Are audiovisual correspondences truly automatic? The influence of top-down effects

Laura Marie Getz
Mechanicsburg, PA

Master of Arts, University of Virginia, 2012
Bachelor of Arts, Elizabethtown College, 2009

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Psychology

University of Virginia
May 2016

Michael Kubovy, Chair
Dennis Proffitt
C. Daniel Meliza
Filip Loncke, Curry School of Education

Abstract

One of the main goals of our perceptual system is to decide when and how to bind information from our different senses to form a single muiltimodal percept. Recently it has been proposed that our knowledge of cross-modal correspondences may be a useful solution to this binding problem (Spence, 2011). If there is a consistent matching between sensory features across modalities, it can guide us to decide when these inputs should be combined. The first goal of my dissertation was to determine the replicability of correspondences between auditory pitch and visual dimensions of size (Size Studies 5-7), height (Height Studies 1a & 1b), spatial frequency (SF Studies 1a & 1b), brightness (Bright Studies 1a & 1b), and sharpness (Sharp Studies 1a & 1b). I failed to replicate seven out of ten correspondences. I conclude that audiovisual correspondences may *not* be a reliable solution to the binding problem.

Another current debate in visual perception is the extent to which perception is "cognitively impenetrable" to higher-order cognition (Firestone & Scholl, in press). The second goal of my dissertation was to determine whether audiovisual correspondences are subject to top-down cognitive influences. To do so, I asked participants to pair dimensions in a way incongruent with "natural" mappings based on environmental correlations or language knowledge (Size Studies 1-4, Height Studies 2-4, SF Studies 2-3, Bright Study 2, Sharp Study 2). I found that differing instructions changed response speed. I conclude that higher-order cognitive processes do influence basic cross-modal perception.

Together, these results point to the fact that audiovisual correspondences either *jointly* rely on bottom-up and top-down processing or are *solely* the result of top-down effects such as task instructions and lexical overlap. My dissertation results therefore strongly question the assumption of automaticity prevalent in the cross-modal correspondence literature. Audiovisual correspondences are a consequence of later decision-level influences rather than being truly automatic perceptual effects.

Keywords: cross-modal correspondence, audiovisual correspondence, top-down processing, automaticity, perception

Acknowledgements

Those who know me well, know that I am an extremely sentimental person, and thus it would be out of character for me not to deliver my heartfelt appreciation to a number of people who made the completion of my dissertation possible.

First and foremost, I would like to thank my advisor, Michael Kubovy. Most of the credit for the scientist I have become is thanks to Michael's mentorship and example. Michael has been a better advisor than I could have ever hoped for over the past six years. He has allowed me to explore a number of topics of interest, always incorporating careful methodological planning and using detailed and sophisticated analytic tools. He is one of the most intelligent people I have ever met, and if I am even 10% as successful and well-respected in the field when I reach Michael's stage, I will count my career as an outstanding success.

Next, thank you to the rest of my dissertation committee members (Denny Proffitt, Dan Meliza, and Filip Loncke), for the helpful advice and feedback. Denny has provided much sage advice over my career at UVa on research topics, how to give good presentations, and how to be an effective teacher. Denny's passion for teaching is infectious, and I will take many "tricks of the trade" with me to my future teaching career. Dan has been a great collaborator and I look forward to continuing our work teaching birds about rhythms. Filip has brought interesting linguistic insights to my work and has always been a cheerful supporter of my progress.

I would also like to thank Rachel Keen, for her research and career advice and emotional support during my time at UVa. She has always been happy to meet to discuss career goals, research ideas, and job market woes, and she is a welcome smiling face to see in the audience of my talks. And I would like to acknowledge Michael Roy, my undergraduate advisor at Elizabethtown College and a continued collaborator and supporter. Mike opened me up to all of the possibilities, excitement, and trials of doing research, is responsible for my most exciting international research experiences, and continues to provide advice and support even years after my days at Etown.

I would also like to acknowledge my labmates Steve Scheid and Minhong Yu for their assistance and feedback on my work, and the plethora of research assistants I have worked with over the years, especially my four DMP students (Priya, Ben, Sophie, and Anna) for their hard work and the excitement they brought to our research.

I am also grateful for my fellow grad students and friends at UVa, especially Elie Hessel, Kelly Hoffman, Diana Dinescu, and David Dobolyi, who shared in all of my struggles and successes over the past six years and helped keep me sane through all the twists and turns.

(I'm almost done, I promise!) I would also like to thank my family and friends for continuing to support me, encourage me, and celebrate with me throughout this long process. To my parents (Dave and Vonny Getz) for always telling me "I'm so proud of you", to my siblings (Sara, Matthew, Miranda, Lauren, Brian) for being my best friends, to my grandparents (Charlie and Geri Nease, Dick and Dottie Kepner) and my in-laws (John and Leslie Herneisey) for the endless support and love, and to my friends who have cheered me on from near and far.

And last but certainly not least, I would like to thank my amazing husband, Adam, for trekking through this journey with me. Adam has been my number one supporter for what feels like forever now. