winter, Perlman, Perry, & Lupyan, 2017). If we find that sensory and multisensory regions exhibit sensitivity to the word-shape congruency examined in the present study, this could bolster the sensory-integrative account of sound symbolism.

Sound-to-shape mappings could reflect multisensory learning in other ways, as well. Ramachandran and Hubbard (2001) proposed that a linkage between auditory and motor representations causes the rounded vowel sounds (which are produced with a round mouth shape) to be associated with rounded forms. While it seems plausible that rounded vowels would be associated with the rounded shape of the lips during articulation, it does not provide a ready account for the association of unrounded vowels with jagged shapes (Nielsen, 2011). Others have theorized that orthographic forms are systematically related to particular speech sounds (e.g., linear spikey letters are more often used in strident consonants or unrounded vowels, whereas rounded letters are more often used to represent continuant consonants or rounded vowels) and that such mappings lead people to associate particular shapes and sounds (Reilly, Westbury, Kean, & Peelle, 2012; Westbury, 2005). While this letter shape-sound association may play a role in some populations, it cannot account for the finding that the bias is shared by individuals from non-literate cultures (Bremner et al., 2013), pre-literate infants (Ozturk et al., 2013), or congenitally blind individuals (Bottini, Barilari, & Collignon, 2019). In short, it is possible that orthographic mappings contribute to the phenomenon in literate populations, but they cannot be solely responsible.

Another possibility is that sound symbolism has its basis in a Magnitude system. Walsh (2003) proposed that dimensional attributes such as space, time, and number could be encoded by a common neural system for magnitude. According to this theory, attributes that can be represented in terms of amount, along a continuous dimensional scale (with *less* and *more* at the dimensional poles), can be encoded in a common neural format according to their respective more-less relations on a magnitude scale (Bueti & Walsh, 2009; Walsh, 2003). Since this general magnitude system was first posited, researchers have found evidence that a number of additional perceptual dimensions (e.g. loudness, luminance) may also be encoded in a domain-general magnitude system (Bueti & Walsh, 2009; Cohen Kadosh & Henik, 2006; Pinel, Piazza, Le Bihan, & Dehaene, 2004). Ahlner and Zlatev (2010) and others have suggested that meanings related to salient perceptual attributes that can be placed along a less/more gradient (e.g. dimensions of roundedness or pointedness) would be good candidates for magnitude-based sound symbolism (Ahlner & Zlatev, 2010; Nielsen & Rendall, 2012). It may be that the sound-meaning correspondences in sound-symbolic language constitute a form of magnitude mapping, whereby some aspect of the sound structure of a word (e.g. fundamental frequency, sonority, voice onset time, duration) is aligned with a corresponding meaning by virtue of their relative positioning on a scaled dimension defined by less-more relations. For example, a word describing a jagged/angular shape could use sounds with more abrupt transitions than a contrasting word for a rounded/blob shape, thus encoding information about spatial frequency of the shape within the temporal frequency of the speech sounds. If magnitude representations underlie the word-shape correspondence examine in this study, we could expect to see audiovisual congruency effects in the intraparietal sulcus (IPS) and possibly the angular gyrus (AG), two regions consistently implicated in magnitude-related processing (Eger et al., 2003; Mock et al. 2018; Pinel et al., 2004; Piazza et al., 2004, 2007).

Another possibility is that we associate these novel pseudowords and shapes because they index or map onto a common conceptual or semantic representation. One way such a mapping could manifest is that a novel pseudoword may be associated with familiar words via lexico-semantic processing. In some cases the pseudowords may sound like actual words with shape-related meanings (Sučević, Savić, Popović, Styles, & Ković, 2015). For example, *bouba* or *lohmoh* sound similar to terms with meanings related to roundedness/curvature such as 'bubble, blob, dome, lump, bulb', whereas *keekay* or *kiki* sound similar to terms connoting jagged/sharp meanings such as 'crooked, crinkle, spike, peak, crack'. If the pseudoword-shape mapping is rooted in this sort of lexico-semantic processing we could expect to see involvement of the semantic system proposed by Fedorenko and team (Fedorenko, Behr, & Kanwisher, 2011; Fedorenko, Duncan, & Kanwisher, 2013). Finding activity related to word-shape stimuli co-localized with activity produced for Fedorenko's semantic localizer would be evidence is support of this theory. Another way that pseudoword-shape mapping could be semantically-mediated is that the sounds in pseudowords may serve to simulate or otherwise bring online a supramodal conceptual representation which would in turn bring online corresponding sensory representations across multiple modalities. Lambon Ralph and colleagues have found evidence for a supramodal conceptual hub in the anterior temporal lobes, evidenced, in part, by a modality-invariant response for semantically-matched signals (e.g. auditory and visual stimuli corresponding to the same meaning/perceptual source) in different sensory channels (Lambon Ralph, 2014; Lambon Ralph, Sage, Jones, & Mayberry, 2010; Visser, Jefferies, Embleton, & Lambon Ralph, 2012). Finding anterior temporal activity for our sound symbolism stimuli could indicate a conceptually-mediated association.

The experiment detailed in Chapter 3 uses functional magnetic resonance imaging (fMRI) to examine the neural mechanisms underlying the word-shape mapping (McCormick, Lacey, Stilla, Nygaard, Sathian, 2018b). By modulating congruency of audiovisual pairings, we identify regions sensitive to sound-symbolic mappings. As described in the previous chapter, we used three functional localizer tasks to identify systems hypothesized to support sound-symbolic mapping between word and shape: magnitude, audiovisual integration, and semantic systems. We reasoned that any loci of activity common to both the word-shape task and any of these functional system(s) would indicate a likely role for these mechanisms in the mapping between pseudoword and shape.

In addition to the mechanisms tested by our functional localizers, other theories offer specific predictions about anatomical regions or systems that may be involved in the word-shape mapping. For example, the proposal by Ramachandran and Hubbard (2001), that the mapping is based in a coupling of sensory and motor representations related to speech sound articulation (e.g. rounding of the mouth associated with rounded vowel sounds and rounded shapes), predicts engagement of the motor and pre-motor regions of cortex in response to our stimuli. Similarly, the proposal that the shape of letters is associated with particular sounds, predicts effects might be found in cortical areas involved in representing visual word form. Thus even without dedicated localizers, the present study could provide preliminary evidence corroborating either (or both) of these theories if we find engagement of these regions for our stimuli.

Method

Participants

Twenty college-age adults were recruited from Emory University and gave written informed consent to participate in the study. One participant was later excluded due to excessive motion during scanning (> 1.5mm). Thus, 19 subjects remained in the final dataset (9 male, 10 female; mean age 25 years, one month, range 19-34 years). All participants were native speakers of English and were right-handed (as determined by a validated subset of the Edinburgh handedness inventory; Raczkowski, Kalat, & Nebes, 1974). Participants reported normal hearing and normal or corrected-to-normal vision, and none reported or showed signs of neurological disorders. Participants were compensated for their time. The research protocol was approved by the Emory University Institutional Review Board.

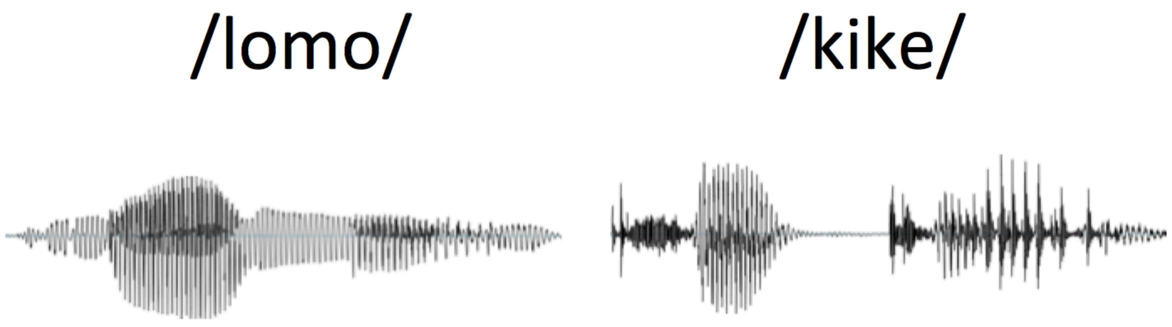
Stimuli. We created two auditory stimuli (pseudowords) and two visual stimuli (novel two-dimensional outline shapes), with each pair contrasting in how rounded/pointed they were rated as either sounding (pseudowords) or appearing (shapes) (McCormick et al., 2015). Auditory stimuli consisted of two pseudowords, *keekay* and *lohmoh* which were rated as sounding extremely pointed and rounded, respectively, in a prior behavioral study (McCormick et al., 2015) (see appendix for details). Both pseudowords were two syllables in length, spoken in a similar prosody by the same talker, a female native speaker of American English. Pseudoword stimuli were recorded in Audacity v2.0.1 (Audacity Team, 2012), using a SHURE 5115D microphone and an EMU 0202 USB external sound card, at a 44.1 kHz sampling rate. Audio recordings were then processed in Audacity, where recordings were edited into separate files, amplitude-normalized, and down-sampled to a 22.05 kHz sampling rate. Stimulus duration was 533ms for *keekay* and 600ms for *lohmoh* (Fig. 16A). The visual stimuli consisted of two shapes (gray line drawings on a black background; see Figure 16B), which were rated as appearing extremely rounded and pointed in a set of norming studies identical to those run for the pseudowords, but with the task of rating a set of 90 line drawings of abstract shapes.

Procedure

General. Scanning for this functional neuroimaging study was conducted in 1-2 sessions depending on whether localizers needed to be run. A subset of the participants (n=14) had already been tested on the three functional localizers for a the study described in Chapter 2 (approximately 4 months prior). The remaining five participants took part in the pseudoword-shape experiment first, followed by the three localizer scans. All experiments were presented using Presentation software (Neurobehavioral Systems Inc., Albany, CA) implemented on a laptop computer, enabling synchronization of stimulus presentation and scan acquisition as well as recording button-press responses and response latency for responses made using a scanner compatible hand-held button box. Additional behavioral testing was conducted following scan sessions in order to determine the strength of each participant's crossmodal pseudoword-shape mapping. Following all scan sessions and behavioral testing, participants completed the Object-Spatial Imagery & Verbal Questionnaire (OSIVQ: Blazhenkova & Kozhevnikov, 2009) to identify individual differences in verbal and object imagery. We reasoned that individuals with a greater propensity for verbal processing could be more prone to assign verbal meanings to the pseudowords in the study, whereas object imagers could be more prone to visualize the shapes associated with the pseudowords.

Pseudoword-shape fMRI task. In the experiment, the shape stimuli subtended ~1° of visual angle and were presented in the center of the screen for 500ms (see Fig. 16B). Visual stimuli were projected onto a screen at the back of the scanner bore and viewed by participants through a mirror angled over the head coil. Auditory and visual stimuli were concurrently presented in couplings that were either crossmodally congruent (*keekay*/pointed shape, *lohmoh*/rounded shape) or incongruent (*keekay*/rounded shape, *lohmoh*/pointed shape; combinations shown in Figure 17). Each of the four multisensory stimulus pairings was presented a total of 80 times (20 times in each of the four functional runs).

A.



B.

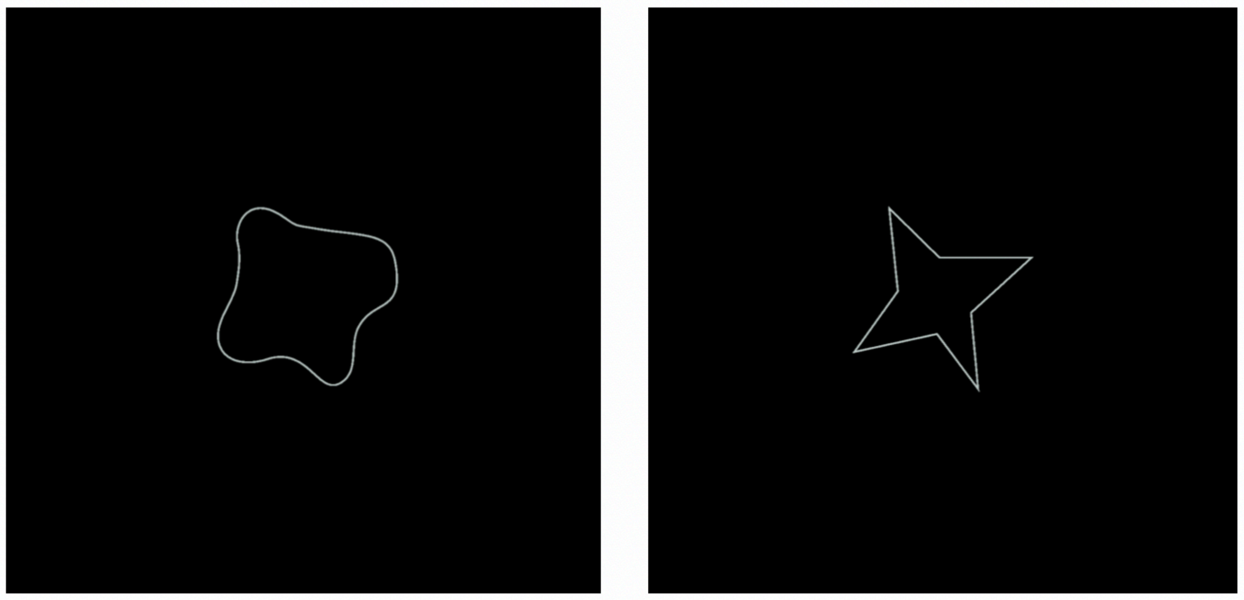


Figure 16. Pseudoword stimuli used in Experiment 2. Phonemic transcriptions and waveform plots of the sound stimuli (A). Rounded and pointed object shape stimuli used in Experiment 2 (B).

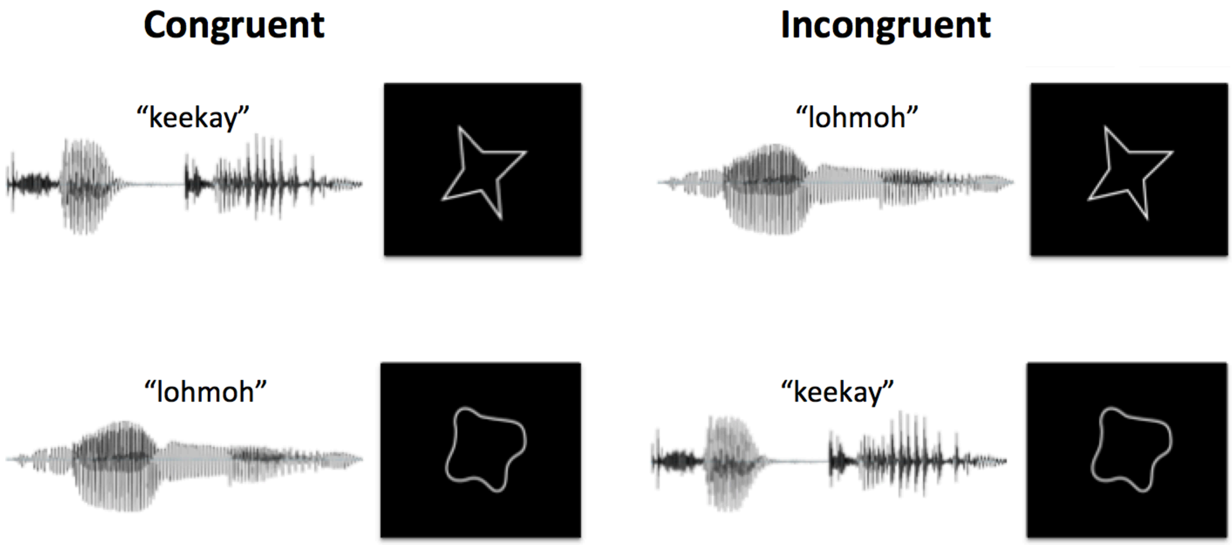


Figure 17. Visual and auditory stimuli for the pseudoword-shape experiment were paired in congruent and incongruent couplings.

Prior to scanning, participants were fitted with earplugs and scanner-compatible headphones. Once inside the scanner, participants were played the pseudoword stimuli and asked to select a volume level that was loud enough to be clearly audible over scanner noise, but not uncomfortable. A high-resolution anatomical volume was collected prior to functional testing.

Functional data were collected over the course of four runs (run duration 6:24), each consisting of eight 30-second task blocks (4 congruent, 4 incongruent) alternating with nine 16-second rest periods. Each block contained 10 x 3-second trials in which an audiovisual stimulus was presented followed by a blank interval of ~2.5 seconds (3 seconds from one stimulus onset to the next). The visual component of the stimuli was presented for 500ms followed by a 2.5-second fixation cross, whereas auditory components of the stimuli had slightly longer total durations (keekay=533ms, lohmoh=600 ms). This timing meant that the stimulus onset occurred at both the beginning and the middle of the 2-second TR. Trials and block conditions were pseudorandomly interleaved (no more than three trials in a row of the same pairing, no more than two blocks in a row of same condition). Each of the four unique audiovisual stimulus pairings was presented 20 times per run (totaling 80 trials across experimental runs). The order of runs was counterbalanced across subjects.

Across experimental blocks, participants engaged in a two-alternative forced-choice (2AFC) task. For two of the four runs, participants were asked to attend to the auditory component of the multisensory stimulus, and for the other two runs, they attended to the visual stimulus component (order of attended modality was counterbalanced across participants).

For the attend-auditory runs, participants pressed one button when they heard the pseudoword keekay and another other button when they heard lohmoh. For the attend-visual runs, participants pressed one button when they saw the pointed shape and pressed the other button when they saw the rounded shape. Participants used right index and middle fingers to make speeded responses on a button box. Response buttons were counterbalanced across participants such that half of the participants used their right index finger to press a button for a particular auditory stimulus, and half used their right middle finger to respond to that stimulus. Response buttons for visual stimuli were similarly counterbalanced. Response key mappings for the attend-auditory and attend-visual segments of the task were counterbalanced between-subjects such that half the subjects were asked to use the same finger/button to respond about auditory and visual stimuli we considered congruent and half used the same finger/button to respond to auditory and visual stimuli we considered incongruent. Performing this task accurately required that participants attend to the perceptual dimensions of interest while keeping extraneous task demands to a minimum.

Functional localizer tasks. To generate data-driven predictions about where the pseudoword-shape crossmodal associations were likely to be represented, three types of functional localizer tasks were run. Localizer experiments are described in detail in Chapter 2. The activity observed for each localizer task established the neural regions involved in multisensory integration, semantic, or magnitude-related processing, and in areas we would expect to be active under the various competing hypotheses as to the basis of the crossmodal association of pseudoword and shape. Functional localizers are described in detail in the previous chapter.

Post-scan behavioral testing: Implicit association of pseudoword and shape. Following scanning, we conducted behavioral testing outside the scanner to establish that individual participants reliably associated the auditory and visual stimuli used in this experiment. The implicit association task was as described for the experiment reported in Chapter 2, but employed the unimodal stimuli from this experiment (pseudowords keekay and lohmoh and jagged and rounded shapes).

MRI image acquisition

Scanning hardware and parameters were as described in Chapter 2, with the exception of volume acquisition in the four multisensory runs. For the present experiment, we collected 192 volumes for each of four multisensory pseudoword-shape runs.

Analysis

In-scanner behavioral data analysis. We investigated whether audiovisual congruency of trials affected task performance (either in overall accuracy or response time latency (RT)). Trials for which there was no response (n=39, 0.6% of trials) were removed from the dataset prior to analysis. The remaining dataset (all trials for which there was a response) was used to calculate overall accuracy. To prepare reaction time data for analysis, we excluded incorrect responses (n=152, 2.5% of all trials). We then trimmed outliers by calculating subjects' mean response times for auditory and visual conditions separately, then trimming responses with latencies in excess of 2.5 standard deviations from each subjects' mean. This resulted in 151 responses (2.6% of correct response trials) being trimmed from the (correct-only) dataset (mean 7.95 responses trimmed per subject, range of 6-12 trials per subject). With the resulting trimmed dataset, we calculated mean RTs for attend-auditory and attend-visual trials in congruent and incongruent conditions for each subject. We then conducted a 2x2 ANOVA within-subject to compare individual mean RTs for the two congruency and attended-modality conditions. Follow-up analyses were paired–samples t-test.

Image processing and analysis. Image processing and analysis were conducted as described in Chapter 2, except using the group-averaged brain based on this set of subjects (n=19).

Multisensory congruency analysis. We conducted univariate analysis of blood-oxygen-level dependent (BOLD) signal change to compare patterns of activity for congruent and incongruent pairings of our multisensory pseudoword-shape stimuli. We contrasted the congruent and incongruent block conditions to identify voxels that were more active in one condition relative to the other. Additional univariate contrasts decomposed the dataset to examine contrasts in the attend-auditory and attend-visual conditions separately.

Functional localizer analyses. Data analysis was as described in the localizer section of Chapter 2. As in the previous study, we compared the distribution of these functionally-localized neural systems to our various multisensory congruency effects. By examining areas where congruency-related activity overlapped with functional localizers, we identified functional systems that could support the cross-sensory mapping between pseudoword and visual object shape.

Results

Behavioral

In-scanner tasks.

Localizer tasks. In the multisensory integration localizer task, participants correctly identified an average of 7 out of 8 oddball target trials (mean ± sem) (hit rate of 87.5± 4.5%). Due to the low number of oddball trials (eight per participant) in the multisensory localizer, we conducted a non-parametric Wilcoxon test of related samples, which indicated that performance was not significantly different for targets in the synchronous (90.1±3.9%) and asynchronous (85.5±5.5%) oddballs (Z = -1.4, *p*=.157). Overall, there was a mean of 8.4 false alarms, which was driven by two subjects who responded to every trial in one or both runs (subjects 06 and 16, respectively). In the magnitude localizer task, accuracy did not differ significantly for the magnitude estimation (92.2±2.0%) and control (96.4±1.3% correct) tasks (t_{18} =-1.771, *p*=.09). However, responses were significantly faster for the control task (900±45ms) than for the magnitude task (991±53ms; t_{18} =3.44, *p*=.003). For the semantic localizer, participants made the proper response to the visual cue at the end of each sentence or non-word string at similar rates. Accuracy did not differ in the sentence (98.4±1.0%) and pseudoword (97.7±1.2% correct) conditions (t_{18} =.908, *p*=.38). Participants made an average of 1.11 false alarms over the course of the two runs.

Pseudoword-shape task. To prepare accuracy and reaction time data for analysis, we first excluded trials for which there was no response (n=39, 0.6% of all trials). A 2x2 (modality x congruency) repeated-measures ANOVA (RM-ANOVA) of accuracy data showed there was no main effect of attended-modality (mean±sem) (attend-auditory 97.4±0.7% versus attend-visual 97.6±0.6%: $F_{1,18}$ = .05, *p* = .82) and no main effect of congruency (congruent trials M= 97.7±0.5% versus incongruent trials M=97.3±0.6%: $F_{1,18}$ = .98, p = .34). There was a significant interaction between attended modality and congruency (F$_{1,18}$ = 9.59, p = .006). Post hoc tests indicated that the interaction was primarily driven by differences in the attend-auditory condition of the congruent (M=98.2±0.6%) vs. incongruent (M=96.5±0.9%, t=2.6, *p*=.036, bonferroni-corrected) comparison.

To prepare reaction time data for analysis, we excluded incorrect responses (n=152, 2.5% of responses), and further excluded trials for which the RT exceeded 2.5 standard deviations from the individual participant mean (n=151, 2.6% of correct response trials). With the trimmed dataset, we calculated mean response times for attend-auditory and attend-visual trials in congruent and incongruent conditions for each subject. A 2x2 (modality x congruency) RM-ANOVA on response time data revealed that mean RTs differed depending on the attended modality, with faster responses for 'attend visual' trials (474±21ms) compared to 'attend auditory' trials (527±23ms: $F_{1,18}$ = 21.33, *p* < .001). There was also a significant difference in response latencies depending on the multisensory congruency of the audiovisual stimulus – with faster responses for congruent trials (489±21ms), compared to incongruent trials (513±22ms: $F_{1,18}$ = 18.75, *p* < .001). The interaction of modality and congruency on RTs was also significant (F$_{1,18}$ = 9.23, *p* < .007), with incongruent conditions slower than congruent in both auditory (545ms vs 509ms) and visual (480ms vs 469ms) modalities (Figs. 18-19). Post-hoc testing showed several group differences underlying this interaction. Overall, there was a more pronounced difference in RTs in the attend-auditory condition, with responses for congruent stimuli 36 milliseconds faster than incongruent (Maud-cong=509±22ms, Maud-incong= 545±26ms, t(18=-4.04,*p* =.004, bonferroni corrected) compared to the attend-visual trials for which responses in the congruent condition were 11 ms faster than for incongruent (Mvis-cong=469± 20ms, Mvis-incong= 480 ± 21ms, t_{18}=-2.95, *p*=.036, bonferroni corrected). Responses were significantly faster in the respond-visual condition compared to the respond-auditory condition, and this pattern held for both congruent (t=-4.02, *p*=.004, bonferroni corrected) and incongruent (t=-4.66, *p*<.004, bonferroni corrected) audiovisual stimulus couplings.

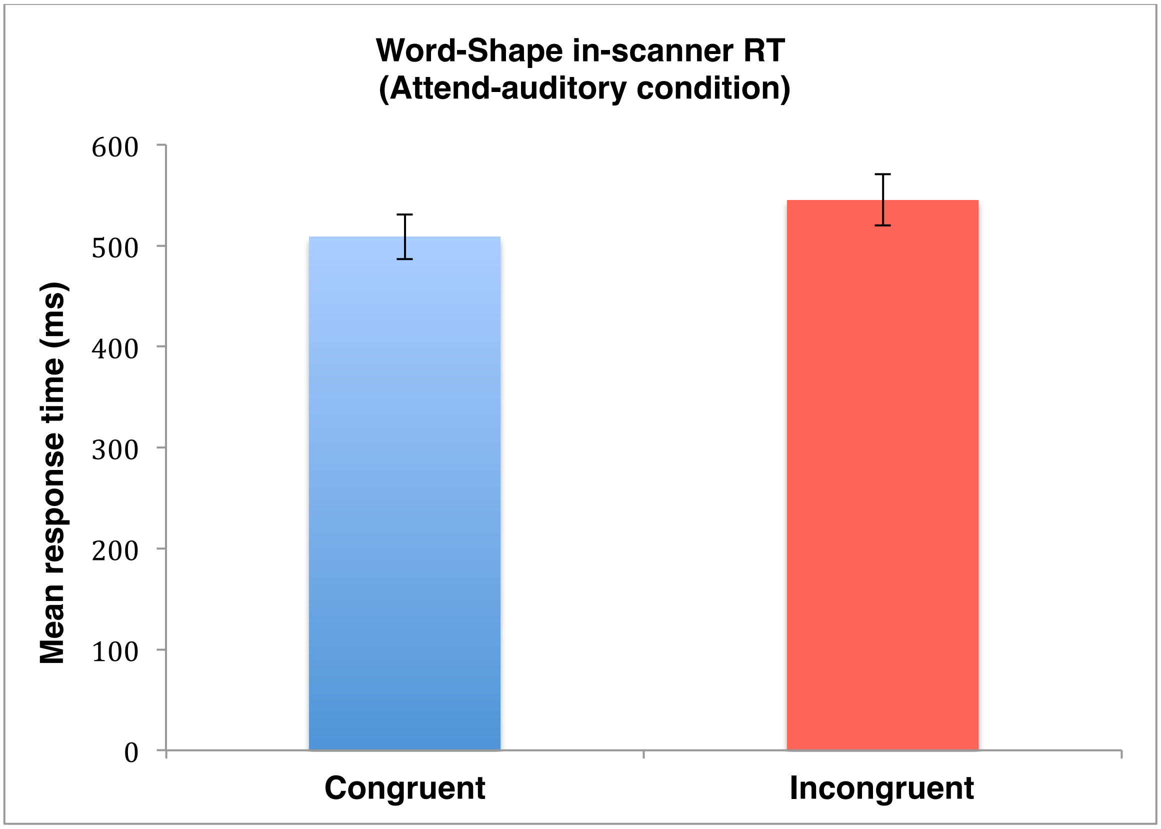


Figure 18. Response times for crossmodally congruent and incongruent trials of the in-scanner task in the attend-auditory condition.

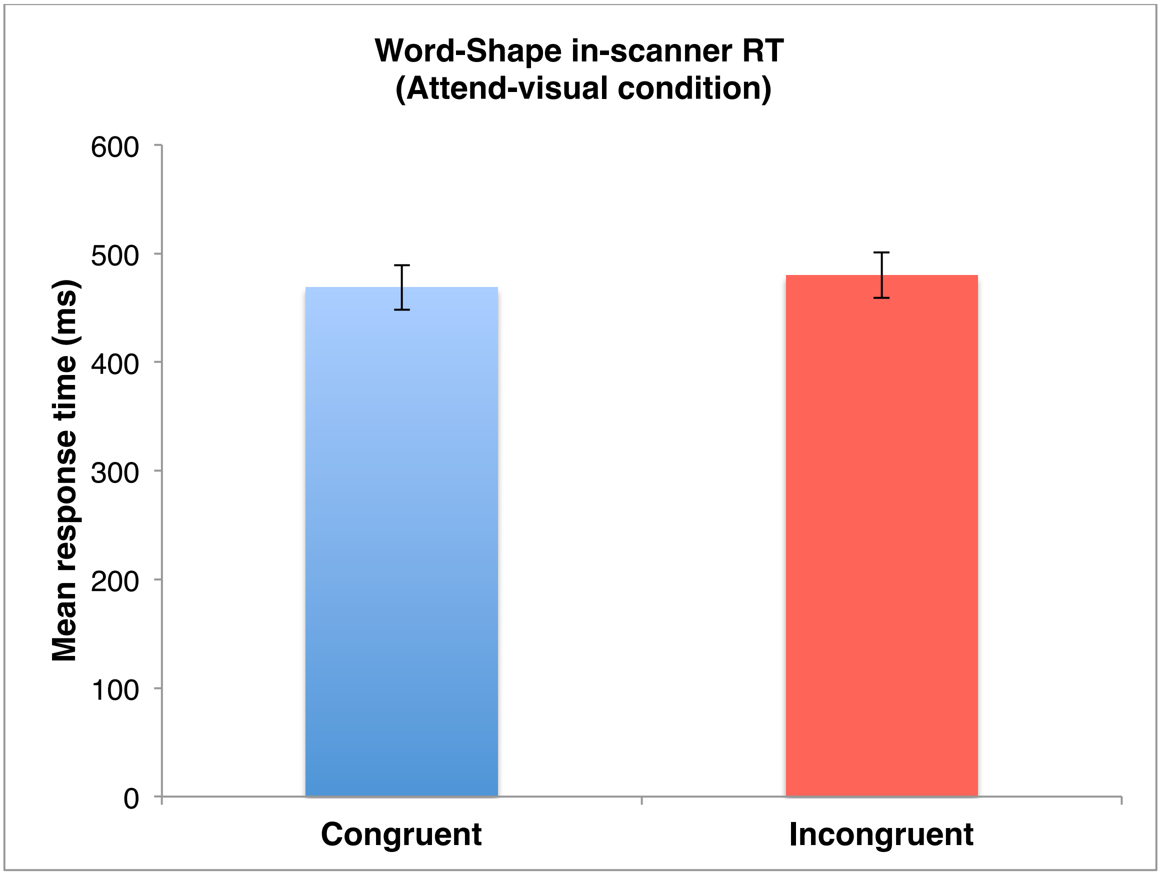


Figure 19. Response times for crossmodally congruent and incongruent trials of the in-scanner task in the attend-visual condition.

Post-scan pseudoword-shape IAT. A logging error resulted in a total of 44 trials not being logged across 8 subjects (from 0 to 15 trials per subject not logged). Trials for which the participant did not make a response (n= 6, .08% of the logged trials) were excluded from the dataset. Overall accuracy on the task was 93.8%. A 2x2 (modality x key mapping congruency) repeated-measures ANOVA (RM-ANOVA) demonstrated main effects of both stimulus modality and congruency of key mapping on response accuracy. Responses were significantly more accurate for the pseudoword stimuli (95.8±0.8%) than the visual shapes (91.9±0.8%: $F_{1,18}$ = 31.86, p < .001) and when response keys were congruently coupled (95.3 ±0.7%) than when they were incongruent (92.4±1.2%: $F_{1,18}$ = 5.73, p = .03). The modality x response key congruency interaction was not significant ($F_{1,18}$ <.01, p = .9).

To prepare reaction time data for analysis, we excluded incorrect responses (6.2% of responses) and trimmed outliers. We calculated subjects' mean response times for auditory and visual conditions separately, trimming responses with latencies in excess of 2.5 SDs from each subjects' mean. For the auditory condition, this resulted in 92 responses (2.6% of the correct responses) being trimmed (mean 4.84 responses trimmed per subject, range of 1-8). For the visual condition, 95 responses (2.9% of the correct responses) were trimmed (mean 5.0 responses trimmed per subject, range of 3-8). Overall, 2.8 % (n=187) of correct trials were trimmed due to excessive latencies (mean 9.8 trimmed per subject, range of 4-14). Mean response times for auditory and visual stimuli were then calculated for each subject using the trimmed dataset. A 2x2 (modality x key mapping congruency) RM-ANOVA compared RTs for trials and demonstrated main effects of both modality and key mapping congruency. Participants responded more quickly for the visual stimuli (606±19ms) than the auditory stimuli (702±21ms: $F_{1, 18}$ = 31.15, *p* < .001), and when response key mappings were congruent (580±18ms) than when they were incongruent (728±22ms: $F_{1,18}$ = 64.53, p < .001) (Fig. 22). The modality x response key congruency interaction was not significant ($F_{1, 18}$ = .5, *p* = .5). Eighteen out of nineteen participants exhibited the expected pattern for the pseudoword stimuli whereas all nineteen subjects showed the expected pattern for the visual stimuli (Figs. 20-21).

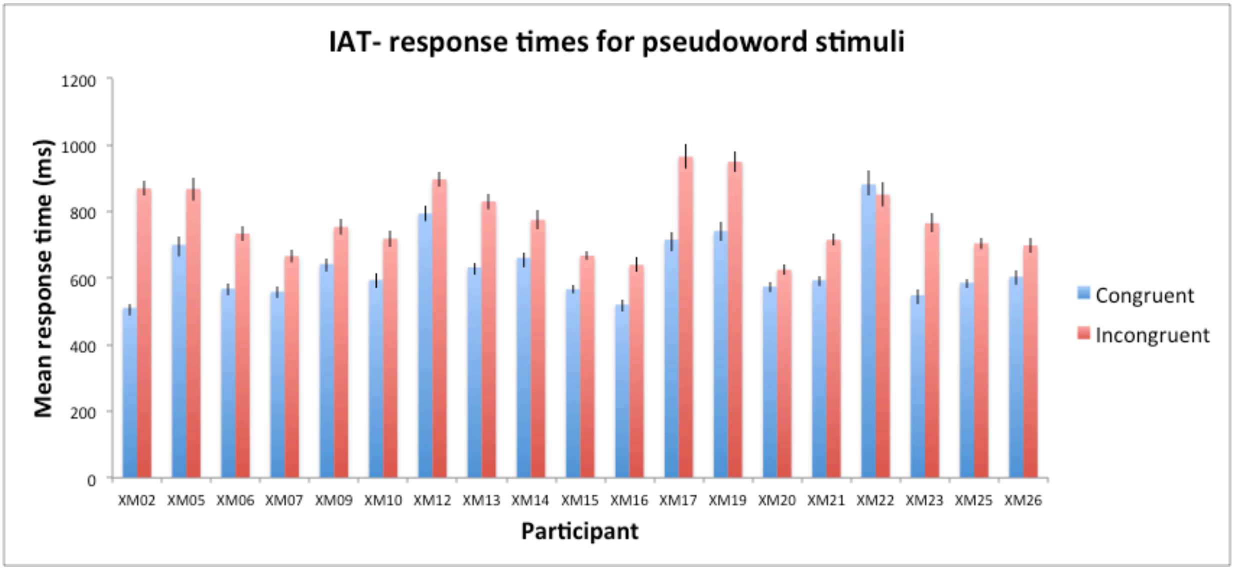


Figure 20. Mean response times for the auditory stimuli on the word-shape IAT by subject. Eighteen of the nineteen subjects responded more quickly in the congruent conditions.

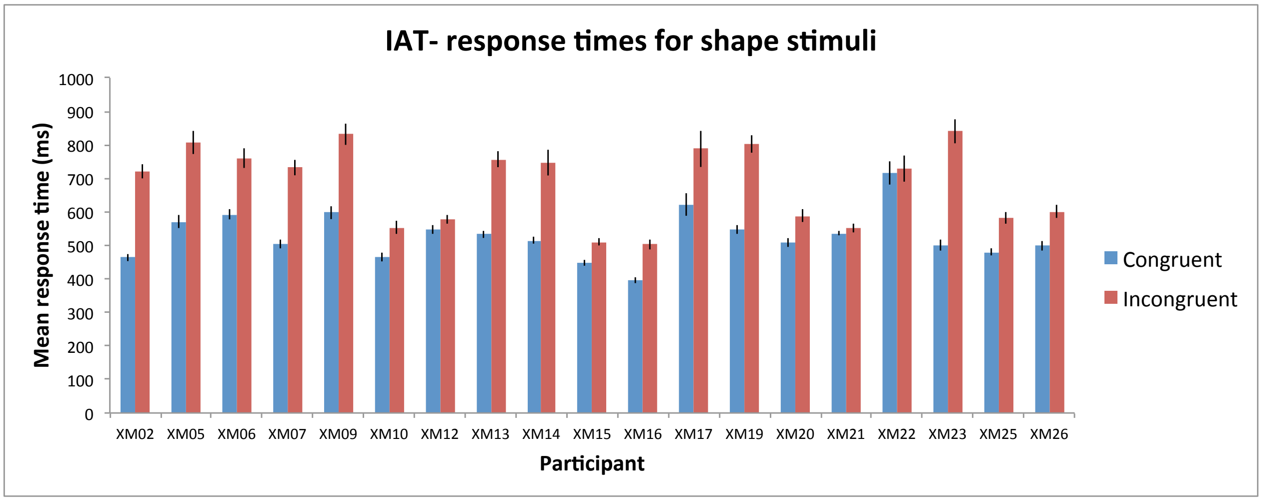


Figure 21. Mean response times for the visual stimuli on the word-shape IAT by subject. All nineteen subjects responded more quickly in the congruent conditions.

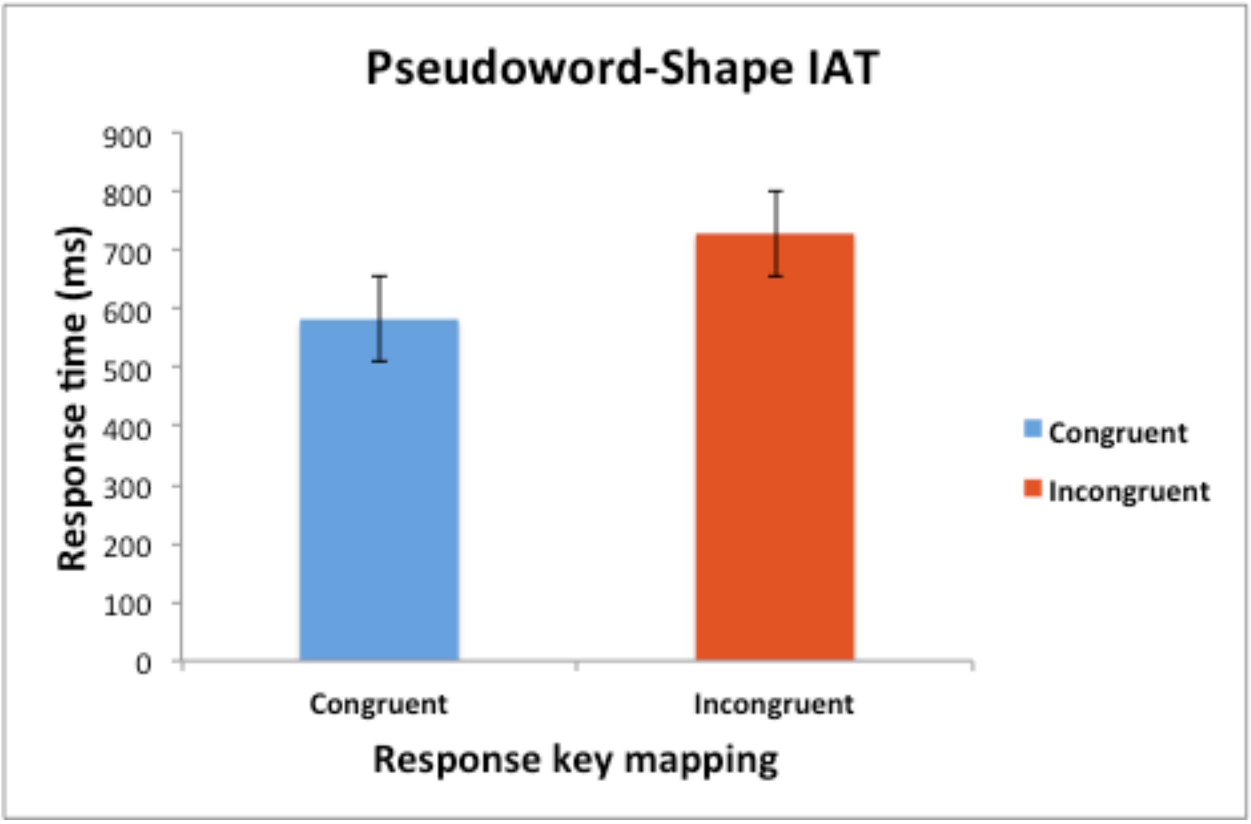


Figure 22. Mean response time on word-shape IAT by congruency of key mapping. Error bars= standard error.

Imaging

Localizer tasks.

Multisensory synchrony. The Synchronous > Asynchronous contrast within the cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels) produced bilateral activations in the anterior calcarine sulcus extending through the posterior occipital fissure (POF) to the cuneus. In the left hemisphere this activation extended to the lingual gyrus (LG) and posterior cingulate gyri (Table 3a; Figure 23). The reverse contrast, Asynchronous>Synchronous produced more widespread activity than the previous contrast, including an active cluster in the right inferior parietal cortex extending across the supramarginal gyrus (SMG) and angular gyrus (AG) and into the anterior intraparietal sulcus (aIPS) and mid-intraparietal sulcus (midIPS). Another right hemisphere cluster was focused in the right inferior frontal gyrus (IFG) (Table 3b; Fig. 23). We did not find activity in regions most widely reported to be sensitive to audiovisual synchrony (e.g. STS/STG), possibly due to differences in stimuli and tasks (see Discussion). However a number of previous studies (Erickson, Heeg, Rauschecker, & Turkeltaub, 2014) have found asynchrony or incongruency responses in proximity to the IFG site produced by the asynchronous>synchronous contrast.

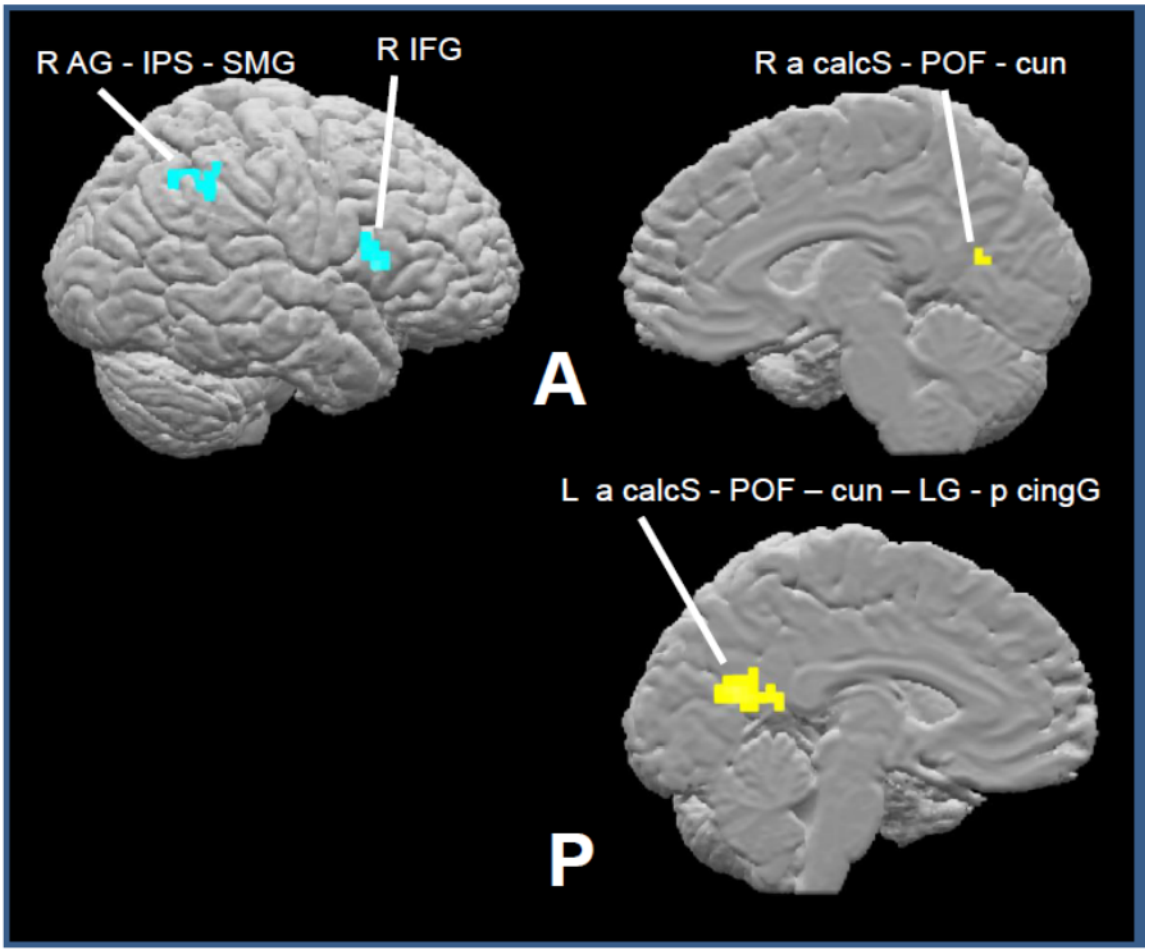


Figure 23. Multisensory integration localizer within cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 8 voxels). Contrast of synchronous > asynchronous (turquoise) reveals putative integration network (Table 3a); asynchronous > synchronous (yellow: Table 3b)..

Magnitude. The magnitude (magnitude > control) contrast within the cortical mask (voxel-wise threshold *p* < .001, cluster-corrected p < .05, cluster threshold 7 voxels) produced three major foci of activity, all in the right hemisphere. Active clusters were located in the right supramarginal gyrus (SMG), the right SPG extending into IPS, and the right middle occipital gyrus (MOG) extending through intra-occipital sulcus (IOS) to the superior occipital gyrus (SOG)(Table 3c, Figure 24). These foci are consistent with activity reported in previous studies on magnitude processing (Dehaene, Piazza, Pinel, & Cohen, 2003; Eger, Sterzer, Russ, Giraud, & Kleinschmidt, 2003; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Pinel et al., 2004).

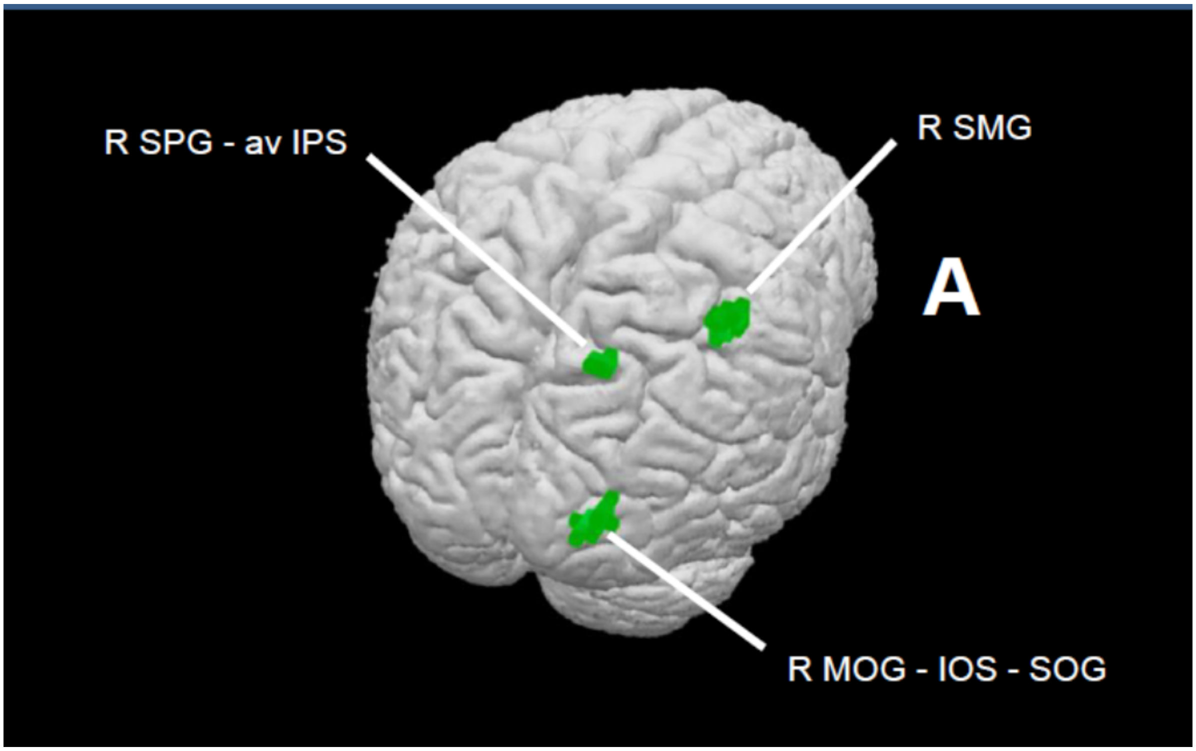


Figure 24. Magnitude localizer within cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 7 voxels). Contrast of magnitude estimation > control reveals magnitude network Table 3c.

Semantic. Contrasting complete sentences>non-words within the cortical mask (voxel-wise threshold *p* < .001, cluster-corrected *p* < .05, cluster threshold 9 voxels) revealed large bilateral sites along the STS extending to parts of superior temporal gyrus (STG), extending more posteriorly on the left than the right. Additional activity in the left hemisphere was found precentrally and in middle occipital gyrus (MOG)(Table 3d; Figure 25). These findings generally replicate those of Fedorenko and colleagues, the team that developed this localizer to identify language-sensitive brain regions (Fedorenko et al., 2011; Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010).

The reverse contrast of pseudowords > complete sentences produced widespread bilateral frontoparietal activations. In the right hemisphere one cluster extended from the SFS to IFG on the right, and in the AG extending to supramarginal gyrus (SMG) on the left and IPS on the right. Additional right hemisphere clusters were found in the posterior cingulate sulcus/gyrus, the precuneus/POF, and the AG/midIPS. Several sites were identified in the left hemisphere including: a cluster in the left medial SFG, one cluster in left posterior cingulate sulcus/gyrus, one cluster in the L MFG and SFS, one cluster with activity extending from the left IFS/IFG, to lateral orbital gyrus and anterior orbital gyrus, one cluster in the left SMG and AG, one cluster in the left SPG, IPS and POF, and a cluster with activity in the left MOG and inferior occipital sulcus (Table 3e; Figure 25).

The contrast of pseudowords>sentences may reflect the additional resources deployed in processing the unfamiliar pseudowords. Similar to our functional localizer, Binder and colleagues contrasted activity for written pseudowords and concrete words and found involvement of the left IFG proximal to one of our sites (Binder, Westbury, McKiernan, Possing, & Medler, 2005). Activity produced by this contrast could be related to phonological processing- a possibility supported by two studies, which implicate sites in the L SMG in phonological processing of written words (Price, Moore, Humphreys, & Wise, 1997; Wilson, Tregellas, Slason, Pasko, & Rojas, 2011). Activity for this contrast may reflect the greater effort required to read unfamiliar pseudowords compared to complete sentences. In line with this possibility, several of the foci produced by our contrast align with the frontoparietal ‘multiple demand’ network proposed to help flexibly balance processing demands (Duncan, 2013; Fedorenko et al., 2013).

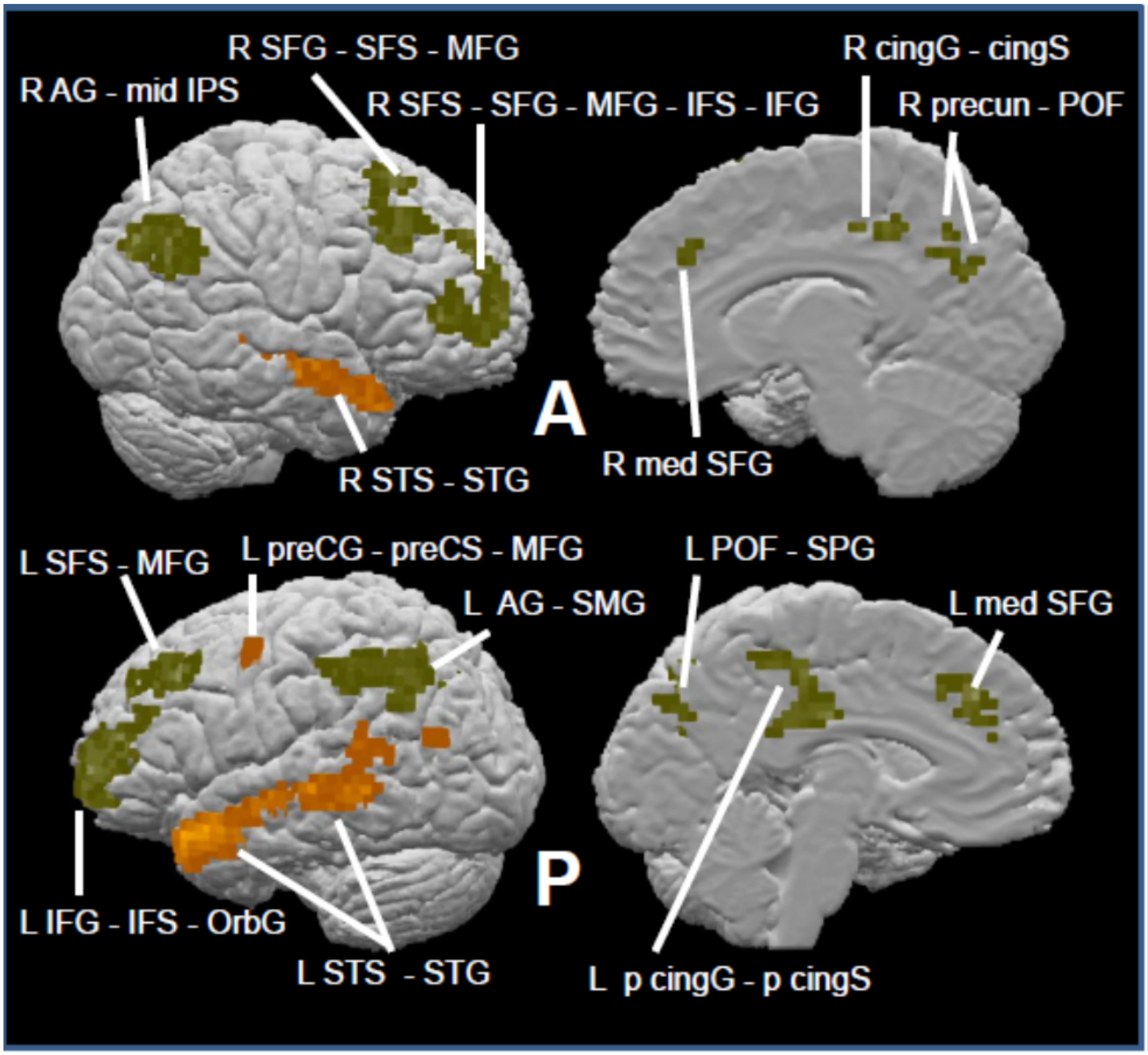


Figure 25. Semantic localizer within cortical mask (voxel-wise threshold p < .001, cluster-corrected p < .05, cluster threshold 9 voxels). Contrast of sentences > pseudowords (orange) reveals semantic network (Table 3d); pseudowords > sentences (olive) likely reflect phonological processing (Table 3e).

Table 3. Localizer activations: multisensory integration (a,b), and magnitude (c) and semantic (d,e); all within cortical mask, voxel-wise threshold *p* < .001, cluster-corrected p < .05, cluster thresholds = semantic, 9 voxels; multisensory integration, 8 voxels; magnitude, 7 voxels; x,y,z Talairach coordinates for centers of gravity.

	Region	x	y	z
(a) Synchronous > asynchronous	R a calcS - POF - cuneus	2	-60	15
	L a calcS - POF - cuneus - LG -p cingG	-4	-58	15
(b) Asynchronous > synchronous	R SMG – a IPS – mid IPS - AG	38	-47	42
	R IFG	53	11	16
(c) Magnitude > control	R SMG	42	-40	47
	R SPG - av IPS	18	-63	42
	R MOG - IOS – SOG	25	-86	8
(d) Sentences > pseudowords	R a STG – a STS –mid STS – p STS	49	-6	-8
	L preCG – preCS - MFG	-44	-5	48
	L MOG	-40	-69	21
	L a STG - a STS - mid STS - p STS - p STG	-51	-19	-3
(e) Pseudowords > sentences	R SFG - SFS - MFG	29	19	47
	R SFS - SFG - MFG - IFS - IFG	32	49	15
	R med SFG	8	33	31

R p cingS - p cingG	4	-31	36
R precun - POF	11	-62	29
R AG – mid IPS	46	-57	35
L med SFG	-2	31	32
L p cingS - p cingG	-4	-29	35
L MFG - SFS	-31	28	38
L IFS - IFG - lat OrbG - a OrbG	-31	48	8
L SMG - AG	-44	-52	41
L SPG - IPS - POF	-12	-68	31
L MOG - InfOS	-36	-79	-3

Pseudoword-Shape incongruency. We tested for regions sensitive to the pseudoword-shape mapping with several contrast analyses, identifying systems modulated by the intersensory attributes of interest.

I first tested for a congruency effect (voxels more active for congruent stimuli than incongruent) by contrasting BOLD activity for congruent and incongruent pairings of our multisensory pseudoword+shape stimuli. Within the cortical mask, at a voxel-wise threshold of *p* < .001, no activations for the contrast of congruent blocks versus incongruent blocks (C>I) or for the reverse contrast of (I>C) survived correction for multiple comparisons.

Follow-up analyses decomposed the dataset, examining the attend-auditory and attend-visual task conditions separately and revealing significant contrasts for the attend-auditory (but not the attend-visual) task condition. While these analyses did not reveal any regions selective for the C>I contrast in either attend-auditory or attend-visual conditions, the reverse contrast (I>C) showed interesting differences.

Only in the attend-auditory condition did the I>C contrast produce significant activations within the cortical mask (voxel-wise threshold of $p < .001$, cluster-corrected $p < .05$, cluster threshold 5 voxels). This incongruency effect appeared in two clusters in the right hemisphere, one in the aIPS and one in the SMG extending to the post-central sulcus. There were three clusters of activity in the left hemisphere, one in the superior parietal gyrus (SPG), one from the L midIPS extending to SMG and aIPS, and a cluster in the middle frontal gyrus (MFG). (Fig. 26).

Table 4. Incongruency effects: pseudoword-shape incongruency-related activations in the 'attend auditory' condition within the cortical mask (voxel-wise threshold $p < .001$, cluster-corrected $p < .05$ cluster threshold 5 voxels); x,y,z, Talairach coordinates for centers of gravity.

Region	x	y	z
R aIPS[1]	33	-40	42
R SMG – poCS[1,2]	45	-38	45
L SPG	-16	-61	58
L mid IPS - SMG – aIPS[3]	-36	-47	44
L MFG[3]	-42	30	28

[1] Overlaps with the asynchronous R aIPS and SMG in Table 3b.

[2] Overlaps with the R SMG and contiguous with the R aIPS magnitude foci in Table 3c.

[3] Overlaps with the semantic control L SMG and MFG in Table 3e.

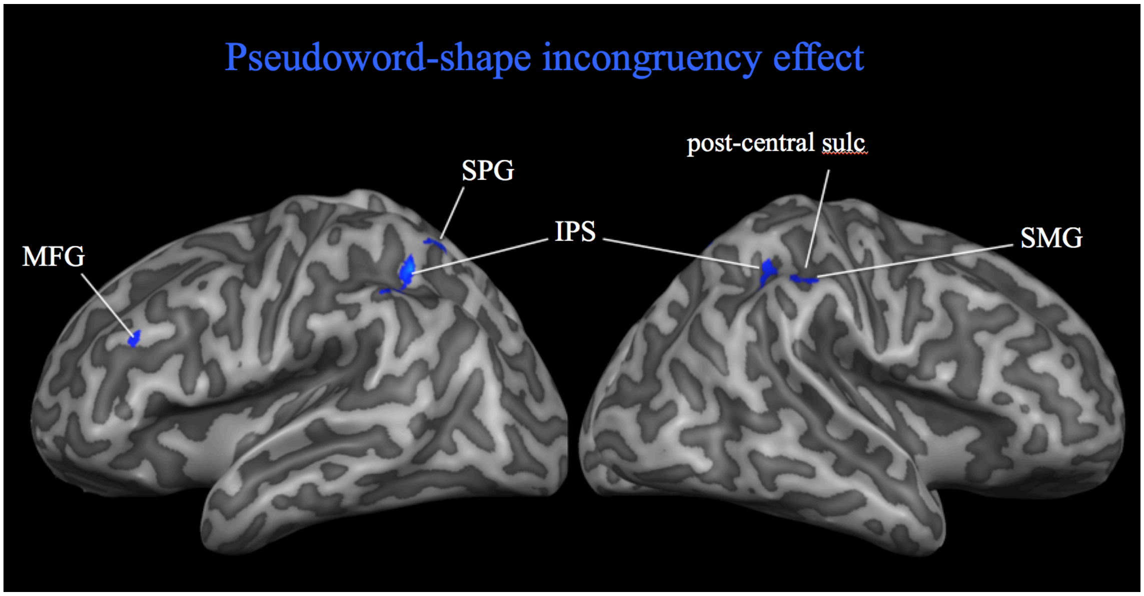


Figure 26. Pseudoword-shape multisensory incongruency effect (for attend-auditory task condition)

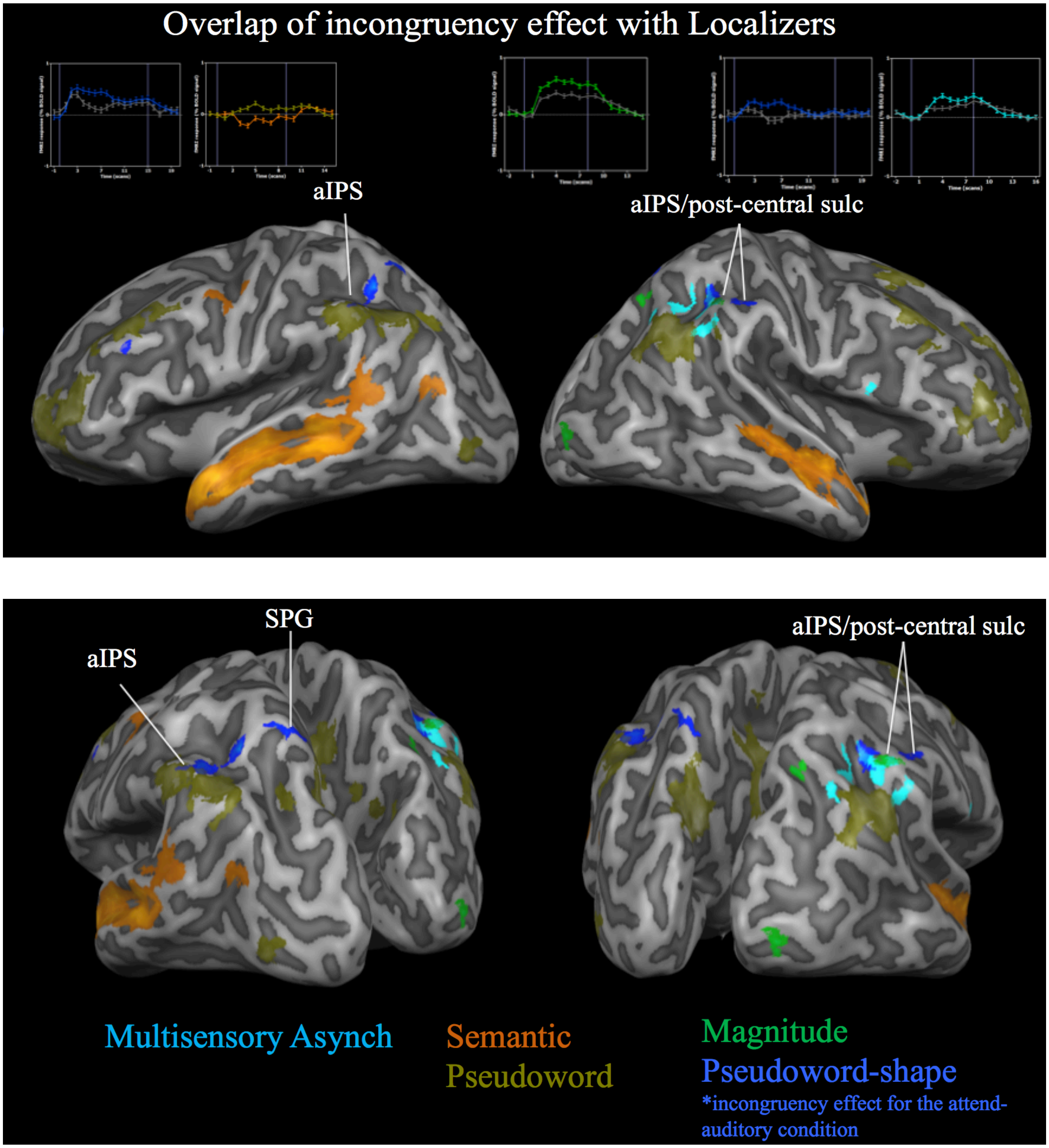


Figure 27. *Overlaps of pseudoword-shape incongruency effect and functional localizers. Abbreviations a anterior; IPS intraparietal sulcus; SPG superior parietal gyrus.*

Overlap of incongruency effect with localizers. In order to assess involvement of multisensory, magnitude, and semantic systems in the pseudoword-shape mapping, we looked for areas of overlap between these functionally localized systems and activity for our multisensory incongruency effect. Comparisons of the pseudoword-shape and localizer maps were made using a strict voxel-wise threshold of $p<.001$ for all maps. Using this stringent threshold reduces the risk of false positives that arises when making multiple comparisons, and constrains the spatial distribution of activity compared to more liberal thresholds.

Activity for the word-shape incongruency effect (attend-auditory condition) was found to overlap with several of the regions identified by our localizers.

Multisensory Integration. In the case of the multisensory integration localizer, only activity for the Asynchronous> Synchronous contrast overlapped with the (attend-auditory) word-shape incongruency effect. Both of the incongruency clusters in the right hemisphere had areas of overlap with a cluster produced by the Asynchronous>Synchronous contrast of the multisensory localizer- these overlaps were in the aIPS and SMG.

Magnitude. For the magnitude localizer, clusters for the Magnitude>control contrast overlapped with pseudoword-shape (attend-auditory) incongruency effect in the right SMG and was adjacent to the incongruency effect in the right aIPS.

Semantic. For the semantic localizer, the complete sentence>pseudoword contrast did not overlap with any word-shape activity. However, the contrast of pseudowords>complete sentences overlapped with the pseudoword-shape incongruency effect in portions of the left SMG and MFG.

Discussion

This study allows us to compare several potential mechanisms underlying the pseudoword-shape mapping. Previous studies have examined an array of sound-symbolic words or pseudowords (Ković et al., 2010; Lockwood et al., 2016; Revill, Namy, DeFife, & Nygaard, 2014; Sučević et al., 2015), which are likely to be based in heterogeneous mechanisms (e.g. pitch-size, pitch-brightness etc.; Sidhu & Pexman, 2017). Other studies have focused on sound-shape correspondence, but relied on reverse inference rather than localizers to conjecture about underlying mechanisms (Peiffer-Smadja & Cohen, 2010; 2019). To our knowledge, this is the first neuroimaging study to employ functional localizers to systematically examine the neural basis of a particular sound-symbolic mapping. Our focus on a single mapping (the association of pseudowords and shapes) and use of functional localizers to isolate possible mechanisms provides new insight into this particular mapping, but also offers a ready framework for testing other cross-sensory and sound-symbolic mappings.

Incongruency effects

We find several regions that respond more strongly to incongruent than congruent couplings when participants are attending to the auditory pseudowords (but no such effect when they are attending to the visual shapes). Corresponding to these imaging findings, we also see stronger behavioral congruency effects for the attend-auditory runs compared to the attend-visual runs, where there was no effect of congruency on accuracy and a less dramatic effect of congruency on RT compared to the attend-auditory runs. This difference in congruency effects depending on the modality attended may arise from differences in timing of processing the two kinds of stimuli. Unlike the visual stimuli, which appear from the start of a trial, the pseudoword stimuli unfold over time, which may make the auditory stimuli more subject to influence of the unattended modality. In the attend-visual condition, a participant can see the shape stimulus and prepare a response immediately, while the pseudoword may still be playing, which could lessen the influence of the unattended pseudoword stimulus. There is substantial evidence indicating that even when attention is focused on a single modality of a multisensory object, information from the unattended (and task-irrelevant) modalities is processed as well (Busse, Roberts, Crist, Weissman, & Woldorff, 2005; Driver & Spence, 2000; Miller, 1991; Molholm & Foxe, 2010; Molholm, Martinez, Shpaner, & Foxe, 2007; Zimmer, Roberts, Harshbarger, & Woldorff, 2010), though recent studies have shown that intermodal processing is modulated by attention relatively late in processing (~500-700ms; Shrem & Deouell, 2017). This relatively late influence of crossmodal attention could account for the asymmetry of the effect, as it could have more time to influence the temporally-unfolding pseudoword stimulus to a greater extent than the visual shape. It has also been suggested that there is a bias toward visual information when there is a conflict between sensory information (Ben-Artzi & Marks, 1995; Molholm, Ritter, Javitt, & Foxe, 2004). Such a bias could manifest as a stronger incongruency response when (unattended) visual information is incongruent with the attended pseudoword than when the reverse is true. Related to this is the proposal that lexical processing takes place further along the processing stream, making it more subject to influence of incoming visual information than vice versa (Melara & Marks, 1990b).

We found several regions that exhibited an incongruency effect- bilaterally in parts of parietal cortex, as well as one site in the left MFG. These regions have been shown to respond to multisensory stimulus attributes by several previous studies (Hein et al., 2007; Noppeney, Josephs, Hocking, Price, & Friston, 2008; Peiffer-Smadja & Cohen, 2010; 2018).

A number of teams have examined the role of crossmodal congruency on the integration of audiovisual stimuli. A key assumption of such testing paradigms is that if an individual has a bias towards associating particular auditory and visual stimuli, presenting them with mismatched or incongruent stimuli should pose a burden on processing and require additional resources relative to a congruent pairing. In contrast, crossmodally-congruent stimuli may require fewer processing resources because congruent pairings represent a type of intersensory redundancy, deriving from, or referring to the same semantic source or object (Bahrick, Lickliter, & Flom, 2004; Shams & Seitz, 2008; Stein et al., 2010). If the observed activity is based in a multisensory system that responds to intersensory redundancy, this could account for the incongruency effect observed in our study and similar findings across several other studies reporting greater activity for the crossmodally incongruent conditions relative to the congruent conditions, possibly due to the added burden of attempting to integrate incongruent or otherwise mismatched sensory inputs (Noppeney et al., 2008).

In separate neuroimaging studies, Hein et al. (2007) and Noppeney et al. (2008) found regions that show greater activity for crossmodally incongruent compared to congruently paired environmental stimuli, and in both cases, they identified sites that were located near (within 6 millimeters) our observed incongruency effect in the L MFG (Hein et al., 2007; Noppeney et al., 2008). Findings from these studies may also help explain the lack of temporal lobe involvement in the congruency mapping in our study. Hein and colleagues found that whereas temporal sites (pSTS and STG) were engaged by couplings of familiar environmental stimuli (slightly stronger for congruent than incongruent), frontal regions responded more to novelty of the AV pairing, with bilateral frontal regions showing activity for incongruent pairings of familiar stimuli (including at a site within 6mm from our word-shape incongruency effect in L MFG), as well as an overlapping site in the right that also responds for novel pairings of unfamiliar stimuli. Hein and colleagues interpreted this frontal activity as reflecting learning upon exposure to a novel or unexpected (incongruent) coupling. Indeed, we see compellingly similar profiles at the whole-brain level when comparing our word-shape incongruency effect alongside the response to unfamiliar novel AV stimuli from Hein and colleagues. On one hand, this is not surprising because, much like the unfamiliar stimuli from Hein at al. (2007), our stimuli combine novel sounds and images (McCormick et al., 2018b, McCormick et al., 2015). On the other hand, however, this finding is somewhat at odds with the behaviorally-documented congruency mapping exhibited by our participants. Although our stimuli are novel, participants demonstrate a clear preference for one AV alignment (what we term ‘congruent’) compared to the other, whereas the unfamiliar novel couplings used by Hein et al. were arbitrarily paired (not congruent or incongruent).

Noppeney et al (2008) used a crossmodal priming paradigm and modulated prime-target congruency to explore the neural basis of symbolic (spoken and written language) and nonsymbolic (sounds and images of familiar objects) stimuli. In their experiment, a trial was crossmodally congruent when the visual prime and auditory target referred to the same object (e.g. an image of a dog and the sound of a dog's bark). Their analysis revealed extensive activity for the incongruent conditions, and no regions where the congruent condition produced greater activity. Of the regions showing the crossmodal incongruency effect, a site in the left IFS was within 6mm from the edge of our incongruency activation focused in the left MFG. This region exhibited an incongruency effect for both the symbolic and non-symbolic auditory targets (spoken words and object sounds). Although not multisensory (auditory and visual stimuli are not presented concurrently), the crossmodal incongruency effects in Noppeney et al. (2008) demonstrate that this inferior frontal region is sensitive to semantically-mediated audiovisual correspondences and responds to meanings as encoded by both words and environmental sounds. Taken together, the findings from Hein et al (2007), Noppeney et al (2008), and now our own group demonstrate a role for the left IFS/MFG in responding to AV incongruency for a range of stimuli. As discussed earlier, this incongruency effect may reflect an added burden on processing as the brain attempts to integrate incongruent or inconsistent stimuli from different modalities (Noppeney, 2012) and may also reflect learning upon exposure to novel multisensory stimuli (Hein et al., 2007).

We find incongruency effects in parietal regions including bilaterally in the IPS and SMG, and the right post-central sulcus, and the left SPG. Previous research has linked these parietal regions with multisensory congruency and binding. Hein et al (2007) identified two left parietal sites that showed the greatest response for novel unfamiliar AV stimuli. Both of these sites are near (<10 mm) the parietal sites of incongruency effects from our study, the left IPS and SPG. In the 2008 study described above, Noppeney et al. also identified a site in the left AG/IPS (within 16mm of our L IPS site), which responds to AV incongruency and shows stronger incongruency effect for environmental sounds compared to words. Another line of research has reported that transcranial magnetic stimulation (TMS) to the right AG/IPS area can interfere with crossmodal congruency, and perceptual pop-out effects in behavioral performance in non-synesthetes (Bien, ten Oever, Goebel, & Sack, 2012; Taylor, Muggleton, Kalla, Walsh, & Eimer, 2011), and can knock out Stroop-like behavioral effects in synesthetes (Muggleton, Tsakanikos, Walsh, & Ward, 2007). In addition, an anecdotal report describes lesioning of angular gyrus (adjacent to the IPS) being associated with a loss of the takete-maluma effect (Ramachandran & Hubbard, 2003). Regardless of whether these studies capture different aspects of a common system, or functionally-distinct systems related to multisensory processing, the right IPS/AG area is broadly implicated in various aspects of crossmodal mappings in both synesthetic- and non-synesthetic individuals.

Notably absent from the areas where we see congruency-related effects are the temporal lobes. A preponderance of previous research has implicated swathes of the temporal lobes in multisensory integrative processes. These previous studies have typically employed audiovisual speech (Atteveldt, Formisano, Blomert, & Goebel, 2007; Erickson et al., 2014; Noesselt et al., 2012; Stevenson & James, 2009) or combinations of familiar environmental images and sounds (Hein et al., 2007; Noppeney et al., 2008), and often compare activity evoked by unisensory and multisensory stimulus conditions. The stimuli in the present experiment are distinct in a few ways. First, we do not have a unisensory condition. Because all experimental conditions in our study are multisensory, there is no BOLD contrast that captures voxelwise selectivity for AV coincidence per se, we can only isolate voxels that respond differently for our two audiovisual conditions. Second, whereas previous AV integration studies have typically employed familiar semantic stimuli, we used novel stimuli presented in congruent or incongruent couplings. Interestingly, Hein (2007) and Noppeney (2008) find crossmodal incongruency effects in close proximity in the STS. While Noppeney et al (2008) find an incongruency effect for both environmental sounds and words in the frontal regions, they find greater modulation of the response for words than environmental sounds in the STS, and the opposite pattern in parietal regions (Noppeney et al., 2008). Given that our own findings align with Noppeney and colleagues' in frontal and parietal, but not temporal regions, this could suggest that our pseudoword stimuli are being processed more like environmental sounds than language. Hein (2007) and others have suggested that the multisensory integrative systems of the temporal lobes may be specialized for familiar semantically-based multisensory couplings, whereas both unfamiliar-, and familiar-but-incongruent audiovisual couplings tend to engage more frontal regions such as the IFC. This pattern is borne out in our data, with an incongruency effect of novel pseudoword and shape in the left MFG, within 6 millimeters of the incongruency site from Hein et al.

Overlaps with localizers

The pseudoword-shape incongruency effect overlapped with the asynchronous>synchronous contrast of the multisensory localizer in a portion of the right aIPS and SMG. In other words, voxels in these overlapping regions were sensitive to both crossmodally-incongruent pairings of shape and pseudoword, and also to temporal mismatch of audiovisual stimuli. The right aIPS site is also close to regions of IPS shown to respond to audiovisual spatial congruency (Sestieri et al., 2006) and the effect overlaps with regions of IPS implicated more generally in multisensory attention (Regenbogen et al., 2018). These findings raise the possibility that a common system may respond to spatial, temporal, and crossmodal incongruency, and that the effect we find here may reflect a generalized mismatch response from a multisensory attention system involved in integrating sensory signals. Such response properties would be an important feature of a multisensory attention system, which serves not only to match and bind stimuli, but also to filter out concurrent but inconsistent signals in different modalities.

The incongruency region in the right SMG overlapped with activity for the magnitude localizer, which extended into the right aIPS. Previous research has consistently implicated the IPS in magnitude-related processing (Eger et al., 2003; Hubbard, Piazza, Pinel, & Dehaene, 2005; Piazza et al., 2004, 2007; Pinel et al., 2004), and provides some evidence for involvement of the SMG (Piazza et al., 2004, 2007) and AG (Cattaneo, Silvanto, Pascual-Leone, & Battelli, 2009; Seghier, 2012) in magnitude representation. Both the pseudowords and the shape stimuli used in the experiment were selected based on norming studies where they were rated as sounding (words) or appearing (shapes) either extremely rounded or pointed (while receiving very low ratings on the other dimension; McCormick et al., 2015). Critically, participants in these norming studies readily assigned magnitude-type ratings to these stimuli, judging where they should be place along the respective dimensions. These rating studies demonstrate that individuals have no problem thinking about these stimuli in terms of magnitude representations, and the overlap between incongruency activity with magnitude localizer further supports the possibility that magnitude representations may underlie the pseudoword-shape mapping.

The word-shape incongruency effect also overlapped with the nonword>complete sentence contrast of the semantic localizer in the left SMG and MFG. The SMG has been implicated in phonological processing of both auditorily- and visually-presented language stimuli (Hartwigsen et al., 2010; Oberhuber et al., 2016; Price et al., 1997; Wilson et al., 2011). Although presented in different modalities, both the pseudowords in the experiment and the nonwords in the semantic localizer are pronounceable but senseless words, so the finding of recruitment of phonological regions for these unfamiliar language-like stimuli makes sense. In addition to the multisensory learning described in the section above, the LMFG site of overlap could reflect involvement of Duncan's frontally focused 'multiple demand system' (Duncan, 2013; Fedorenko, Duncan, & Kanwisher, 2012; Fedorenko et al., 2013), which could support the more effortful processing of the pseudowords (in the semantic localizer) and the crossmodally-incongruent AV stimuli relative to their contrasting conditions, which likely require less processing effort.

Because our pseudoword stimuli resemble actual words,wehad expected that the pseudoword-shape mapping would engage a semantic system at least partially overlapping with classic language areas (left perisylvian regions including inferior frontal, inferior parietal, superior temporal regions). However, there was a lack of overlap between our incongruency effect and the semantic localizer, suggesting the phenomenon likely has its basis outside the language system. The pseudowords in this experiment are marked by differences in sound structure that are paralleled by non-linguistic auditory stimuli, which could provide a basis for the mapping. For example, previous research has found that square wave tones, which have a jarring, disrupted quality (much like the stop consonants in keekay), are associated with jagged shapes, whereas the sine wave tones which have a smooth, tonal quality (much like the sonorant consonants in lohmoh), are associated with rounded shapes (Parise & Spence, 2012). It could be that sound-symbolic language directly engages these perceptual systems, without direct contact with the canonical language system. If this is the case, the pseudowords keekay and lohmoh could be expected to bring online representations that are largely overlapping with multisensory binding system (implicated in integrating meaningful sounds from the environment) in the brain, as would the pairing of the word lohmoh or the sine wave tone with rounded or pointed shapes. However, it could be that sound-to-meaning mappings in language are supported by very different mechanisms. One possibility is that we associate the word lohmoh with roundness because we are familiar with the rounded feeling or appearance of the mouth when articulating these sounds (Maurer et al., 2006; Ramachandran & Hubbard, 2001). Thus, it could be that a motor representation of orofacial posture is brought online when we hear sounds in language that we know how to pronounce.

Limitations and future directions

The block design of the word-shape experiment raises a few issues. Because BOLD activity is averaged across an entire block of ten trials, we are unable to examine the modulation of the brain responses to each trial. In addition, the homogeneous nature of stimuli within a block (all congruent or incongruent) could have suppressed extant BOLD effects. Because back-to-back repetition of stimuli of the same condition can suppress the magnitude of BOLD response in voxels, the block design of this experiment may have reduced the magnitude of the observed BOLD signal (Grill-Spector, Henson, & Martin, 2006; Grill-Spector & Malach, 2001; Sawamura & Orban, 2006).

In addition to the limitations posed by the block design of our study, our small set of stimuli limits our ability to test for different mechanisms underlying mappings of vowels and consonants. Nielsen and Rendall (2011) posited one such hybrid account of the word-shape mappings, suggesting that vowels may drive the matching to rounded forms and consonants may drive the matching to jagged shapes. If it is the case that different attributes of the pseudowords (e.g., vowels and consonants) drive the mappings of pseudowords to rounded and pointed shapes, these may be supported by different neural mechanisms, which we are unable to resolve with the present experiment design. To test this, it would be ideal to use a richer set of stimuli, varying in their degree of rounded or pointedness and to be able to extract event-related activity at the trial-level in order to disambiguate systems underlying rounded and jagged trials.

Another consideration concerning the association of our stimuli is that auditory and visual stimuli used may not have been equally distinguishable. In behavioral tasks both in- and outside the scanner, participants were faster to respond to the visual stimuli compared to auditory stimuli. Previous research on crossmodal correspondences has suggested that perceptual dimensions are more likely to interact when processing latencies are similar for both dimensions (Ben-Artzi & Marks, 1995). So although there is clearly a bias in terms of how these dimensions are associated both within our participants (performance on the IAT) and across the greater population (bouba and kiki are matched to shapes across diverse populations), it may not be the early and automatic dimensional association we see in other dimensional interactions.

Neuroscientific research on sound-symbolic mappings is still in its infancy. Sound-symbolic mappings such as the one examined in this study are believed to be privileged in early language learning (Perry et al., 2015, 2017), and, at least in some cases appear to be scaffolded on widely shared crossmodal mappings (Spector & Maurer, 2009). Given this evidence, sound-symbolic language could be useful for language rehabilitation as well. The finding that BOLD activity for the sound-symbolic pseudoword-shape mapping is distinct from the canonical language system indicates a possible therapeutic application in the event that language becomes impaired. Because these systems are dissociated, either channel could be exploited as a means of tapping into meaningful representations if the other channel showed a deficit. New information can be applied to improve therapies for patient populations who are impaired in their ability to integrate certain sensory information. When language is impaired, as is often the case with stroke patients, other systems may offer a means for accessing meaningful representations. Much in the same way that hearing a tool sound can remind patients with apraxia how to use a given tool (Hanna-Pladdy, 2012; Worthington, 2016), there is preliminary evidence to suggest that hearing a sound-symbolic word may help aphasics access meaning in a way that they could not for more arbitrary language (Meteyard, Stoppard, Snudden, Cappa, & Vigliocco, 2015).

The results of the present study may help explain how sound-symbolic words may be able to tap into routes of communication that are spared when much of language has been disrupted (Bieńkiewicz et al., 2012; Worthington, 2016). Kita (1997) has argued that it may be possible to use arbitrary language and conceptualize meanings without bringing online the full, grounded meanings of words, but has proposed that mimetic terms provide the listener a firsthand sensory/affective experience such that listeners (at least native speakers) feel the meanings of ideophonic words (see also Dingemanse, 2011a, 2011b). If this is the case, sound-symbolic language could be an effective way to rehabilitate language or even communicate with individuals who have impaired language comprehension but spared sensory-receptive capacities.

Language research will benefit from improved understanding of how sounds evoke meaningful representations in the brain. The ability to use language is a powerful skill, which confers selective advantages upon humans and in many ways sets us apart from other animals. Although sound-symbolic words may not make up a majority of our language, understanding them is important for our understanding of how sound can encode and communicate information about a range of sensory and affective experiences which would be an important insight into of verbal communication. Many researchers believe that non-arbitrary mappings similar to the kiki-bouba phenomenon may have been an early step in the evolution of language. In a sense, sound-symbolic words could be seen as representing a first step in symbolic displacement, a defining feature of symbolic language. In symbolic displacement, a symbol stands in, as a semantic place-holder for something that it is not, and may be a more readily manipulable handle for meaning than the original referent. In the case of the pseudoword-shape mapping, sounds stand in for object shape, which can be experienced in several other senses. By using sounds that are inherently connected to the other senses, sound-symbolic words could be effective for communicating information without prior experience or learning necessary.

Conclusion

The present study provides a window into sound symbolism in the brain and allows us to examine the extent to which several neural systems support sound to meaning mappings in language. To my knowledge, this is the first within-subject comparison of neural responses for sound-symbolic stimuli and systems underlying magnitude, synchrony, and semantic processing. The present findings provide preliminary evidence for sound-symbolic language having a basis outside the classic language system, which has implications for our understanding of a possible role for sound-symbolic language in language learning. These findings serve to clarify the role of multiple systems in, and the neural and psychological organization of, the cross-sensory association and multisensory processing system. This study complements findings from research on sound symbolism in natural languages and provides us with neuroscientific evidence as to the possible accounts for sound symbolism. Further work is needed to understand how sound-symbolic language may be tapping into grounded cognitive processes.

In addition to elucidating the bouba/kiki effect, specifically, this study provides a framework for exploring many of the sound-symbolic correspondences found in language. By systematically modulating and carefully controlling the structural characteristics of language and other perceptual stimuli, we can gain insight into how specific attributes of sound map to different perceptual dimensions, or encode meaningful representations in the brain.

Chapter 4. Crossmodal association of auditory pitch and visual object size.

Extensive research finds that individuals reliably associate auditory pitch and visual object size, with relatively higher-pitched sounds associated with smaller objects and lower-pitched sounds associated with relatively larger objects (Bonetti & Costa, 2017; Evans & Treisman, 2010; Gallace & Spence, 2006; Marks, Hammeal, Bornstein, & Smith, 1987; Boyle & Tarte, 1980; Ohala, 1997; Parise & Spence, 2009; Walker, Walker, & Francis, 2012; Walker & Smith, 1984; 1985, but see Eitan et al., 2014 and Krugliak & Noppeney, 2015).

The pitch-size mapping has been demonstrated across a number of behavioral studies using a range of paradigms. Several of these studies have employed speeded classification (Bonetti & Costa, 2017; Evans & Treisman, 2010; Gallace & Spence, 2006) and implicit association (Parise & Spence, 2012) paradigms to examine the dimensional interaction between pitch and size mapping as well as the direction of the perceptual influences between the two dimensions. Bonetti and Costa (2017) tested speeded classification for tones in a range of auditory frequencies presented with a range of size stimuli and found that auditory pitch is inversely mapped to size (e.g. high-pitched tones were classified most quickly and accurately when presented with small visual objects). In another study, Walker, Walker, & Francis (2012) used an explicit task, asking individuals to rate the similarity of various stimuli on a number of perceptual dimensions (Walker et al., 2012), and found a robust association of high and low pitches with small and large sizes, respectively.

Mondloch and Maurer (2004) found that 3-year-old children demonstrate the mapping, matching high-pitched sounds to images of a small ball, and low-pitched sounds to a large ball. The finding that even young children exhibit the mapping led Mondloch and Maurer to suggest that it may have its basis in innate wiring or early perceptual-cognitive experience (Mondloch & Maurer, 2004). More recently, Fernández-Prieto, Navarra, & Pons (2015) reported evidence for the pitch-size correspondence in infants as young as 6 months of age, but did not find the pattern in younger infants, suggesting that experience in the first months of life may be critical for establishing the mapping (Fernández-Prieto et al., 2015).

A form of the pitch-size correspondence also appears in sound-symbolic language, with high front vowel sounds (which are higher-pitched) associated with small meanings, and low back vowel sounds (which are lower-pitched) associated with larger meanings (Berlin, 1994; Ohala, 1983, 1994; Tarte, 1974; Tarte & Barritt, 1971; Thompson & Estes, 2011). This correspondence has been documented in studies examining sound patterns in natural languages as well as studies employing non-words, and ratings of individual vowel sounds. The phenomenon was first systematically studied by Sapir (1929), who asked listeners to match pseudowords mil and mal, with images of two tables, one large and the other small. Thus, this matching task examined how the terms, which contrasted only in a single vowel were matched to referents which differed only in size (Sapir, 1929). Participants favored pseudowords containing the /a/ sound to represent the large table and words containing the /i/ sound to represent the small table. Shortly thereafter, Newman (1933) conducted a comprehensive study, presenting hundreds of individuals with pairs of nonsense words and asking them to match the words to size-related meanings. Expanding on the preliminary findings from Sapir (1929), Newman found that both height and backness of vowel articulation were correlated with mappings to size-related meanings. Words containing higher/front vowel sounds (i.e. /i/) were most consistently matched to small meanings, followed by words containing vowels such as /e/ and /ɛ/, whereas the lower vowel sounds (i.e. /a/, /u/, /o/ in this respective order) tended to be matched to large meanings (Newman, 1933). Thompson and Estes (2011) followed up on this research, employing a large set of three-syllable pseudowords to determine how vowel and consonant content systematically related to matching to size-related meanings. They found evidence that sound-symbolic pitch-size mappings can be graded, with pseudowords containing the most high-pitched vowels matched with the smallest of a range of novel images, and increasingly matched to larger images as a pseudoword contained more low-pitched vowels. Since the foundational studies in sound symbolism, which tended to employ pseudowords, the vowel-size mapping has also been attested in a number of natural languages and across diverse cultures, the most consistent finding being that high front vowels (e.g., /i/ and /e/) more often appear in terms with small/diminutive meanings, whereas low back vowels (e.g., /o/ and /a/) more often appear in terms with large/augmentative meanings (Berlin, 1994; Ohala, 1983, 1984; Tarte & Barritt, 1971 but see Diffolth, 1994). As with the non-linguistic form of the pitch-size correspondence, young children are sensitive to the sound-symbolic vowel-size correspondence in spoken language with researchers finding evidence of the vowel-size mapping in infants as young as 4 months (Peña et al., 2011).

As was the case for the crossmodal mappings examined in previous chapters, there are diverse theories as to the possible origins and underlying mechanisms that could give rise to the pitch-size correspondence. These various accounts make differing predictions about the neural systems likely to be involved in the pitch-size mapping. In Chapter 2,weoffered a detailed review and discussed numerous functional loci we could expect to play a role in processing under the multisensory, magnitude, and semantic accounts of crossmodal mappings. For the pitch-size mapping, we had the same overall predictions about potential neural mechanisms, but considered this one the most likely of the three mappings examined to have a basis in multisensory statistical learning. Researchers have long theorized that crossmodal correspondences in the environment may be a basis for the pitch-size mapping since pitch and size are correlated across perceptual experience in the natural world with larger objects resonating at lower frequencies than small objects, and larger animals of a given species producing lower frequency calls or sounds than smaller animals (Ohala, 1984, 1994) Thus, many believe the correspondence of pitch and size is a form of intersensory redundancy processing of signals that would be likely to originate from the same source in the environment (Evans & Treisman, 2010; L. Walker et al., 2012). In his ‘frequency coding hypothesis’ Ohala theorized that it was the audiovisual coincidence of stimuli that led to the association of pitch and object size (1994). If the pitch-size association is a form of perceptual learning based on statistical experience in the multisensory environment, we could expect to find engagement of the multisensory temporal synchrony system or other multisensory integrative systems described in detail in Chapter 2.

Investigators studying the vowel-size variant of the pitch-size association have suggested that correlated multisensory experience could also explain the sound-symbolic association of particular vowels and word meanings related to object size (Ozturk et al., 2013). One such correlational account focuses on the embodied sensory-motor experience of speech sound production, suggesting that it may be the shape of the mouth and constriction of the vocal tract as these sounds are produced that leads individuals to associate particular speech sounds with object size (Newman, 1933; Ramachandran & Hubbard, 2001; Sapir, 1929). Researchers have noted that words or pseudowords associated with small concepts tend to contain high front vowels, which are produced by constricting the oral cavity by raising the tongue and narrowing the vocal tract, and in the case of unround vowels such as /i/ and /e/, pulling the lips taught. In contrast, vowels associated with largeness are produced with the tongue low and the mouth open wider (Newman, 1933; Ramachandran & Hubbard, 2001; Sapir, 1929). Ramachandran & Hubbard (2001) termed this type of crossmodal mapping between mouth shape and corresponding speech sounds 'Synkinesia' and propose that such a mapping would likely have a neural basis in sensorimotor regions and multisensory convergence zones including the angular gyrus.

There is evidence indicating that pitch-size correspondence can modulate integration of audiovisual stimuli. For example, Parise and Spence (2008) asked participants to judge the temporal order of presentation visual stimuli that were immediately preceded or followed by tones. They found that synesthetically congruent couplings of auditory pitch and visual size produced a more robust effect of perceptual unity. That is, congruently pairing pitch and size stimuli led to a stronger auditory capture effect, perceptually pulling apart visual stimuli presented in rapid succession, and increasing participants' sensitivity to the temporal order in which the stimuli were presented. However, in another study, Parise and Spence (2009) show that this perceptual unity effect can also blunt perceptual sensitivity. In this experiment, auditory and visual stimuli were presented either synchronously or asynchronously, with varying lags in onset and participants were asked to report which of the two stimuli was presented second. In this case, participants were more sensitive (exhibiting a smaller just-noticeable-difference) when pitch-size stimulus pairings were synesthetically incongruent than when they were congruent. These findings of enhanced integration and perceptual unity for congruent pitch-size stimuli indicate a functional connection between pitch-size processing and systems involved in temporal binding of audiovisual stimuli. Building on these findings, Bien at al. (2012) reported that applying TMS to right parietal regions in the vicinity of the right IPS could disrupt the (typically enhanced) integration of congruent pitch-size stimuli thereby eliminating the temporal ventriloquism effect. Together, these studies indicate a functional link between multisensory binding systems (and in particular a right parietal contribution) and pitch-size congruency processing. While the precise nature of this functional interplay is as yet unclear, these findings offer important clues as to the neural basis of pitch-size congruency processing.

Method

Participants

A subset of the participants (N=10) that were tested on the pitch-elevation dimensions were also tested on the coupling of pitch and visual object size. One participant was excluded from analyses due to excessive movement during scanning (>1.5mm), leaving a total of 9 participants (four male, five female) in our dataset (mean age 25.7 years, range 20-33 years).

Stimuli

We created two sets of stimuli (one auditory, one visual), which contrasted along perceptual dimensions of auditory pitch and visual object size, respectively. Auditory stimuli consisted of two tones, one low-pitched (180 Hz), and one high-pitched (1440 Hz), and visual stimuli consisted of two gray circles, one large (diameter=170 pixels, subtending ~2.4° of visual angle) and one small (diameter=34 pixels, subtending ~0.5° of visual angle). These basic stimuli were used in computer-based behavioral testing (outside the scanner) and were combined to create multisensory (audio-visual) stimuli used in the neuroimaging experiment. Auditory and visual stimuli were combined to create audiovisual triplets, comprising three repetitions of identical stimuli (200ms on, 200 ms off) presented over the course of 1000 milliseconds (Fig. 28B). Multisensory stimulus pairings were either crossmodally congruent (high pitch+small or low pitch+large) or incongruent (high pitch+large, low pitch+small) with respect to the crossmodal pitch-size correspondence (see Fig. 28A). For each participant, the auditory stimuli were those they matched for apparent loudness for the pitch-elevation experiment (loudness matching procedure described in Chapter 2).

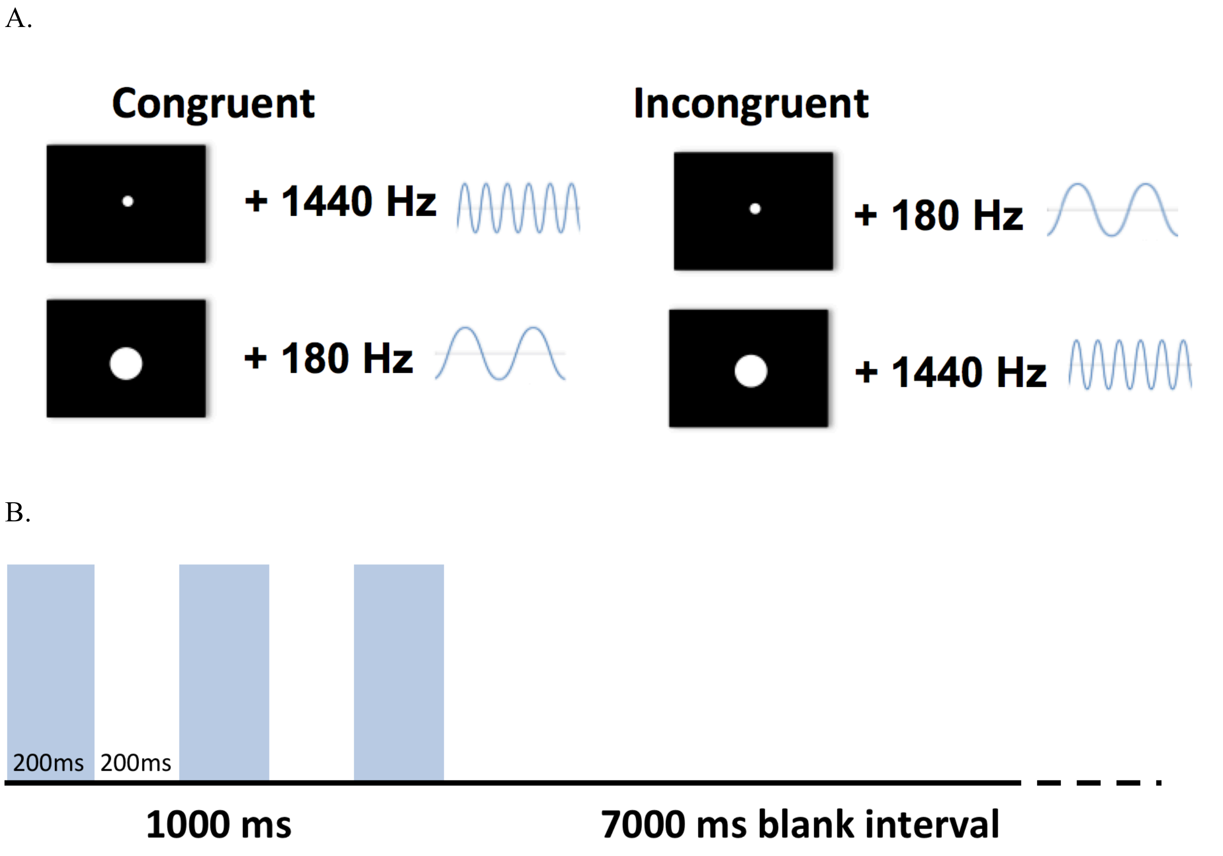


Figure 28. *Trial structure in the Pitch-Size experiment. Visual and auditory stimuli were paired in congruent and incongruent couplings(A) Audio-visual stimuli were presented in a triplet pulse over the course of 1000 milliseconds (B).*

Procedure

Functional localizer tasks. Methods for conducting the three functional localizer experiments were as described for Experiment 1 (pitch-elevation).

Pitch-size fMRI task. For the in-scanner task, participants engaged in a one-back repetition recognition task as described for pitch-elevation but using the pitch-size stimuli described above.

Post-scan behavioral testing: Implicit association of pitch and size. Methods for the IAT were identical to those described for pitch-elevation except the visual stimuli were large/small circles centered on the screen. As discussed above, a number of previous studies have employed speeded classification and implicit association paradigms to demonstrate an interaction between pitch and object size processing. In such studies, the pairing of congruent or compatible stimuli results in faster response times than the pairing of incongruent or incompatible stimuli. In these studies, an association or bias in processing as reflected by faster response times of high-pitched tones when coupled with smaller visual objects and low-pitched tones with larger visual objects (Parise & Spence, 2012). Such congruency effects can obtain either when presenting multisensory couplings of stimuli that are synesthetically congruent, but also when using the same response keys to make responses about unimodally presented auditory and visual stimuli that are in some way compatible.

Results

Behavioral

In–scanner tasks.

Localizer tasks. Analysis of functional localizer data was as described for Experiment 1, but was restricted to the subset of 9 subjects who participated in the pitch-size experiment.

Pitch-size task. To prepare accuracy and reaction time data for analysis, we first excluded trials for which there was no response. This included initial trials in each test block, for which no response was expected (n=144, 10.0% of all trials) and trials for which a response was expected but was not made (n=8, 0.6% of all trials). The remaining dataset (all trials for which there was a response) was used to calculate overall accuracy (correct/correct+incorrect). Additional analyses decomposed the dataset to compare accuracy for i) task conditions: same versus different trials and ii) stimulus conditions: congruent versus incongruent trials. To prepare reaction time data for analysis, we then filtered incorrect responses (n=60, 4.7 % of responses). We then trimmed outliers from the remaining dataset (comprising 95.3% of responses) by calculating subjects' mean response times, then trimming responses with latencies in excess of ±2.5 standard deviations from each subjects' mean. This resulted in the exclusion of 38 responses or 3.2% of the correct trials (mean 4.2 responses trimmed per subject, range of 2-7 trials per subject). With the trimmed dataset, we calculated mean RTs for congruent and incongruent conditions for each subject.

We then analyzed the dataset as described for Experiment 1 to determine whether audiovisual congruency of trials affected in-scanner task performance (either in terms of overall accuracy or response time latency (RT)). The difference in accuracy for congruent trials (94.6±1.9%,) versus incongruent trials (96.1±2.3%) approached, but did not attain, significance (paired samples t test; t_8 =-2.077, *p*=0.071, two-tailed; Fig. 29). Although there was not a significant effect of congruency on accuracy at the group level, individuals exhibited differences in patterns of responses across the two conditions. Six subjects were more accurate in the incongruent condition, while one was more accurate in the incongruent condition, and two were equally accurate in the two conditions (Fig. 30).

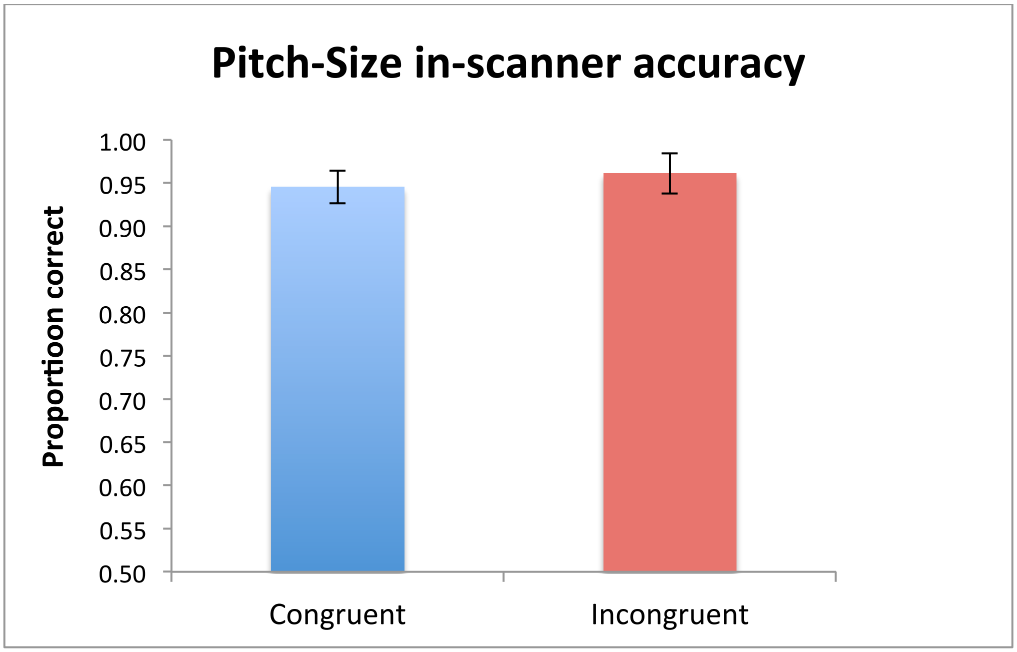


Figure 29. Mean accuracy on in-scanner pitch-size task by congruency of key mapping. Error bars = standard error.

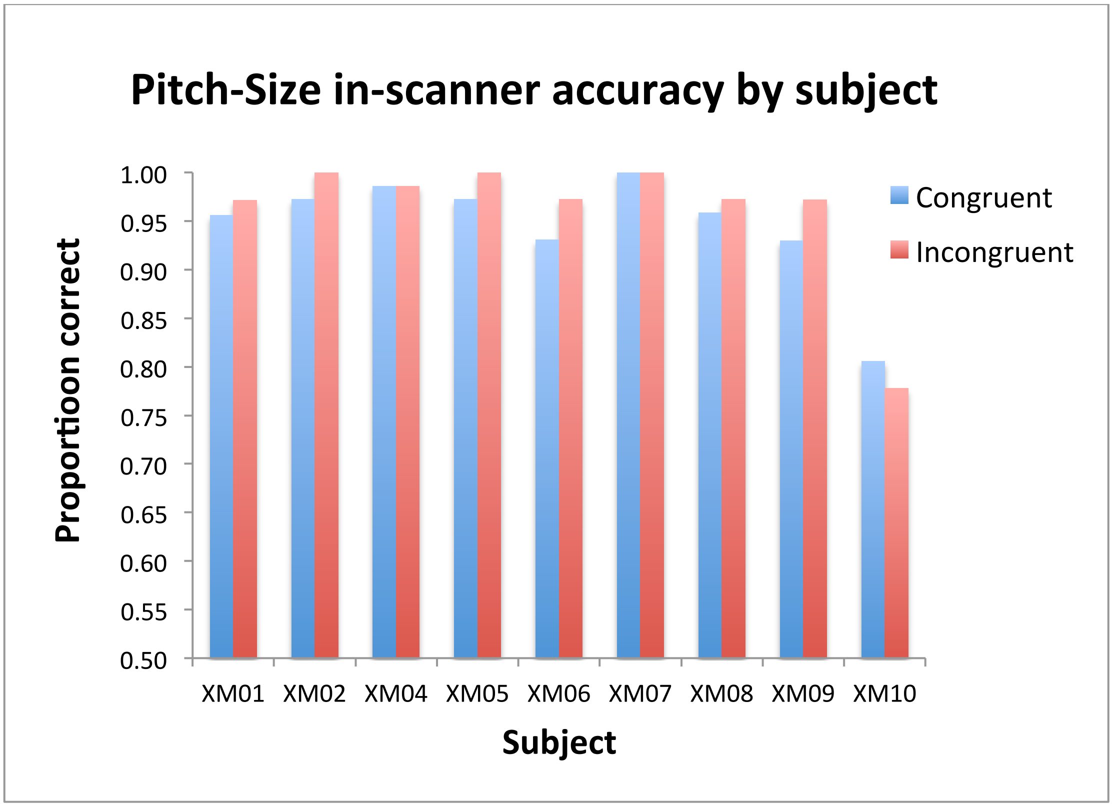


Figure 30. Mean accuracy on in-scanner pitch-size task by congruency of multisensory stimuli for each subject.

A within-subject paired-samples t-test revealed a marginally significant difference in RTs for congruent (1224±114ms) and incongruent task conditions (1200±120ms) (paired samples t test; t_8 = 2.199, *p*=.059). Eight of the subjects had faster RTs for the incongruent condition and one subject was faster for the congruent condition (Fig. 31).

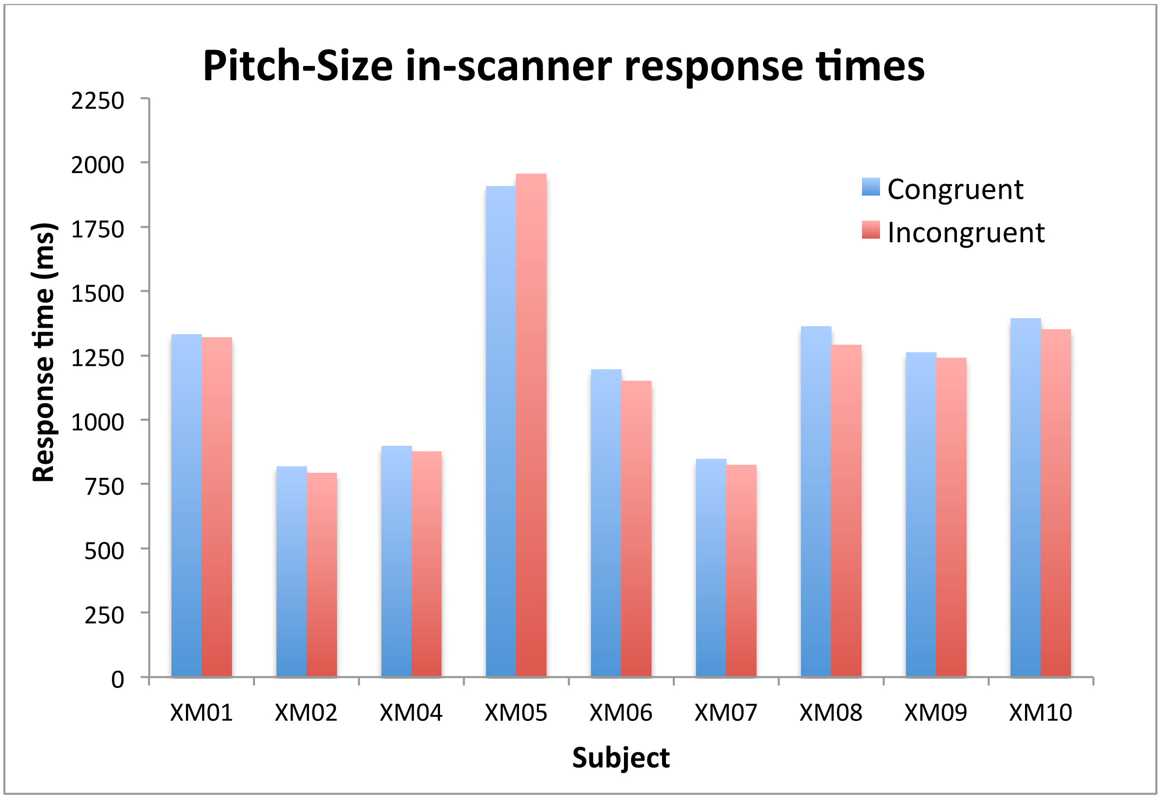


Figure 31. Mean response time on in-scanner pitch-size task by congruency of multisensory stimuli for each subject.

Post-scan pitch-size IAT. A logging error resulted in a total of 11 trials not being logged across 4 subjects (from 0 to 7 trials per subject were not logged). Overall accuracy on the task was 96.1±0.7%. A 2x2 (modality x key mapping congruency) repeated-measures ANOVA (RM-ANOVA) did not reveal significant main effects of stimulus modality nor congruency of key mapping on response accuracy. Accuracy was not significantly different for the pitch stimuli (96.3±1.0%) than the size stimuli (95.8±0.8%: $F_{1,8}$ =.224, p = .649) nor when response key mappings were congruent (96.1 ±0.8%) compared to when they were incongruent (96.0±0.8%: $F_{1,8}$ = .026, p = .876) (Fig. 32). Examining response patterns on an individual basis, five subjects were more accurate in the congruent condition, one was equally accurate in both conditions, and three were more accurate in the incongruent condition (Fig. 33).

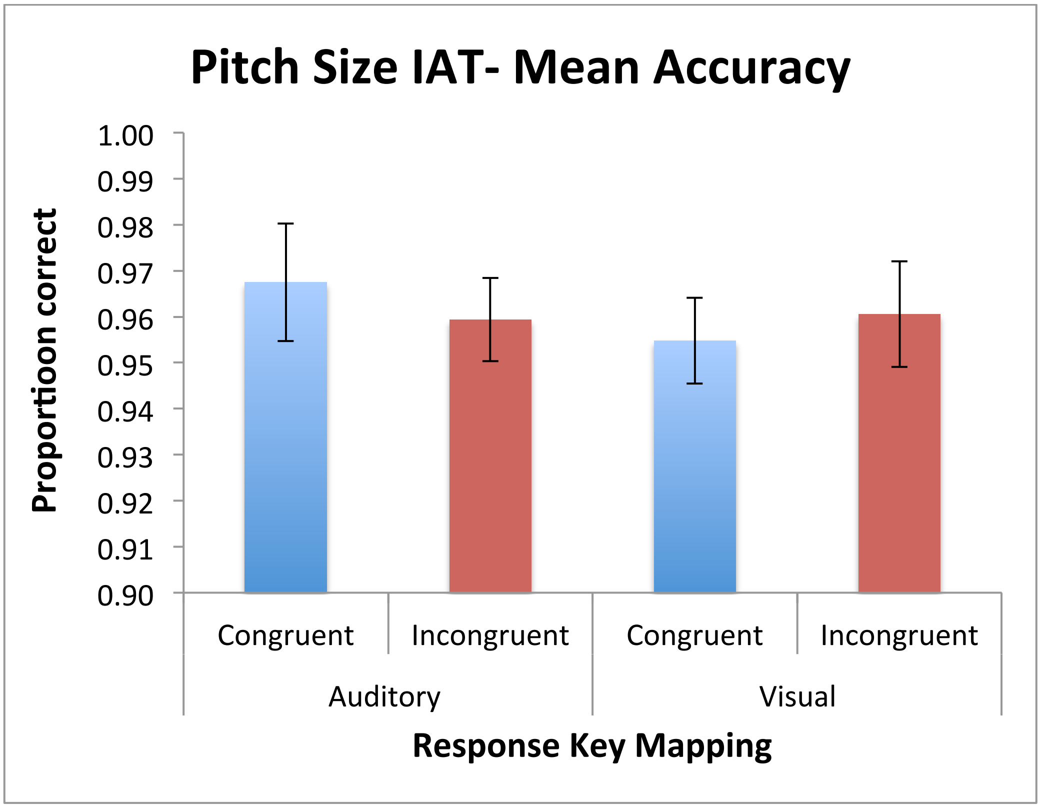


Figure 32. Mean accuracy on pitch-size IAT by congruency of key mapping. Error bars = standard error.

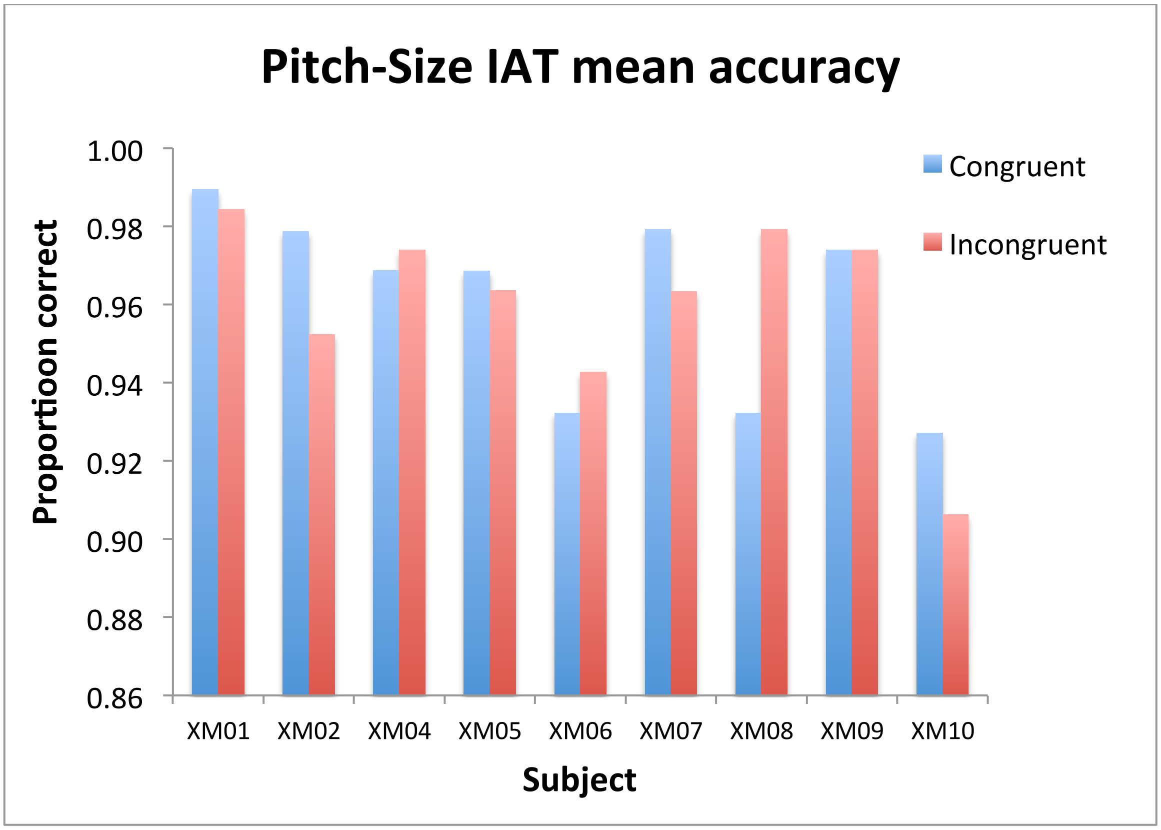


Figure 33. Mean accuracy on pitch-size IAT by congruency of key mapping for each subject.

To prepare reaction time data for analysis, we excluded incorrect responses (3.9% of responses) and trimmed outliers. We calculated subjects' mean response times for auditory and visual conditions separately, trimming responses with latencies in excess of 2.5 SDs from each subjects' mean. Overall, this resulted in 2.5% (n=82) of correct responses being trimmed due to excessive latencies (mean 9.1 responses trimmed per subject, range of 6-14). Mean response times for auditory and visual stimuli were then calculated for each subject using the trimmed dataset. A 2x2 (modality x key mapping congruency) RM-ANOVA compared RTs for trials and demonstrated a main effect of stimulus modality but not key mapping congruency. Participants responded more quickly for the visual stimuli (524±24ms) than the auditory stimuli (614±27ms: $F_{1,8} = 34.60$, $p < .001$), but were not significantly faster when response key mappings were congruent (552±19ms) than when they were incongruent (586±37ms: $F_{1,8} = 1.04$, $p = .337$; Fig. 34). This was somewhat surprising given the robust interactions between dimensions of pitch and size that have been demonstrated in previous studies (Bonetti & Costa, 2017; Evans & Treisman, 2010; Gallace & Spence, 2006; Marks, Ben-Artzi, & Lakatos, 2003; Parise & Spence, 2008, 2009, 2012; Spence & Parise, 2012).

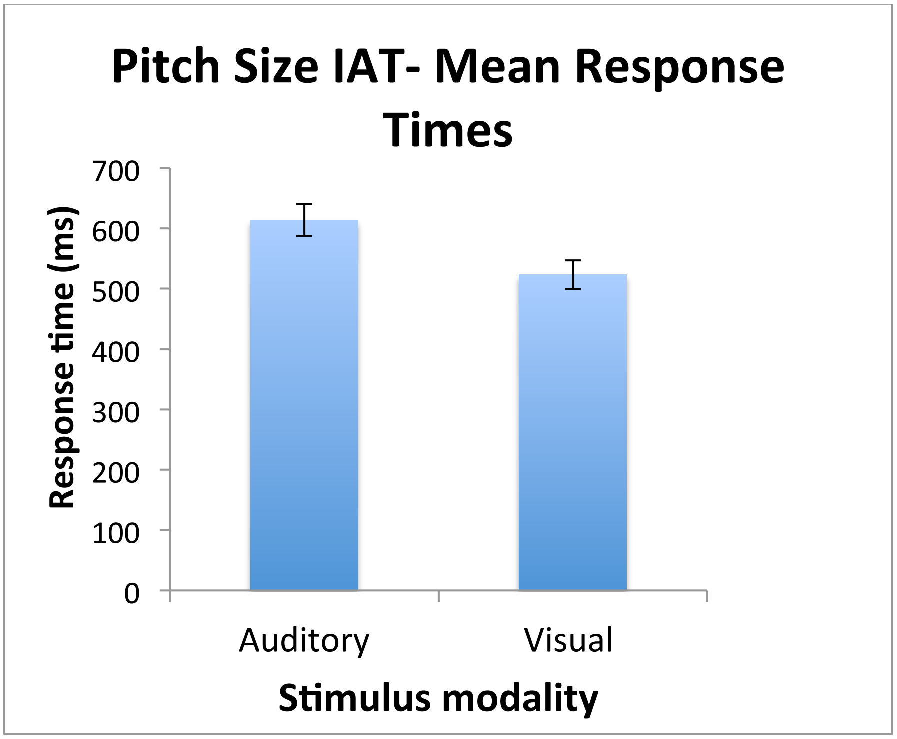


Figure 34. Mean response times on pitch-size IAT by congruency of key mapping. Error bars = standard error.

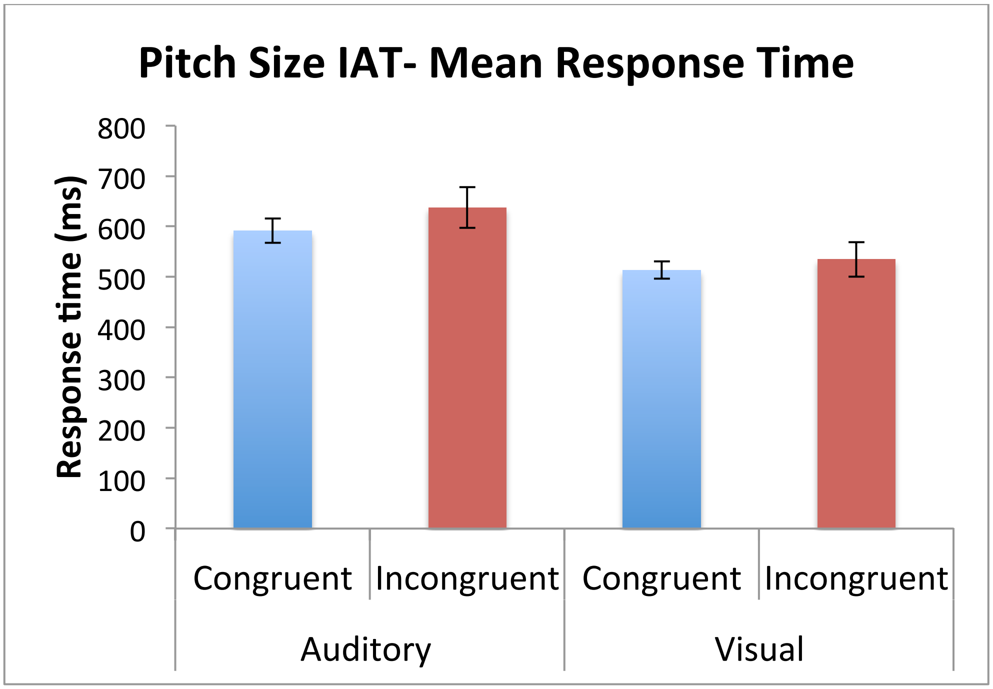


Figure 35. Mean response time on pitch-size IAT by congruency of key mapping. Error bars = standard error.

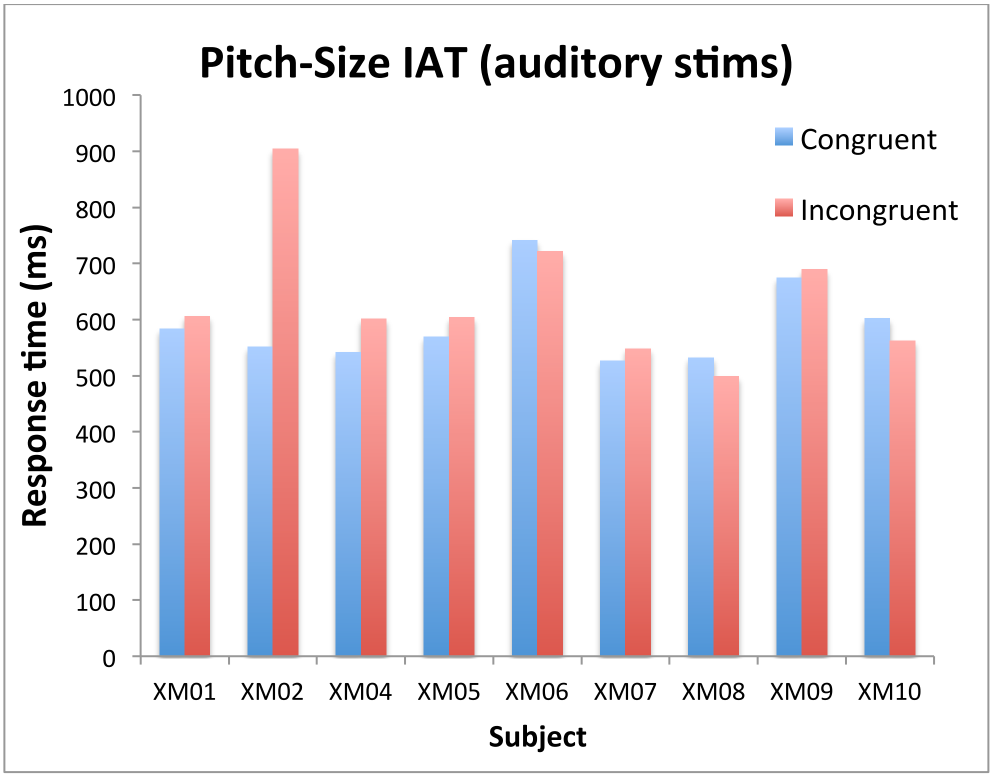


Figure 36. Mean response time for auditory stimuli in the pitch-size IAT by congruency of key mapping for each subject.

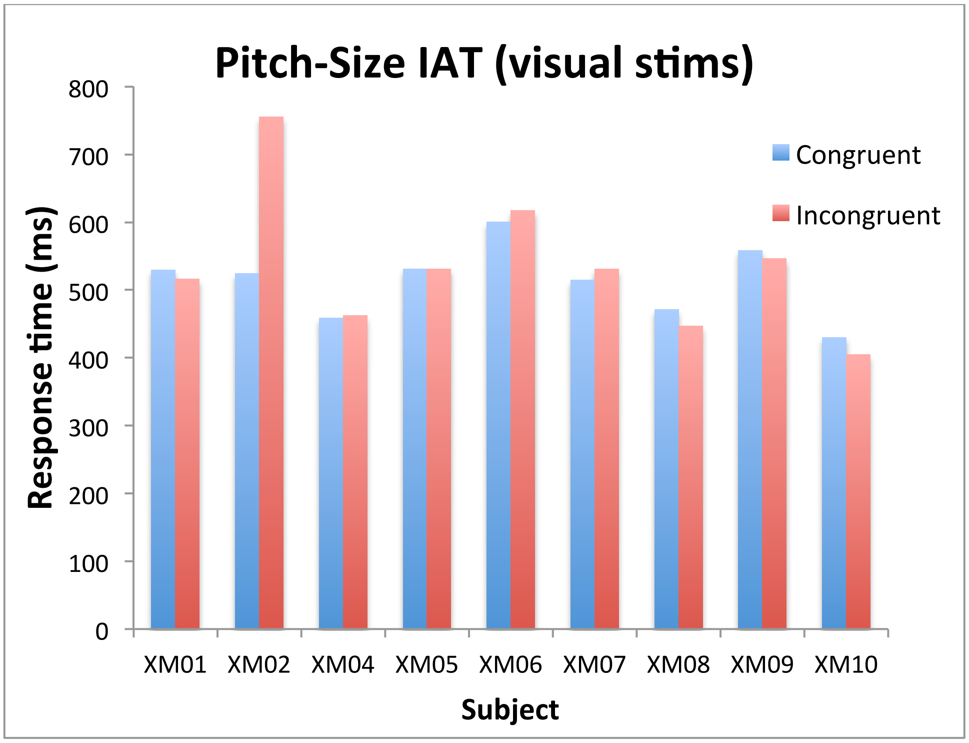


Figure 37. Mean response time for visual stimuli in the pitch-size IAT by congruency of key mapping for each subject.

We had expected that response patterns would be consistent with previous research, and that participants would show an RT congruency effect (responding more quickly when using the same key to make responses for synesthetically congruent stimuli than for incongruent stimuli). Although there was not an effect at the group level, RTs trended in the expected direction (Fig. 35). For the auditory condition, six out of nine subjects responded more quickly for the congruent key-mappings compared to incongruent key-mappings of the pitch-size IAT (Fig. 36). However, only four out of nine subjects responded more quickly for congruent mappings in the visual condition (Fig. 37). Whereas most subjects exhibited small to moderate differences between conditions, subject 02 showed a pronounced congruency effect (showing a difference of 231ms between congruent and incongruent visual stimuli, and a difference of 352ms for two auditory conditions). Five subjects were more accurate in the congruent condition. Behavioral results for the IAT task are summarized in Table 5.

Table 5. Summary of Pitch-Size IAT data, indicating for which subjects and conditions the anticipated congruency effects obtained (0= predicted pattern not observed, 1=subject exhibited predicted pattern,).

Subject	Accuracy	RT	
	C>I	C<I Auditory	C<I Visual
XM01	1	1	0
XM02	1	1	1
XM04	0	1	1
XM05	1	1	0
XM06	0	0	1
XM07	1	1	1
XM08	0	0	0
XM09	0	1	0
XM10	1	0	0

Imaging

Localizer tasks. Functional localizers were as described for Experiment 1, but analyses were restricted to the subset of 9 individuals who participated in the present study.

Pitch-size task. As was the case for the pitch-elevation experiment, the contrast of BOLD activity for all congruent trials compared to all incongruent trials (C>I) did not reveal any sites of activation. A follow-up analysis isolated trials that were preceded by the same congruency condition, either two congruent trials (CC) or two incongruent trials (II) presented back-to-back. This 'CC>II' contrast produced only negative activity (foci for which II modulated BOLD more than CC). Activations for this contrast were located in the right STS/STG, the right postcentral gyrus (post CG), the left dorsal premotor cortex (PMd) with extension from precentral gyrus (pre CG) to the postcentral sulcus (post CS) and postcentral gyrus, and bilateral activity in the rostral portion of the anterior cingulate.

Overlap between incongruency effect and localizers. The incongruency effect overlapped with all three functional localizers as well as pseudoword condition of the semantic localizer and the asynchronous condition of the temporal synchrony localizer.

Multisensory integration. The pitch-size incongruency effect overlapped with the Synchronous>Asynchronous contrast of the multisensory localizer in several areas, including portions of the right post CG, R STS/STG, and the L PMd extending into the post CG. There was also a single voxel overlap between Asynchronous>Synchronous contrast conditions of the multisensory localizer in the right post CG.

Magnitude. The incongruency effect overlapped with the magnitude>control activation in the right post CG.

Semantic. The incongruency effect overlapped with both conditions of the semantic localizer-sentences>pseudowords in the R STS/STG, and pseudowords>sentences in the L PMd and the R post CG.

Discussion

There was a marginally significant behavioral incongruency effect on response times for the in-scanner task. These results were somewhat surprising in light of previous research reporting faster responses for congruent audiovisual stimulus pairings than incongruent pairings in an array of speeded response paradigms. As noted in Chapter 2, a major difference between these experiments and the in-scanner task is that studies reporting congruency effects typically involved a task in which subjects made responses about attributes of the immediate stimuli (e.g. classifying whether a tone was high or low in pitch). In our task, subjects were asked to compare the stimuli in the immediate trial with the preceding trial. It is possible that the demands of the task may render the multisensory congruency of stimuli less salient than in the classification tasks.

Although the analysis of the neuroimaging data from the pitch-size task did not show a congruency effect overall, the CC>II contrast revealed that when back-to-back trials were of the same condition, there was an incongruency effect. However, the response profile for the IAT was more inconsistent, with congruency effects in both directions across subjects. This heterogeneous response pattern was especially noteworthy in its contrast to the pitch-elevation and pseudoword-shape IAT for which the same subjects had shown a consistent and robust congruency effect (including all 9 for pitch-elevation, and all 6 who also participated in the pseudoword-shape experiment). One possible reason for the apparent reverse mapping observed in some subjects is that some may have perceived stimuli as visually looming. Previous research has shown that when participants interpret stimuli as visually looming, they exhibit a reversed pitch-size mapping (associating high pitch sounds with large visual objects and low pitch with small). Due to the inconsistent behavioral response patterns on the implicit association task (and lack of convergence with performance on in-scanner tasks), we discontinued testing on the Pitch-Size after 10 subjects.

Individual differences

For the in-scanner task, eight out of nine subjects responded more slowly for the congruent condition. Only XM05 responded more quickly for congruent stimuli. Six out of the nine subjects responded more accurately for the incongruent condition compared to the congruent condition. For the IAT, most subjects showed only a small difference between congruent and incongruent key mappings, with 6 out of 9 showing a congruency effect for the auditory stimuli, and 4 out of 9 showing a congruency effect for the visual stimuli. Subject XM02 exhibited dramatic congruency effects on the IAT, responding more quickly in the congruent key mapping condition for both auditory (congruency effect 352ms) and the visual condition (congruency effect of 231ms). This subject also responded more accurately for congruent compared to incongruent key mappings (mean difference 2.7%).

Mechanism

The pitch-size incongruency effect task overlapped with functional activity for all three localizers. However, in the absence of consistent behavioral effects for the in-scanner and IAT tasks, it is difficult to interpret this neural activity. While we cannot be sure as to the nature of the congruency processing in-scanner, it is at least compelling that we find extensive overlap between the pitch-size incongruency activity and the multisensory localizer (including both synchronous and asynchronous conditions). In addition to overlapping extensively with the multisensory localizer for the present study, the incongruency effect colocalized with multisensory integration regions (STG/STS) implicated by substantial previous research (see multisensory review in Chapter 2 for review and discussion of Beauchamp, 2005a; Beauchamp, Lee, et al., 2004; Calvert et al., 2000). In addition, the activation in the right post CG appeared to be located near the site (EEG electrode P4) stimulated in a TMS study by Bien et al. (2012) and found to modulate multisensory binding of pitch-size stimuli (see introduction of this chapter for further discussion of the binding system). Together, the current findings, and previous research provide convergent evidence for the possibility that multisensory binding system could be a basis for the pitch-size mapping.

Conclusion

Further research is needed to resolve or account for the inter-subject differences we found in the behavioral portion of this study.Weexpected the pitch-size mapping to be more consistent across individuals in the study. It may be useful to include additional debriefing questions to discover whether individuals interpreted visual stimuli as looming (or had some other basis for a reverse-mapping of pitch-size congruency). With a more uniform behavioral effect outside the scanner, functional activity for the task would be more interpretable and a meaningful account of the phenomenon could be developed.

Chapter 5. Conclusion

A fundamental question in cognitive science is how incoming information from the senses is woven into the unified fabric of cognitive experience and consciousness (Deroy, Chen, & Spence, 2014; Morsella, Godwin, Jantz, Krieger, & Gazzaley, 2016). The present cohort of experiments provides preliminary evidence as to the neural systems supporting and shaping the merging and interaction of the different senses as well as the representation of conflicting information from different sensory channels. In so doing, these studies provide new insight and expand our understanding of cross-sensory mappings and interactions in the brain.

The experiments for this dissertation were designed with two distinct goals in mind. The first goal was to test three working hypotheses about the neurocognitive mechanisms involved in each of three audio-visual mappings: auditory pitch and visuospatial elevation; auditory pitch and visual object size, between pseudowords and object angularity.Wesought to identify systems involved in these audiovisual mappings by presenting both crossmodally-congruent, and incongruent stimulus couplings and examining response patterns in these areas, our reasoning being that a system involved in aligning auditory and visual stimuli would likely be sensitive to an audiovisual mismatch. Employing functional localizer tasks for each subject,weisolated three systems previously hypothesized to play a role in representing audiovisual congruency, these included regions responding to audiovisual synchrony, magnitude-related processing, and semantic processing. In addition to these functionally-localized regions,weassessed the involvement of known unisensory, multisensory, and supramodal regions. By examining BOLD response profiles for each the three audiovisual mappings, and seeking overlap with systems hypothesized to play a role,weexamined the possible cognitive mechanisms underlying each association. The second goal of these experiments was not ultimately achieved.wehad reasoned that overlap in activity produced by the three crossmodal stimulus pairings could represent part of a core network for representing crossmodal associations. For this reason,wehad initially planned to enroll the same subjects for all three experiments, to allow for a direct, within-subject comparison of results for the three experiments. However, following early data collection for pitch-elevation and pitch-size experiments,wegrew concerned that the nature of the one-back task could be contaminating our congruency-related responses (e.g. if making a 1-back 'same'/'different' judgment engages systems overlapping with audiovisual congruency). For this reason, the research team decided to change the task to a two-way forced-choice with subjects reporting the identity of immediate stimulus, as well as changing to a blocked experiment design. These substantial changes to the format of the experiment preclude direct pooling of the data from three experiments (e.g. we cannot average across the experiments to find a common locus of cross-sensory processing). However, one can still make meaningful inferences by comparing the regions engaged in our three experiments. To my knowledge, this is the first study to compare neural responses to cross-sensory congruency and systems underlying magnitude, synchrony, and semantic processing and to make this comparison within-subjects. These findings serve to clarify the role of multiple systems in, and the neural and psychological organization of, the cross-sensory association and multisensory processing system.

These studies also have implications for our broader understanding of both arbitrary and systematic symbolic representation in the brain. The field currently lacks a productive model for how sound-symbolic language serves to bring online meaningful representations in the brain. The word-shape experiment offers an approach for examining neural systems underlying sound to meaning mappings in language and building on current understanding of how sound-symbolic language may bring online meaningful representations in the brain. Compellingly,wefind evidence that word-shape mapping may have its basis in multisensory attention and temporal synchrony systems. The regions modulated by crossmodal congruency in the word-shape experiment overlapped extensively with the putative multisensory attention and salience system described by Menon and colleagues. This activity profile lends support to the theory from Evans and Treisman (2010), that multisensory congruency effects between the auditory signal and visual features such as size, location, or spatial frequency may reflect the brain boosting saliency of crossmodally-redundant (congruent) inputs as it works to integrate input into a maximally coherent or 'unified' representation (Evans & Treisman, 2010). The results from Experiment 3 provide compelling preliminary evidence that symbolic language may engage this same system, at least in the case of the word-shape mapping. Future research could more directly compare systems for multisensory attention and salience with those recruited in processing sound-symbolic language.

The fact that a sound can symbolically stand-in for sensory attributes in the other senses could be considered a most basic form of abstraction (Gallese & Lakoff, 2005; Johnson, 2013; Santiago, Román, & Ouellet, 2011). More generally, the ability for a stimulus in one sensory modality to bring online a host of sensory and semantic representations is a powerful but poorly understood aspect of cognition that could serve as a foundation for complex cognitive processes such as crossmodal and contextual priming and symbol use. When considered as such, these studies provide a window into how such first-order abstractions can be instantiated in the brain. These cross-sensory mappings are, in many ways, akin to metonymy and semantic chaining, two forms of analogical extension described by metaphor experts (Lakoff & Johnson, 1980). Metonymy is a phenomenon wherein a constituent part of an entity is symbolically used as a stand-in to symbolize a whole. In the case of the word-shape mapping, a speaker iconically conveys information about a part of the entity or experience to stand for the whole thing. The second, related phenomenon is semantic chaining, in which a transitive mapping is applied such that a signifier is related to the signified via a series of semantic links. It is possible that the audiovisual mappings in the current experiments are linked through a similar phenomenon. For example, the word-shape mapping has been conceptualized by Ramachandran and Hubbard (2001) as a form of 'Synkinesia', linking the process of articulating a word with the sensory meaning it embodies.

Regardless of whether sound-symbolic mappings are achieved through articulatory-motor representations, or other means, it is clear that they represent a unique form of language, the neural basis of which could to be distinct from other forms of language processing. Experiment 3 allows us to examine the extent to which activation related to word-shape mapping overlaps with the language system, and the extent to which linguistic and other functional systems (such as temporal synchrony processing) overlap.

For both the pitch-elevation mapping and the word-shape mapping, we found foci of congruency-related activity in the inferior frontal regions. These foci were strikingly close to frontal sites identified by

Hein et al. (2007) which responded to both semantically incongruent AV stimuli, as well as arbitrarily-paired unfamiliar stimuli (8mm from pitch-elevation site in right IFG, 6mm from word-shape site in left MFG). It is widely accepted that frontal activity is strongly modulated by task demands and format (Miller & Cohen, 2001), yet here we find a relatively focal region exhibiting a reliable response to audiovisual incongruency for a range of tasks and diverse stimuli. Our findings extend on research of Hein et al. (2007) and Noppeney et al. (2008). Interestingly, Hein et al. (2007) report that BOLD response for congruent and incongruent stimuli were highly overlapping in classic multisensory binding regions, whereas frontal and parietal systems responded more similarly for incongruent and novel unfamiliar stimuli. This raises the interesting possibility that some systems treat these stimuli as incongruent whereas other systems may simply treat them as 'not congruent' or orthogonally matched. Two parietal regions where Hein et al. (2007) found activity for unfamiliar incongruent stimuli were close to our word-shape incongruency effect. In these foci, the brain may be treating auditory and visual stimuli as either belonging together or novel, and may not be especially burdened or disrupted by the stimuli we designed to be incongruent or mismatched. From a multisensory binding perspective, this response profile makes ecological sense- while some auditory and visual stimuli are matched, a vast majority of incoming signals originate from different and disparate sources and are not readily alignable. Thus, the behavioral congruency effect observed in the out of scanner IAT, could reflect more of a facilitation in processing congruent signals, rather than an interference effect (see limitations section for caveat).

Limitations and future directions

A major limitation of the current studies is that they rely on two different in-scanner tasks, which limits us in our ability to compare findings from the experiments in Chapters 2 and 3. At the outset of these studies, our aim was to directly compare results of the three experiments within-subjects. However, one experiment was terminated, and the changes in both experimental task (one-back to 2-AF forced choice) and design (event-related to block design) to word-shape experiment preclude direct comparison between the two experiments. Thus, future studies are needed if we wish to directly compare the neural basis of different cross-sensory mappings on a within-subject basis.

Unfortunately, the IAT we used to assess the strength (and direction) of crossmodal mappings is unable to disambiguate the extent to which any differences in the two conditions are driven by facilitation in the congruent condition versus interference in the congruent condition, or a combination of the two. Future studies should take care to include an orthogonal or baseline condition of congruency testing, to help determine whether observed congruency effects reflect processing facilitation, interference, or both.

Many relevant questions are beyond the scope of the present experiments. For example, the present study does not address the question as to whether these mappings are learned or innate. There is some suggestion that the parietal lobe activations seen for associative activity in adults reflects processing that gradually emerges over the course of development (Holloway & Ansari, 2010; Nieder & Dehaene, 2009), and that through learning and experience in the world and with cultural training, these more generalized representations mature out of sensory integrative systems. A related question is to what extent intersensory connections are plastic, or able to be shaped and reshaped on the basis of perceptual and cognitive experience or training. Our understanding of functional plasticity can benefit from distinguishing between cross-sensory mappings for which humans are predisposed as opposed to mappings that arise solely on the basis of sensory experience (Seitz et al., 2007; Shams & Seitz, 2008). It is possible that such mappings could be entrained on a relatively short timescale.

The present cohort of studies is also limited in that it does not allow us to generalize findings to individuals from other cultures. There are some hints in the literature that the pitch-elevation mapping might be more culturally idiosyncratic than other mappings (pitch-size- more robust). Future research can examine the neural basis cross-sensory mappings across different developmental stages and across cultures different cultures. Another outstanding question relates to the timecourse and trajectory of processing. With a TR of 2000 ms we do not have the temporal resolution necessary to characterize the trajectory and flow of information with a functional connectivity analysis.

Conclusion

Neuroscientific research on crossmodal mappings is still in its infancy. Identifying how sensory information combines and interacts is a first step in understanding myriad cognitive processes built upon these processes, and in consciousness writ large. By gaining insights into intersensory interactions we can build on current approaches for fluently conveying information, focusing attention, and perhaps modulating other kinds of cognitive processing. The present study offers a systematic empirical framework for examining the neural basis of crossmodal mappings. Future research can readily apply our approach to examine additional crossmodal mappings. By drawing approaches and integrating insights from diverse disciplines we hope that the present studies can inform research in fields ranging from cognitive psychology to sensory neuroscience.

References

Adam, R., & Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *NeuroImage*, *52*(4), 1592–1602. http://doi.org/10.1016/j.neuroimage.2010.05.002

Ahlner, F., & Zlatev, J. (2010). Crossmodal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies*, *38*(1), 298–348.

Aleman, A., & Rutten, G. (2001). Activation of striate cortex in the absence of visual stimulation: an fMRI study of synesthesia.

Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002). Convergence of Visual and Tactile Shape Processing in the Human Lateral Occipital Complex. *Cerebral Cortex*, *12*(11), 1202–1212. http://doi.org/10.1093/cercor/12.11.1202

Amedi, A., von Kriegstein, K., van Atteveldt, N., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, *166*(3–4), 559–71. http://doi.org/10.1007/s00221-005-2396-5

Arata, M., Imai, M., Okuda, J., Okada, H., & Matsuda, T. (2010). Gesture in language : How sound-symbolic words are processed in the brain. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society.* (pp. 1374–1379).

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature Neuroscience*, *6*(2), 115–116. http://doi.org/10.1038/nn1003

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in cognitive sciences*, *8*(4), 170-177.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *18*(4), 177–185. http://doi.org/10.1016/j.tics.2013.12.003

Arrington, C. M., Carr, T. H., Mayer, A. R., & Rao, S. M. (2000). Neural Mechanisms of Visual Attention: Object-Based Selection of a Region in Space. *Journal of Cognitive Neuroscience*, *12*(supplement 2), 106–117. http://doi.org/10.1162/089892900563975

Arsalidou, M., & Taylor, M. J. (2011). Is 2+ 2= 4? Meta-analyses of brain areas needed for numbers and calculations. *Neuroimage*, *54*(3), 2382-2393.

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*, *63*, 196–205. http://doi.org/10.1016/j.cortex.2014.08.025

Atteveldt, N. M. Van, Formisano, E., Blomert, L., & Goebel, R. (2007). The Effect of Temporal Asynchrony on the Multisensory Integration of Letters and Speech Sounds. *Cerebral Cortex*, *17*, 962–974. http://doi.org/10.1093/cercor/bhl007

Aveyard, M. E. (2011). Some consonants sound curvy : Effects of sound symbolism on object recognition. http://doi.org/10.3758/s13421-011-0139-3

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory Redundancy Guides the Development of Selective Attention , Perception, and Cognition in Infancy. *Child Development*, *13*(3), 99–102.

Bankieris, K., & Simner, J. (2014). Sound symbolism in synesthesia: evidence from a lexical-gustatory synesthete. *Neurocase*, *20*(6), 640–51. http://doi.org/10.1080/13554794.2013.826693

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577-609-660.

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. Language and Cognitive Processes, 18(5–6), 513–562. http://doi.org/10.1080/01690960344000026

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. http://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W., Pecher, D., Zeelenberg, R., Simmons, W. K., & Hamann, S. B. (2005). Multimodal simulation in conceptual processing. *Categorization inside and outside the lab: Festschrift in honor of Douglas L. Medin*, 249-270.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12584027

Beauchamp, M. S. (2005a). See me, hear me touch me: multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, *15*(2), 145–153. http://doi.org/10.1016/j.cognition.2010.08.016

Beauchamp, M. S. (2005b). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics*, 93–113. http://doi.org/10.1385/NI

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*(11), 1190–2. http://doi.org/10.1038/nn1333

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), 809–23. http://doi.org/S0896627304000704 /pii/

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(11). http://doi.org/10.1073/pnas.1014818108

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*(8), 1151–1162. http://doi.org/10.3758/BF03208371

Ben-Artzi, E., & Marks, L. E. (1999). Processing linguistic and perceptual dimensions of speech: interactions in speeded classification. *Journal of Experimental Psychology. Human Perception and Performance*, *25*(3), 579–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10385980

Bergen, B. K. (2007). Experimental methods for simulation semantics *. In M. (Eds.). Gonzalez-Marquez, M., Mittleberg, I., Coulson, S., & Spivey (Ed.), *Methods in cognitive linguistics*. John Benjamins.

Bergen, B. K., & Chang, N. (2003). Embodied Construction Grammar in Simulation-Based Language Understanding. In Constructional Approaches to Language 3 (pp. 147–190). http://doi.org/10.1075/cal.3.08ber

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and Linguistic Aspects of Visual Imagery in Sentence Comprehension. Cognitive Science, 31, 733–764. http://doi.org/10.1080/03640210701530748

Berlin, B. (1994). Evidence for pervasive synaesthetic sound symbolism is ethnozoological nomenclature. In L. Hinton, J. Nichols, & J. Ohala (Eds.), Sound Symbolism (pp. 77–93).

Berlin, B., & O'Neill, J. P. (1981). The pervasiveness of onomatopoeia in Aguaruna and Huambisa bird names. Journal of Ethnobiology, 1(2), 238–261.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of Some Variations in Auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*(2), 241–247.

Bien, N., ten Oever, S., Goebel, R., & Sack, A. T. (2012). The sound of size. Crossmodal binding in pitch-size synesthesia: A combined TMS, EEG and psychophysics study. NeuroImage, 59(1), 663–672. http://doi.org/10.1016/j.neuroimage.2011.06.095

Bieńkiewicz, M., Goldenberg, G., Cogollor, J. M., Ferre, M., Hughes, C., & Hermsdörfer, J. (2012). Use of biological motion based cues and ecological sounds in the neurorehabilitation of apraxia. In Proceedings. 2013 IEEE International Conference on Healthcare Informatics ICHI 2013.

Binder, J.R., Desai RH. 2011. The neurobiology of semantic memory. Trends Cogn Sci 15(11):527–36.

Binder JR, Desai RH, Graves WW, Conant LL. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19(12): 2767–96.

Binder, J. R., Westbury, C. F., McKiernan, K. a, Possing, E. T., & Medler, D. a. (2005). Distinct brain systems for processing concrete and abstract concepts. Journal of Cognitive Neuroscience, 17(6), 905–17. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16021798

Binney, R. J., Parker, G. J. M., & Lambon Ralph, M. A. (2012). Convergent Connectivity and Graded Specialization in the Rostral Human Temporal Lobe as Revealed by Diffusion-Weighted Imaging Probabilistic Tractography. *Journal of Cognitive Neuroscience*, *24*(10), 1998–2014. http://doi.org/10.1162/jocn_a_00263

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Sciences, 113(39), 10818–10823. http://doi.org/10.1073/pnas.1605782113

Blazhenkova, O., & Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. Applied Cognitive Psychology, 23(5), 638–663. http://doi.org/10.1002/acp.1473

Bonetti, L., & Costa, M. (2017). Pitch-verticality and pitch-size crossmodal interactions. *Psychology of Music*. http://doi.org/10.1177/0305735617710734

Bottini, R., Barilari, M., & Collignon, O. (2019). Sound symbolism in sighted and blind. The role of vision and orthography in sound-shape correspondences. *Cognition, 185,* 62-70.

Boyle, M. W., & Tarte, R. D. (1980). Implications for phonetic symbolism: The relationship between pure tones and geometric figures. Journal of Psycholinguistic Research, 9, 535–544.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition*, *126*(2), 165–172. http://doi.org/10.1016/j.cognition.2012.09.007

Brunel, L., Carvalho, P. F., & Goldstone, R. L. (2015). It does belong together: crossmodal correspondences influence crossmodal integration during perceptual learning. *Frontiers in Psychology*, *6*(April), 1–10. http://doi.org/10.3389/fpsyg.2015.00358

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1831–1840. http://doi.org/10.1098/rstb.2009.0028

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., & Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. Proceedings of the National Academy of Sciences, 102(51), 18751–18756. http://doi.org/10.1073/pnas.0507704102

Cabrera, D., & Morimoto, M. (2007). Influence of fundamental frequency and source elevation on the vertical localization of complex tones and complex tone pairs. *The Journal of the Acoustical Society of America*, *122*(1), 478-488.

Cai, W., Ryali, S., Chen, T., Li, C.-S. R., & Menon, V. (2014). Dissociable Roles of Right Inferior Frontal Cortex and Anterior Insula in Inhibitory Control: Evidence from Intrinsic and Task-Related Functional Parcellation, Connectivity, and Response Profile Analyses across Multiple Datasets. *Journal of Neuroscience*, *34*(44), 14652–14667. http://doi.org/10.1523/JNEUROSCI.3048-14.2014

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex*, *11*(12), 1110–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11709482

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., … David, A. S. (1997). Activation of Auditory Cortex During Silent Lipreading. *Science*, *276*(5312), 593–596. http://doi.org/10.1126/science.276.5312.593

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol*, *10*, 649–657. http://doi.org/S0960-9822(00)00513-3 /pii/

Calvert, G. A., Spence, C., & Stein, B. E. (Eds.). (2004). *The handbook of multisensory processing*. Cambridge, MA: MIT Press.

Cantlon, J. F., Brannon, E. M., Carter, E. J., & Pelphrey, K. a. (2006). Functional imaging of numerical processing in adults and 4-y-old children. *PLoS Biology*, *4*(5), e125. http://doi.org/10.1371/journal.pbio.0040125

Cattaneo, Z., Silvanto, J., Pascual-Leone, A., & Battelli, L. (2009). The role of the angular gyrus in the modulation of visuospatial attention by the mental number line. *NeuroImage*, *44*(2), 563–568. http://doi.org/10.1016/j.neuroimage.2008.09.003

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, *129*(3), 564-583.

Chambers, C. D., Payne, J. M., & Mattingley, J. B. (2007). Parietal disruption impairs reflexive spatial attention within and between sensory modalities. *Neuropsychologia*, *45*(8), 1715–1724. http://doi.org/10.1016/j.neuropsychologia.2007.01.001

Chen, T., Michels, L., Supekar, K., Kochalka, J., Ryali, S., & Menon, V. (2015). Role of the anterior insular cortex in integrative causal signaling during multisensory auditory-visual attention. *European Journal of Neuroscience*, *41*(2), 264–274. http://doi.org/10.1111/ejn.12764

Chiou, R., & Rich, A. N. (2012). Crossmodality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, *41*(3), 339–353. http://doi.org/10.1068/p7161

Cohen Kadosh, R, & Henik, A. (2006). A common representation for semantic and physical properties: A cognitive-anatomical approach. Experimental Psychology, 53(2), 87–94. http://doi.org/10.1027/1618-3169.53.2.87

Cohen Kadosh, R., & Henik, A. (2007). Can synaesthesia research inform cognitive science? *Trends in Cognitive Sciences*, *11*(4), 177–184. http://doi.org/10.1016/j.tics.2007.01.003

Cohen Kadosh, R., Henik, A., Rubinsten, O., Mohr, H., Dori, H., Van De Ven, V., … Linden, D. E. J. (2005). Are numbers special? The comparison systems of the human brain investigated by fMRI. *Neuropsychologia*, *43*(9), 1238–1248. http://doi.org/10.1016/j.neuropsychologia.2004.12.017

Cohen Kadosh, R., Henik, A., & Walsh, V., (2007). Small is bright and big is dark in synaesthesia. *Current Biology*, *17*(19), 834–835. http://doi.org/http://dx.doi.org/10.1016/j.cub.2007.07.048

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience*, *3*(3), 292–7. http://doi.org/10.1038/73009

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron*, *58*(3), 306–324. http://doi.org/10.1016/j.neuron.2008.04.017

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–15. http://doi.org/10.1038/nrn755

Crottaz-Herbette, S., & Menon, V. (2006). Where and when the anterior cingulate cortex modulates attentional response: combined fMRI and ERP evidence. *Journal of Cognitive Neuroscience*, *18*(5), 766–80. http://doi.org/10.1162/jocn.2006.18.5.766

Dahl, C. D., Logothetis, N. K., & Kayser, C. (2009). Spatial organization of multisensory responses in temporal association cortex. *The Journal of Neuroscience*, *29*(38), 11924–11932. http://doi.org/10.1523/jneurosci.3437-09.2009

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, *8*(2–3), 109–14. http://doi.org/10.1002/(SICI)1097-0193(1999)8:2/3&lt;109::AID-HBM7&gt;3.0.CO;2-W

Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, *90*(11), 4957-4960.

Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, *90*(11), 4957-4960.

Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., & Damasio, A. (2004). Neural systems behind word and concept retrieval. *Cognition*, *92*(1-2), 179-229.

Davis, R. (1961a). The fitness of names to drawings. A cross-cultural study in Tanganyika. British Journal of Psychology, 52(3), 259–268. http://doi.org/10.1111/j.2044-8295.1961.tb00788.x

Davis, R. (1961b). The Fitness of names to Drawings. A Cross-Cultural Study in Tanganyika. British Journal of Psychology, 52(3), 259–268.

de Saussure, F. (1916/2009). *Course in General Linguistics*. Open Court Classics: Peru, IL.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398. http://doi.org/10.1016/j.neuron.2007.10.004

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3–6), 487–506. http://doi.org/10.1080/02643290244000239

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: behavioral and brain-imaging evidence. *Science (New York, N.Y.)*, *284*(5416), 970–4. http://doi.org/10.1126/science.284.5416.970

Deroy, O., Chen, Y., & Spence, C. (2014). Multisensory constraints on awareness.

Deshpande, G., Hu, X., Lacey, S., Stilla, R., & Sathian, K. (2010). Object familiarity modulates effective connectivity during haptic shape perception. *Neuroimage*, *49*(3), 1991-2000.

Diffolth, G. (1994). i: big, a: small. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), Sound symbolism (pp. 107–114). Cambridge: Cambridge University Press.

Dingemanse, M. (2011a). Ideophones and the Aesthetics of Everyday Language in a West-African Society. The Senses and Society, 6(1), 77–85. http://doi.org/10.2752/174589311X12893982233830

Dingemanse, M. (2011b). The Meaning and Use of Ideophones in Siwu. PhD dissertation, Nijmegen: Radboud University.

Dingemanse, M. (2012). Advances in the Cross-Linguistic Study of Ideophones. Language and Linguistics Compass, 6(10), 654–672. http://doi.org/10.1002/lnc3.361

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences, 19*(10), 603–615. http://doi.org/10.1016/j.tics.2015.07.013

Dingemanse, M., & Majid, A. (2010). The semantic structure of sensory vocabulary in an African language. Dimension Contemporary German Arts And Letters, 1.

Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2014). Prelinguistic infants are sensitive to space-pitch associations found across cultures. *Psychological Science*, *25*(6), 1256–61. http://doi.org/10.1177/0956797614528521

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2002). A cortical network for the detection of novel events across multiple sensory modalities. *Journal of Neurophysiology*, *87*, 615–620. http://doi.org/10.1016/S1053-8119(01)91653-2

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on “sensory-specific” brain regions, neural responses, and judgments. *Neuron*, *57*(1), 11–23. http://doi.org/10.1016/j.neuron.2007.12.013

Driver, J., & Spence, C. (2000). Multisensory perception : Beyond modularity and convergence. Current Biology, 10–12.

Duncan, J. (2013). The Structure of Cognition: Attentional Episodes in Mind and Brain. *Neuron*, *80*(1), 35–50. http://doi.org/10.1016/j.neuron.2013.09.015

Duvernoy, H.M. (1999). The Human Brain. Surface, Blood Supply and Three-dimensional Sectional Anatomy. New York: Springer.

Eckert, M. A., Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2009). At the heart of the ventral attention system: The right anterior insula. *Human Brain Mapping*, *30*(8), 2530–2541. http://doi.org/10.1002/hbm.20688

Eger, E., Sterzer, P., Russ, M. O., Giraud, A.-L., & Kleinschmidt, A. (2003). A supramodal number representation in human intraparietal cortex. *Neuron*, *37*(4), 719–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12597867

Eitan, Z., Schupak, A., Gotler, A., & Marks, L. E. (2014). Lower pitch is larger, yet falling pitches shrink: Interaction of pitch change and size change in speeded discrimination. *Experimental Psychology*, *61*(4), 273–284. http://doi.org/10.1027/1618-3169/a000246

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: cross-domain mappings of auditory pitch in a musical context. *Cognition*, *114*(3), 405–22. http://doi.org/10.1016/j.cognition.2009.10.013

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905. http://doi.org/10.1073/pnas.1602413113

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–6. http://doi.org/10.1016/j.tics.2004.05.003

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582. http://doi.org/10.1111/j.1551-6709.2009.01023.x

Erickson, L. C., Heeg, E., Rauschecker, J. P., & Turkeltaub, P. E. (2014). An ALE meta-analysis on the audiovisual integration of speech signals. *Human Brain Mapping*, *35*(11), 5587–5605. http://doi.org/10.1002/hbm.22572

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–9. http://doi.org/10.1016/j.tics.2004.02.002

Evans, K. K., & Treisman, A. (2010). Natural crossmodal mappings between visual and auditory features. *Journal of Vision*, *10*, 1–12. http://doi.org/10.1167/10.1.6.Introduction

Farooqui, A. A., Mitchell, D., Thompson, R., & Duncan, J. (2012). Hierarchical Organization of Cognition Reflected in Distributed Frontoparietal Activity. *Journal of Neuroscience*, *32*(48), 17373–17381. http://doi.org/10.1523/JNEUROSCI.0598-12.2012

Fedorenko, E., Behr, M. K., & Kanwisher, N. G. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16428–33. http://doi.org/10.1073/pnas.1112937108

Fedorenko, E., Duncan, J., & Kanwisher, N. G. (2012). Language-selective and domain-general regions lie side by side within Broca's area. Current Biology : CB, 22(21), 2059–62. http://doi.org/10.1016/j.cub.2012.09.011

Fedorenko, E., Duncan, J., & Kanwisher, N. G. (2013). Broad domain generality in focal regions of frontal and parietal cortex. Proceedings of the National Academy of Sciences of the United States of America, 110(41), 16616–21. http://doi.org/10.1073/pnas.1315235110

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. G. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. Journal of Neurophysiology, 104(2), 1177–1194. http://doi.org/10.1152/jn.00032.2010

Fedorenko, E., & Kanwisher, N. G. (2009). Neuroimaging of Language: Why Hasn't a Clearer Picture Emerged? *Language and Linguistics Compass*, *3*(4), 839–865. http://doi.org/10.1111/j.1749-818X.2009.00143.x

Fedorenko, E., & Kanwisher, N. G. (2011). Functionally Localizing Language-Sensitive Regions in Individual Subjects With fMRI: A Reply to Grodzinsky's Critique of Fedorenko and Kanwisher (2009). *Language and Linguistics Compass*, *5*(2), 78–94. http://doi.org/10.1111/j.1749-818X.2010.00264.x

Feldman, J., & Narayanan, S. (2004). Embodied meaning in a neural theory of language. *Brain and language*, *89*(2), 385-392.

Fernández, L.M., Visser, M., Ventura-Campos, N., Ávila, C. & Soto-Faraco, S. (2015). Top-down attention regulates the neural expression of audiovisual integration. NeuroImage, 116:272- 285.

Fernández-Prieto, I., Navarra, J., & Pons, F. (2015). How big is this sound? Crossmodal association between pitch and size in infants. *Infant Behavior and Development*, *38*, 77-81.

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., ... & Seidenberg, M. S. (2015). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, *26*(5), 2018-2034.

Fort, M., Martin, A., & Peperkamp, S. (2014). Consonants are More Important than Vowels in the Bouba-kiki Effect. Language and Speech, 122–140. http://doi.org/10.1177/0023830914534951

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*(7), 1191–203. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17355042

Gallese, V., & Lakoff, G. (2005). The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3–4), 455–479. http://doi.org/10.1080/02643290442000310

Ganis, G., Thompson, W.L. & Kosslyn, S.M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. Cognitive Brain Research, 20:226-241.

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology*, *8*(1), 98–123. http://doi.org/10.1016/0010-0285(76)90006-2

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*(3), 225–241. http://doi.org/10.1016/0010-0285(70)90016-2

Gasser, M. (2004). The Origins of Arbitrariness in Language. In Proceedings of the 26th Annual Conference of the Cognitive Science Society (pp. 1–6).

Ghahremani, A., Rastogi, A., & Lam, S. (2015). The Role of Right Anterior Insula and Salience Processing in Inhibitory Control. *Journal of Neuroscience*, *35*(8), 3291–3292. http://doi.org/10.1523/JNEUROSCI.5239-14.2015

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic bulletin & review*, *9*(3), 558-565.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. Journal of Memory and Language, 43(3), 379–401. http://doi.org/10.1006/jmla.2000.2714

Göbel, S., Walsh, V., & Rushworth, M. F. S. (2001). The Mental Number Line and the Human Angular Gyrus The Mental Number Line and the Human Angular Gyrus. *NeuroImage*, *14*(February), 1278–1289. http://doi.org/10.1006/nimg.2001.0927

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*(2–3), 231–62. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9557384

Greenwald, G., Mcghee, D. E., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition : The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain : neural models of stimulus-specific effects. Trends in Cognitive Science, 10(1), 14–23. http://doi.org/10.1016/j.tics.2005.11.006

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. Acta Psychologica, 107(1–3), 293–321. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11388140

Hamano, S. (1994). Palatalization of Japanese Sound Symbolism. In Sound Symbolism (pp. 148–157).

Hanna-Pladdy, B. (2012). Therapies for Ideomotor Apraxia.

Hampshire, A., Chamberlain, S. R., Monti, M. M., Duncan, J., & Owen, A. M. (2010). The role of the right inferior frontal gyrus: inhibition and attentional control. *NeuroImage*, *50*(3), 1313–1319. http://doi.org/10.1016/j.neuroimage.2009.12.109

Hartwigsen, G., Baumgaertner, A., Price, C. J., Koehnke, M., Ulmer, S., & Siebner, H. R. (2010). Phonological decisions require both the left and right supramarginal gyri. Proceedings of the National Academy of Sciences of the United States of America, 107(38), 16494–16499. http://doi.org/10.1073/pnas.1008121107

Hashimoto, T., Usui, N., Taira, M., Nose, I., Haji, T., & Kojima, S. (2006). The neural mechanism associated with the processing of onomatopoeic sounds. NeuroImage, 31(4), 1762–1770. http://doi.org/10.1016/j.neuroimage.2006.02.019

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(30), 7881–7. http://doi.org/10.1523/JNEUROSCI.1740-07.2007

Hikosaka, K., Iwai, E., Saito, H., & Tanaka, K. (1988). Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *Journal of Neurophysiology*, *60*(5), 1615–1637.

Hockett, C. (1960). The Origin of Speech. Scientific American, 203(3), 88–96.

Holloway, I. D., & Ansari, D. (2010). Developmental specialization in the right intraparietal sulcus for the abstract representation of numerical magnitude. *Journal of Cognitive Neuroscience*, *22*(11), 2627–37. http://doi.org/10.1162/jocn.2009.21399

Holloway, I. D., Price, G. R., & Ansari, D. (2010). Common and segregated neural pathways for the processing of symbolic and nonsymbolic numerical magnitude: an fMRI study. *NeuroImage*, *49*(1), 1006–17. http://doi.org/10.1016/j.neuroimage.2009.07.071

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews. Neuroscience*, *6*(6), 435–48. http://doi.org/10.1038/nrn1684

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369, 20130298-. http://doi.org/10.1098/rstb.2013.0298

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109(1), 54–65. http://doi.org/10.1016/j.cognition.2008.07.015

Indovina, I., & Macaluso, E. (2007). Dissociation of Stimulus Relevance and Saliency Factors during Shifts of Visuospatial Attention. *Cerebral Cortex*, *17*(7), 1701–1711. http://doi.org/10.1093/cercor/bhl081

Jamal, Y., Lacey, S., Nygaard, L. & Sathian, K. (2017). Interactions between auditory elevation, auditory pitch, and visual elevation during multisensory perception. Multisensory Research, 30:287-306

James, T. W., Stevenson, R. A., Kim, S., VanDerKlok, R. M., & James, K. H. (2011). Shape from sound: evidence for a shape operator in the lateral occipital cortex. *Neuropsychologia*, *49*(7), 1807-1815.

Johnson, M. (2013). The body in the mind: The bodily basis of meaning, imagination, and reason. University of Chicago Press.

Kadunce, D.C., Vaughan, W., Wallace, M.T., Benedek, G., Stein, B.E. Mechanisms of within- and crossmodality suppression in the superior colliculus. J Neurophysiol 1997, 78:2834-2847.

Kakehi, H., Tamori, I., & Schourup, L. (1996). Dictionary of Iconic Expressions in Japanese: Vol I: A-J. Vol II: K-Z. Walter de Gruyter.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–85. http://doi.org/10.1038/nn1444

Kamke, M. R., Vieth, H. E., Cottrell, D., & Mattingley, J. B. (2012). Parietal disruption alters audiovisual binding in the sound-induced flash illusion. *NeuroImage*, *62*(3), 1334–1341. http://doi.org/10.1016/j.neuroimage.2012.05.063

Kantartzis, K., Kita, S., & Imai, M. (2011). Japanese sound symbolism facilitates word learning in English speaking children. Cognitive Science, 35(3), 575–586.

Kassuba, T., Menz, M. M., Röder, B., & Siebner, H. R. (2012). Multisensory Interactions between Auditory and Haptic Object Recognition. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(5), 1–11. http://doi.org/10.1093/cercor/bhs076

Kaufmann, L., Vogel, S. E., Wood, G., Kremser, C., Schocke, M., Zimmerhackl, L.-B. B., & Koten, J. W. (2008). A developmental fMRI study of nonsymbolic numerical and spatial processing. *Cortex*, *44*(4), 376–385. http://doi.org/10.1016/j.cortex.2007.08.003

Kilian-Hatz, C. (2001). Universality and diversity: ideophones from Baka and Kxoe. In F. K. E. Voeltz & C. Kilian-Hatz (Eds.), Ideophones (pp. 155–163). Amsterdam: John Benjamins.

Kirby, S. (1996). Fitness and the selective adaptation of language. Processing, 1–19.

Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. Linguistics, 35(2), 379–415. http://doi.org/10.1515/ling.1997.35.2.379

Köhler; W. (1929). *Gestalt Psychology*. New York: Liveright Publishing Corporation.

Köhler, W. (1947). Gestalt Psychology: An Introduction to New Concepts in Modern

Psychology. Liveright: New York, NY.

Ković, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, *114*(1), 19–28. http://doi.org/10.1016/j.cognition.2009.08.016

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. http://doi.org/10.1016/j.neuron.2008.10.043

Krugliak, A., & Noppeney, U. (2016). Synaesthetic interactions across vision and audition. *Neuropsychologia*, *88*, 65-73.

Kunihira, S. (1971). Effects of the expressive voice on phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior*, *10*(4), 427-429.

Lacey, S., Flueckiger, P., Stilla, R., Lava, M., & Sathian, K. (2010). Object familiarity modulates the relationship between visual object imagery and haptic shape perception. *Neuroimage*, *49*(3), 1977-1990.

Lacey, S., Martinez, M., McCormick, K., & Sathian, K. (2016). Synesthesia strengthens sound-symbolic crossmodal correspondences, 1–6. http://doi.org/10.1111/ejn.13381

Lacey, S., Stilla, R., Sreenivasan, K., Deshpande, G., & Sathian, K. (2014). Spatial imagery in haptic shape perception. *Neuropsychologia*, *60*, 144-158.

Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bipolar concepts presented in vertical space. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 726–736. http://doi.org/10.1037/a0024955

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago press.

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*, 20120392. http://doi.org/10.1098/rstb.2012.0392

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(6), 2717–22. http://doi.org/10.1073/pnas.0907307107

Lecce, F., Walsh, V., Didino, D., & Cappelletti, M. (2015). “How many” and “how much” dissociate in the parietal lobe. *Cortex*, *73*(September), 73–79. http://doi.org/10.1016/j.cortex.2015.08.007

Lenartowicz, A., Verbruggen, F., Logan, G. D., & Poldrack, R. a. (2011). Inhibition-related activation in the right inferior frontal gyrus in the absence of inhibitory cues. *Journal of Cognitive Neuroscience*, *23*(11), 3388–99. http://doi.org/10.1162/jocn_a_00031

Leopold, D. a, Bondar, I. V, & Giese, M. a. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), 572–5. http://doi.org/10.1038/nature04951

Lewis, J. W. (2010). Audio-Visual Perception of Everyday Natural Objects – Hemodynamic Studies in Humans. In J. Kaiser & M. J. Naumer (Eds.), *Multisensory Object Perception in the Primate Brain* (pp. 155–190). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4419-5615-6

Lewkowicz, D. J., & Minar, N. J. (2014). Infants are not sensitive to synesthetic crossmodality correspondences: a comment on Walker et al. (2010). *Psychological Science*, *25*(3), 832–834. http://doi.org/10.1177/0956797613516011

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: a review of behavioral, developmental, and neuroimaging research into sound-symbolism. Frontiers in Psychology, 6(August), 1–14. http://doi.org/10.3389/fpsyg.2015.01246

Lockwood, G., Hagoort, P., & Dingemanse, M. (2016). How Iconicity Helps People Learn New Words: Neural Correlates and Individual Differences in Sound-Symbolic Bootstrapping. Collabra, 2(1), 7. http://doi.org/10.1525/collabra.42

Lourenco, S. F., & Longo, M. R. (2010). General magnitude representation in human infants. *Psychological Science : A Journal of the American Psychological Society / APS*, *21*(6), 873–81. http://doi.org/10.1177/0956797610370158

Lourenco, S. F., & Longo, M. R. (2011). *Origins and Development of Generalized Magnitude Representation. Space, Time and Number in the Brain* (First Edit, Vol. 1). Elsevier Inc. http://doi.org/10.1016/B978-0-12-385948-8.00015-3

Lourenco, S.F., Bonny, J.W., Fernandez, E. P. & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. Proceedings of the National Academy of Sciences USA, 109:18737-18742.

Lu, C. H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin and Review*, *2*, 174–207.

Luppino, G., & Rizzolatti, G. (2000). The organization of the frontal motor cortex. *Physiology*, *15*(5), 219-224.

Lupyan, G. (2012). What Do Words Do? Toward a Theory of Language-Augmented Thought. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 57, pp. 255–297). Elsevier. http://doi.org/10.1016/B978-0-12-394293-7.00007-8

Lupyan, G., & Bergen, B. K. (2015). How Language Programs the Mind. *Topics in Cognitive Science*, *8*(2), 408–424. http://doi.org/10.1111/tops.12155

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychology*, 1–10. http://doi.org/10.1177/0963721415570732

Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends in Neurosciences*, *28*(5), 264–271. http://doi.org/10.1016/j.tins.2005.03.008

Macaluso, E., Frith, C. D., & Driver, J. (2002). Supramodal effects of covert spatial orienting triggered by visual or tactile events. *Journal of Cognitive Neuroscience*, *14*(3), 389–401. http://doi.org/10.1162/089892902317361912

Maeda, F., Kanai, R., & Shimojo, S. (2004). Changing pitch induced visual motion illusion. *Current Biology*, *14*(23), 990–991. http://doi.org/10.1016/j.cub.2004.11.018

Maglio, S. J., Rabaglia, C. D., & Feder, M. A. (2014). Vowel Sounds in Words Affect Mental Construal and Shift Preferences for Targets, 143(3), 1082–1096. http://doi.org/10.1037/a0035543

Magnus, M. (2001). What's in a Word? Studies in Phonosemantics. PhD dissertation, NTNU.

Man, K., Damasio, A., Meyer, K., & Kaplan, J. T. (2015). Convergent and Invariant Object Representations for Sight , Sound , and Touch, *3640*, 3629–3640. http://doi.org/10.1002/hbm.22867

Marchant, J. L., Ruff, C. C., & Driver, J. (2012). Audiovisual synchrony enhances BOLD responses in a brain network including multisensory STS while also enhancing target-detection performance for both modalities. *Human Brain Mapping*, *33*(5), 1212–1224. http://doi.org/10.1002/hbm.21278

Marks, L. E. (1974). On associations of light and sound: the mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, *87*(1–2), 173–188. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4451203

Marks, L. E. (1987). On crossmodal similarity: auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology. Human Perception and Performance*, *13*(3), 384–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2958587

Marks, L. E., Ben-Artzi, E., & Lakatos, S. (2003). Crossmodal interactions in auditory and visual discrimination. *International Journal of Psychophysiology*, *50*(1–2), 125–145. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=15233672

Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving Similarity and Comprehending Metaphor. *Monographs of the Society for Research in Child Development*, *52*(1), i. http://doi.org/10.2307/1166084

Martin, A. (2007). The Representation of Object Concepts in the. *Brain*. http://doi.org/10.1146/annurev.psych.57.102904.190143

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*(7), 903–923. http://doi.org/10.1068/p2866

Martino, G., & Marks, L. E. (2000). Crossmodal interaction between vision and touch: the role of synesthetic correspondence. *Perception*, *29*(6), 745–754. http://doi.org/10.1068/p2984

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and Weak. *Current Directions in Psychological Science*, *10*(2), 61–65. http://doi.org/10.1111/1467-8721.00116

Maurer, D., & Mondloch, C. J. (2005). Neonatal Synesthesia: a Reevaluation. In L. C. Robertson & N. Sagiv (Eds.), *Synesthesia: Perspectives from Cognitive Neuroscience* (pp. 193–213). New York, NY US: Oxford University Press, Oxford.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006a). The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316–322. http://doi.org/10.1111/j.1467-7687.2006.00495.x

Maurer, D., Pathman, T., & Mondloch, C. J. (2006b). The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science*, *3*, 316–322.

McCormick, K., Kim, J. Y., List, S., & Nygaard, L. C. (2015). Sound to Meaning Mappings in the Bouba - Kiki Effect. In *CogSci 2015*.

McCormick, K., Lacey, S., Stilla, R., Nygaard, L. C., & Sathian, K. (2018). Neural basis of the crossmodal correspondence between auditory pitch and visuospatial elevation. *Neuropsychologia*, *112*, 19-30.

McCormick, K., Lacey, S., Stilla, R., Nygaard, L. C., & Sathian, K. (2018). Neural Basis Of The Sound-Symbolic Crossmodal Correspondence Between Auditory Pseudowords And Visual Shapes. *bioRxiv*, 478347.

Mcguire, P. K., Silbersweig, D. A., & Frith, C. D. (1996). Functional Neuroanatomy of Verbal Self-Monitoring. *Brain*, *119*, 907–917.

Melara, R. D., & Algom, D. (2003). Driven by information: a tectonic theory of Stroop effects. *Psychol Rev*, *110*(3), 422–471. http://doi.org/10.1037/0033-295X.110.3.422

Melara, R. D., & Marks, L. E. (1990a). Dimensional interactions in language processing: investigating directions and levels of crosstalk. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*(4), 539–54. http://doi.org/10.1037/0278-7393.16.4.539

Melara, R. D., & Marks, L. E. (1990b). Processes underlying dimensional interactions: Correspondences between linguistic and nonlinguistic dimensions. *Memory & Cognition*, *18*(5), 477–495. http://doi.org/10.3758/BF03198481

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, *116*(4), 323–336. http://doi.org/10.1037//0096-3445.116.4.323

Menon, V., Adleman, N. E., White, C. D., Glover, G. H., & Reiss, A. L. (2001). Error-related brain activation during a Go/NoGo response inhibition task. *Human Brain Mapping*, *12*(3), 131–143. http://doi.org/10.1002/1097-0193(200103)12:3<131::AID-HBM1010>3.0.CO;2-C

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. Brain Structure and Function, 214(5–6), 655–667. http://doi.org/10.1007/s00429-010-0262-0

Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F., & Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. Neuropsychologia, 76, 264–275. http://doi.org/10.1016/j.neuropsychologia.2015.01.042

Meyer, K., & Kaplan, J. T. (2011). Crossmodal Multivariate Pattern Analysis. *Journal of Visualized Experiments : JoVE*, (57), 1–6. http://doi.org/10.3791/3307

Meyer, K., Kaplan, J. T., Essex, R., Damasio, H., & Damasio, A. (2011). Seeing Touch Is Correlated with Content-Specific Activity in Primary Somatosensory Cortex. *Cerebral Cortex*, *21*(9), 2113–2121. http://doi.org/10.1093/cercor/bhq289

Meyer, K., Kaplan, J. T., Essex, R., Webber, C., Damasio, H., & Damasio, A. (2010). Predicting visual stimuli on the basis of activity in auditory cortices. *Nature Neuroscience*, *13*(6), 667–8. http://doi.org/10.1038/nn.2533

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, *11*, 735–59. http://doi.org/10.1006/nimg.2000.0568

Miller, J. (1991). Channel Interaction and the Redundant-Targets Effect in Bimodal Divided Attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 160–169. http://doi.org/10.1037/0096-1523.17.1.160\

Mock, J., Huber, S., Bloechle, J., Dietrich, J. F., Bahnmueller, J., Rennig, J., ... & Moeller, K. (2018). Magnitude processing of symbolic and non-symbolic proportions: an fMRI study. *Behavioral and Brain Functions*, *14*(1), 9.

Molholm, S., & Foxe, J. J. (2010). Editorial: Making sense of multisensory integration. European Journal of Neuroscience, 31(10), 1709–1712. http://doi.org/10.1111/j.1460-9568.2010.07238.x

Molholm, S., Martinez, A., Shpaner, M., & Foxe, J. J. (2007). Object-based attention is multisensory: Co-activation of an object's representations in ignored sensory modalities. European Journal of Neuroscience, 26(2), 499–509. http://doi.org/10.1111/j.1460-9568.2007.05668.x

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. http://doi.org/10.1093/cercor/bhh007

Monaghan, P., Mattock, K. & Walker, P. (2012). The role of sound symbolism in language learning. Journal of Experimental Psychology: Learning, Memory, & Cognition, 38:1152-1164.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369(1651). http://doi.org/10.1098/rstb.2013.0299

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369(1651). http://doi.org/10.1098/rstb.2013.0299

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective & Behavioral Neuroscience*, *4*(2), 133–6. http://doi.org/10.3758/cabn.4.2.133

Moore, B. C. J. (2012). *Psychology of Hearing* (6th ed.). Leiden, The Netherlands: Brill.

Morsella, E., Godwin, C. A., Jantz, T. K., Krieger, S. C., & Gazzaley, A. (2016). Homing in on consciousness in the nervous system: An action-based synthesis. *Behavioral and Brain Sciences*, *39*.

Mossbridge, J., Zweig, J., Grabowecky, M., & Suzuki, S. (2017). An Association between Auditory–Visual Synchrony Processing and Reading Comprehension: Behavioral and Electrophysiological Evidence. *Journal of Cognitive Neuroscience*, *29*(3), 435–447. http://doi.org/10.1162/jocn_a_01052

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology*, *66*, 347–352.

Muggleton, N., Tsakanikos, E., Walsh, V., & Ward, J. (2007). Disruption of synaesthesia following TMS of the right posterior parietal cortex. Neuropsychologia, 45(7), 1582–1585. http://doi.org/10.1016/j.neuropsychologia.2006.11.021

Mulvenna, C. M., & Walsh, V. (2006). Synaesthesia: supernormal integration? *Trends in Cognitive Sciences*, *10*(8), 350–352. http://doi.org/10.1016/j.tics.2006.06.004

Namy, L. L., & Nygaard, L. C. (2008). Perceptual-motor constraints on sound to meaning correspondence in language. *Behavioral and Brain Sciences*, *31*, 528–529.

Nee DE, Wager TD, Jonides J. 2007. Interference resolution: insights from a meta-analysis of neuroimaging tasks. Cogn Affect Behav Neurosci 7(1):1–17.

Newman, S. S. (1933). Further Experiments in Phonetic Symbolism. The American Journal of Psychology , 45(1), 53–75.

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual review of neuroscience*, *32*, 185-208.

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology*, *65*(2), 115–24. http://doi.org/10.1037/a0022268

Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. Language and Cognition, 4(2), 115–125. http://doi.org/10.1515/langcog-2012-0007

Noesselt, T., Bergmann, D., Heinze, H.-J. J., Münte, T. F., Spence, C., & Munte, T. (2012). Coding of multisensory temporal patterns in human superior temporal sulcus. *Frontiers in Integrative Neuroscience*, *6*(August), 64. http://doi.org/10.3389/fnint.2012.00064

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005

Noppeney, U. (2012). Characterization of Multisensory Integration with fMRI: Experimental Design, Statistical Analysis, and Interpretation. In M. Murray & M. Wallace (Eds.), The Neural Bases of Multisensory Processes. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22593861

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. Cerebral Cortex, 18(3), 598–609. http://doi.org/10.1093/cercor/bhm091

Nuckolls, J. B. (1996). Sounds like life: Sound-symbolic grammar, performance, and cognition in Pastaza Quechua (Vol. 2). Oxford University Press on Demand.

Nuckolls, J. B. (1999). The Case for Sound Symbolism. Annual Review of Anthropology, 28(1), 225–252. http://doi.org/10.1146/annurev.anthro.28.1.225

Nuckolls, J. B. (2003). To be or not to be ideophonically impoverished. In W. F. Chiang, E. Chun, L. Mahalingappa, & S. Mehus (Eds.), SALSA XI: Proceedings of the Eleventh Annual Symposium about

Language and Society (pp. 131–142). Austin. Retrieved from http://studentorgs.utexas.edu/salsa/proceedings/2003/nuckolls.pdf

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2008). Sound Symbolism in Word Learning. In V. M. Sloutsky, B. C. Love, & K. McRae (Eds.), Proceedings of the 30th Annual Meeting of the Cognitive Science Society (pp. 1912–1917). Washington D.C.

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. Cognition, 112(1), 181–186. http://doi.org/10.1016/j.cognition.2009.04.001

Oberhuber, M., Hope, T. M. H., Seghier, M. L., Parker Jones, O., Prejawa, S., Green, D. W., & Price, C. J. (2016). Four Functionally Distinct Regions in the Left Supramarginal Gyrus Support Word Processing. Cerebral Cortex, 26(11), 4212–4226. http://doi.org/10.1093/cercor/bhw251

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. Phonetica, 41(1), 1–16. http://doi.org/10.1159/000261706

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. *Sound Symbolism*.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (No. 47). University of Illinois press.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound-shape crossmodal correspondences in 4-month-olds. Journal of Experimental Child Psychology, 114(2), 173–86. http://doi.org/10.1016/j.jecp.2012.05.004

Parise, C. V, Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(16), 6104–8. http://doi.org/10.1073/pnas.1322705111

Parise, C. V, & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters*, *442*(3), 257–261. http://doi.org/10.1016/j.neulet.2008.07.010

Parise, C. V, & Spence, C. (2009). “When birds of a feather flock together”: Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PloS One*, *4*(5), e5664. http://doi.org/10.1371/journal.pone.0005664

Parise, C. V, & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, *220*(3–4), 319–333. http://doi.org/10.1007/s00221-012-3140-6

Parise, C. V, Spence, C., & Ernst, M. O. (2012). When Correlation Implies Causation in Multisensory Integration. *Current Biology*, *22*(1), 46–49. http://doi.org/http://dx.doi.org/10.1016/j.cub.2011.11.039

Parkinson, C., Kohler, P. J., Sievers, B., & Wheatley, T. (2012). Associations between auditory pitch and visual elevation do not depend on language: Evidence from a remote population. *Perception*, *41*(7), 854–861. http://doi.org/10.1068/p7225

Patching, G. R., & Quinlan, P. T. (2002). Garner and Congruence Effects in the Speeded Classification of Bimodal Signals. *Journal of Experimental Psychology. Human Perception and Performance*, *28*(4), 755–775. http://doi.org/10.1037//0096-1523.28.4.755

Peiffer-Smadja, N., & Cohen, P. L. (2010). *Exploring the bouba / kiki effect : a behavioral and fMRI study*. *Small*. Universite Paris V–Descartes, France.

Peiffer-Smadja, N., & Cohen, L. (2019). The cerebral bases of the bouba-kiki effect. *NeuroImage*, *186*, 679-689.

Peña, M., Mehler, J., & Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological science*, *22*(11), 1419-1421.

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages. Frontiers in Psychology, 1(December), 227. http://doi.org/10.3389/fpsyg.2010.00227

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in english and Spanish and its relation to lexical category and age of acquisition. PLoS ONE, 10(9). http://doi.org/10.1371/journal.pone.0137147

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2017). Iconicity in the speech of children and adults. Developmental Science, (September 2016), e12572. http://doi.org/10.1111/desc.12572

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–55. http://doi.org/10.1016/j.neuron.2004.10.014

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, *53*(2), 293–305. http://doi.org/10.1016/j.neuron.2006.11.022

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*(6), 983–93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15046729

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology*, *13*(3), 278–285. http://doi.org/10.1037/h0072651

Price, C. J., Moore, C. J., Humphreys, G. W., & Wise, R. J. S. (1997). Segregating Semantic from Phonological Processes during Reading. Journal of Cognitive Neuroscience, 9(6), 727–733. http://doi.org/10.1162/jocn.1997.9.6.727

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442. http://doi.org/10.1037/0033-2909.132.3.416

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and Language*, *127*(1), 86–103. http://doi.org/10.1016/j.bandl.2013.05.015

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. Retrieved from http://dx.doi.org/10.1038/nrn2811

Raczkowski, D., Kalat, J. W., & Nebes, R. (1974). Reliability and validity of some handedness questionnaire items. *Neuropsychologia*, *12*(1), 43–47. http://doi.org/10.1016/0028-3932(74)90025-6

Raij, T., Ahveninen, J., Lin, F.-H., Witzel, T., Jääskeläinen, I. P., Letham, B., … Belliveau, J. W. (2010). Onset timing of cross-sensory activations and multisensory interactions in auditory and visual sensory cortices. *European Journal of Neuroscience*, *31*(10), 1772–1782. http://doi.org/10.1111/j.1460-9568.2010.07213.x

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A Window Into Perception, Thought and Language. *Journal of Consciousness Studies*, *8*(12), 3–34.

Ramachandran, V. S., & Hubbard, E. M. (2003). Hearing Colors, Tasting Shapes. Scientific American, 288(5), 52–59. http://doi.org/10.1038/scientificamerican0503-52

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews. Neuroscience*, *5*(3), 184–194. http://doi.org/10.1038/nrn1343

Regenbogen, C., Seubert, J., Johansson, E., Finkelmeyer, A., Andersson, P., & Lundstr, J. N. (2018). The intraparietal sulcus governs multisensory integration of audiovisual information based on task difficulty, (December 2017), 1313–1326. http://doi.org/10.1002/hbm.23918

Reilly, J., Westbury, C., Kean, J., & Peelle, J. E. (2012). Arbitrary symbolism in natural language revisited: when word forms carry meaning. PloS One, 7(8), e42286. http://doi.org/10.1371/journal.pone.0042286

Revill, K. P., Namy, L. L., DeFife, L. C., & Nygaard, L. C. (2014). Cross-linguistic sound symbolism and crossmodal correspondence: Evidence from fMRI and DTI. *Brain and Language*, *128*(1), 18–24. http://doi.org/10.1016/j.bandl.2013.11.002

Revill, K. P., Namy, L. L., & Nygaard, L. C. (2018). Eye movements reveal persistent sensitivity to sound symbolism during word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(5), 680-698.

Riemer, M., Diersch, N., Bublatzky, F., & Wolbers, T. (2016). Space, time, and numbers in the right posterior parietal cortex: Differences between response code associations and congruency effects. *NeuroImage*, 1–8. http://doi.org/10.1016/j.neuroimage.2016.01.030

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, *27*, 169-192.

Rizzolatti, G., Fogassi, L., & Gallese, V. (1997). Parietal cortex: from sight to action. *Current opinion in neurobiology*, *7*(4), 562-567.

Rizzolatti, G., Luppino, G., & Matelli, M. (1998). The organization of the cortical motor system: new concepts. *Electroencephalography and clinical neurophysiology*, *106*(4), 283-296.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature reviews neuroscience*, *11*(4), 264.

Rogers, S. K., & Ross, A. S. (1975). A Cross-Cultural Test of the Maluma—Takete Phenomenon. Perception, 4(1), 105–106. http://doi.org/10.1068/p040105

Rouw, R., & Scholte, H. S. (2007). Increased structural connectivity in grapheme-color synesthesia. *Nature Neuroscience*, *10*(6), 792–7. http://doi.org/10.1038/nn1906

Rumelhart, D. E. (1979). Some problems with the notion of literal meanings. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 71–82). Cambridge University Press.

Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., & Butterworth, B. (2006). Spatial representation of pitch height : The SMARC effect Spatial representation of pitch height : the SMARC effect. *Cognition*, *99*(April), 113–129. http://doi.org/10.1016/j.cognition.2005.01.004

Sagiv, N., & Ward, J. (2006). Chapter 15 Crossmodal interactions: lessons from synesthesia. *Progress in Brain Research*, *155 B*(2001), 259–271. http://doi.org/10.1016/S0079-6123(06)55015-0

Santens, S., Roggeman, C., Fias, W., & Verguts, T. (2010). Number processing pathways in human parietal cortex. *Cerebral cortex*, *20*(1), 77-88.\

Santiago, J., & Lakens, D. (2015). Can conceptual congruency effects between number, time, and space be accounted for by polarity correspondence? *Acta Psychologica*, *156*, 179–191. http://doi.org/10.1016/j.actpsy.2014.09.016

Santiago, J., Román, A., & Ouellet, M. (2011). Flexible foundations of abstract thought: A review and a theory. *Spatial dimensions of social thought*, 41-110.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of experimental psychology*, *12*(3), 225.

Sathian, K., Simon, T. J., Peterson, S., Patel, G. A., Hoffman, J. M., & Grafton, S. T. (1999). Neural evidence linking visual object enumeration and attention. *Journal of Cognitive Neuroscience*, *11*(1), 36–51. http://doi.org/10.1162/089892999563238

Sathian, K., Zangaladze, A., Hoffman, J. M., & Grafton, S. T. (1997). Feeling with the mind's eye. *Neuroreport*, *8*(18), 3877-3881.

Sawamura, H., & Orban, G. A. (2006). Selectivity of Neuronal Adaptation Does Not Match Response Selectivity : A Single-Cell Study of the fMRI Adaptation Paradigm, 307–318. http://doi.org/10.1016/j.neuron.2005.11.028

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., … Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *J Neurosci*, *27*(9), 2349–2356. http://doi.org/10.1523/JNEUROSCI.5587-06.2007

Seghier, M. (2012). The Angular Gyrus: Multiple Functions and Multiple Subdivisions. *The Neuroscientist*, *19*(1), 43–61. http://doi.org/10.1177/1073858412440596

Seitz, A. R., Kim, R., van Wassenhove, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception*, *36*(10), 1445-1453.

Sestieri, C., Di Matteo, R., Ferretti, A., Del Gratta, C., Caulo, M., Tartaro, A., … Romani, G. L. (2006). "What" versus "Where" in the audiovisual domain: An fMRI study. *NeuroImage*, *33*(2), 672–680. http://doi.org/10.1016/j.neuroimage.2006.06.045

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–7. http://doi.org/10.1016/j.tics.2008.07.006

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, *55*(2), 167–177. http://doi.org/10.1016/j.jml.2006.03.002

Shor, R.E. (1975). An auditory analog of the Stroop test. Journal of General Psychology, 93:281-288.

Shrem, T., & Deouell, L. Y. (2017). Hierarchies of Attention and Experimental Designs: Effects of Spatial and Intermodal Attention Revisited. Journal of Cognitive Neuroscience, 29(1), 203–219. http://doi.org/10.1162/jocn_a_01030

Sidhu, D. M., & Pexman, P. M. (2017). Five mechanisms of sound-symbolic association. Psychonomic Bulletin & Review. http://doi.org/10.3758/s13423-017-1361-1

Simon, O., Mangin, J.-F., Cohen, L., Le Bihan, D., & Dehaene, S. (2002). Topographical Layout of Hand, Eye, Calculation, and Language-Related Areas in the Human Parietal Lobe. *Neuron*, *33*(3), 475–487. http://doi.org/10.1016/S0896-6273(02)00575-5

Smith, L. B., & Sera, M. D. (1992). A Developmental Analysis of the Polar Structure of Dimensions. *Cognitive Psychology*, *24*(1), 99–142. http://doi.org/10.1016/0010-0285(92)90004-l

Sokolowksi, H.M., Fias, W., Ononye, C.B. & Ansari, D. (2017). Are numbers grounded in a general magnitude processing system? A functional neuroimaging meta-analysis. *Neuropsychologia*, 105:50-69.

Spector, F., & Maurer, D. (2008). The colour of Os: Naturally biased associations between shape and colour. *Perception*, *37*(6), 841–847. http://doi.org/10.1068/p5830

Spector, F., & Maurer, D. (2009). Synesthesia: a new approach to understanding the development of perception. *Developmental Psychology*, *45*(1), 175–89. http://doi.org/10.1037/a0014171

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995. http://doi.org/10.3758/s13414-010-0073-7

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration ? Evaluating the spatial rule, 1–19. http://doi.org/10.1111/nyas.12121

Spence, C., & Parise, C. V. (2012). The cognitive neuroscience of crossmodal correspondences. *I-Perception*, *3*(7), 410–412. http://doi.org/10.1068/i0540ic

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P. J., Alex Meredith, M., Perrault, T. J., … Lewkowicz, D. J. (2010). Semantic confusion regarding the development of multisensory integration: A practical solution. European Journal of Neuroscience, 31(10), 1713–1720. http://doi.org/10.1111/j.1460-9568.2010.07206.x

Stevenson, R.A., Altieri, N.A., Kim, S., Pisoni, D.B. & James, T.W. (2010). Neural processing of asynchronous audiovisual speech perception. NeuroImage, 49:3308-3318.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. NeuroImage, 44(3), 1210–1223. http://doi.org/10.1016/j.neuroimage.2008.09.034

Stivers, T. (2008). Stance, Alignment, and Affiliation During Storytelling: When Nodding Is a Token of Affiliation. Research on Language & Social Interaction, 41(1), 31–57. http://doi.org/10.1080/08351810701691123

Styles, S. J., & Gawne, L. (2017). When Does Maluma/Takete Fail? Two Key Failures and a Meta-Analysis Suggest That Phonology and Phonotactics Matter. I-Perception, 8(4), 204166951772480. http://doi.org/10.1177/2041669517724807

Sučević, J., Savić, A. M., Popović, M. B., Styles, S. J., & Ković, V. (2015). Balloons and bavoons versus spikes and shikes: ERPs reveal shared neural processes for shape-sound-meaning congruence in words, and shape-sound congruence in pseudowords. Brain and Language, 145–146, 11–22. http://doi.org/10.1016/j.bandl.2015.03.011

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, *116*(2), 918–933.

Talairach, J. & Tournoux, P. (1988). Co-planar Stereotaxic Atlas of the Human Brain. Thieme Medical Publishers; New York.

Tarte, R. D. (1974). Phonetic symbolism in adult native speakers of Czech. Language and Speech, 17(1), 87–94.

Tarte, R. D., & Barritt, L. S. (1971). Phonetic Symbolism in Adult Native Speakers of English: Three Studies. Language and Speech, 14(2), 158–168.

Taylor, P. C. J., Muggleton, N. G., Kalla, R., Walsh, V., & Eimer, M. (2011). TMS of the right angular gyrus modulates priming of pop-out in visual search: combined TMS-ERP evidence. *Journal of Neurophysiology*, *106*(6), 3001–3009. http://doi.org/10.1152/jn.00121.2011

Thompson, P. D., & Estes, Z. (2011). Sound-symbolic naming of novel objects is a graded function. *Experimental Psychology*, *64*(October), 37–41. http://doi.org/10.1080/17470218.2011.605898

Trimble, O. C. (1934). Localization of sound in the anterior posterior and vertical dimensions of auditory space. *British Journal of Psychology*, *24*, 320–334.

Tufvesson, S. (2011). Analogy-making in the Semai Sensory World. The Senses and Society, 6(1), 86–95. http://doi.org/10.2752/174589311X12893982233876

Turkeltaub PE, Eden GF, Jones KM, Zeffiro TA. 2002. Metaanalysis of the functional neuroanatomy of single-word reading: method and validation. Neuroimage 16(3, Pt 1): 765–80.

Tzeng, C. Y., Nygaard, L. C., & Namy, L. L. (2017). The Specificity of Sound-symbolic Correspondences in Spoken Language. Cognitive Science, 41, 2191–2220. http://doi.org/10.1111/cogs.12474

Van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of Letters and Speech Sounds in the Human Brain. *Neuron*, *43*(2), 271–282.

Van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2006). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex*, *17*(4), 962-974.

Van Wanrooij, M. M., Bremen, P., & John Van Opstal, A. (2010). Acquired prior knowledge modulates audiovisual integration. *European Journal of Neuroscience*, *31*(10), 1763–1771. http://doi.org/10.1111/j.1460-9568.2010.07198.x

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics*, *69*(5), 744–756. http://doi.org/10.3758/BF03193776

Venezia, J. H., Fillmore, P., Matchin, W., Lisette Isenberg, A., Hickok, G., & Fridriksson, J. (2016). Perception drives production across sensory modalities: A network for sensorimotor integration of visual speech. *NeuroImage*, *126*, 196–207. http://doi.org/10.1016/j.neuroimage.2015.11.038

Vigliocco, G., & Kita, S. (2006). Language-specific properties of the lexicon: Implications for Learning and Processing. Language and Cognitive Processes, 21(7–8), 790–816. http://doi.org/10.1080/016909600824070

Visser, M., Jefferies, E., Embleton, K. V., & Lambon Ralph, M. a. (2012). Both the Middle Temporal Gyrus and the Ventral Anterior Temporal Area Are Crucial for Multimodal Semantic Processing: Distortion-corrected fMRI Evidence for a Double Gradient of Information Convergence in the Temporal Lobes. *Journal of Cognitive Neuroscience*, *24*(8), 1766–1778. http://doi.org/10.1162/jocn_a_00244

Visser, M., & Lambon Ralph, M. A. (2011). Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *Journal of Cognitive Neuroscience*, *23*(10), 3121–3131. http://doi.org/10.1162/jocn_a_00007

Von Kriegstein, K., Giraud, A.-L. AL, & Kriegstein, K. Von. (2006). Implicit Multisensory Associations Influence Voice Recognition. *PLoS Biology*, *4*(10), e326. http://doi.org/10.1371/journal.pbio.0040326

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony : a tutorial review. *Attention, Perception, & Psychophysics*, *72*(4), 871–884. http://doi.org/10.3758/APP

Wager TD, Sylvester CY, Lacey SC, Nee DE, Franklin M,

Jonides J. 2005. Common and unique components of response inhibition revealed by fMRI. Neuroimage 27(2):323–40.

Walker, L., Walker, P., & Francis, B. (2012). A common scheme for cross-sensory correspondences across stimulus domains. *Perception*, *41*(10), 1186–1192. http://doi.org/10.1068/p7149

Walker, P. (2012). Cross-sensory correspondences and cross talk between dimensions of connotative meaning : Visual angularity is hard , high-pitched , and bright, 1792–1809. http://doi.org/10.3758/s13414-012-0341-9

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic crossmodality correspondences. *Psychological Science*, *21*(1), 21–5. http://doi.org/10.1177/0956797609354734

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2014). Preverbal Infants Are Sensitive to Cross-Sensory Correspondences: Much Ado About the Null Results of Lewkowicz and Minar (2014). *Psychological Science*, *25*(3), 835–836. http://doi.org/10.1177/0956797613520170

Walker, P., & Smith, S. (1984). Stroop interference based on the synaesthetic qualities of auditory pitch. *Perception*, *13*(1), 75–81. http://doi.org/10.1068/p130075

Wallace, M. T., Meredith, M. A., & Stein, B. E. (1992). Integration of multiple sensory modalities in cat cortex. *Experimental Brain Research*, *91*(3), 484–488.

Wallace, M. T., Meredith, M. A., & Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *Journal of Neurophysiology*, *80*(2), 1006–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9705489

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*(2), 252–258. http://doi.org/10.1007/s00221-004-1899-9

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. http://doi.org/10.1016/j.tics.2003.09.002

Ward, J. (2013). Synesthesia. In B. E. Stein (Ed.), *The New Handbook of Multisensory Processing* (Vol. 64, pp. 319–343). Annual Reviews. http://doi.org/10.1146/annurev-psych-113011-143840

Ward, J., & Simner, J. (2003). Lexical-gustatory synaesthesia: Linguistic and conceptual factors. *Cognition*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027703001227

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. Brain and Language. http://doi.org/10.1016/j.bandl.2004.07.006

Westermann, D. H. (1927). Laut, Ton und Sinn in westafrikanischen Sudansprachen. In F. Boas (Ed.), Festschrift Meinhof (pp. 315–28). Hamburg.

Westermann, D. H. (1937). Laut und Sinn in einigen westafrikanischen Sudan-Sprachen. Archiv Für Vergleichende Phonetik, 1, 154–172.

Willems, R. M., Özyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage*, *47*(4), 1992–2004. http://doi.org/10.1016/j.neuroimage.2009.05.066

Wilson, L. B., Tregellas, J. R., Slason, E., Pasko, B. E., & Rojas, D. C. (2011). Implicit phonological priming during visual word recognition. NeuroImage, 55(2), 724–731. http://doi.org/10.1016/j.neuroimage.2010.12.019

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? Iconicity in English sensory words. Interaction Studies, 18(3), 433–454. Retrieved from http://sapir.psych.wisc.edu/papers/winter_perlman_perry_lupyan_interaction-studies.pdf

Wolff, P., & Gentner, D. (2011). Structure-mapping in metaphor comprehension. *Cognitive Science*, *35*(8), 1456–1488. http://doi.org/10.1111/j.1551-6709.2011.01194.x

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419. http://doi.org/10.1016/j.neuroimage.2013.12.058

Worthington, A. (2016). Treatments and technologies in the rehabilitation of apraxia and action disorganisation syndrome : A review. NeuroRehabilitation, 39, 163–174. http://doi.org/10.3233/NRE-161348

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio–visual simultaneity judgments. *Perception & Psychophysics*, *67*(3), 531–544.

Zangaladze, A., Epstein, C. M., Grafton, S. T., & Sathian, K. (1999). Involvement of visual cortex in tactile discrimination of orientation. *Nature*, *401*(6753), 587–590.

Zimmer, U., Roberts, K. C., Harshbarger, T. B., & Woldorff, M. G. (2010). Multisensory conflict modulates the spread of visual attention across a multisensory object. NeuroImage, 52(2), 606–616. http://doi.org/10.1016/j.neuroimage.2010.04.245

Zwaan, R. A. (2003). The immersed experiencer: Toward an Embodied theory of Language Comprehension. Psychology of Learning and Motivation, 44, 35–62. http://doi.org/0079-7421/04

Zwaan, R. A., & Kaschak, M. P. (2008). Language in the brain, body, and world. *The Cambridge handbook of situated cognition*, *368*.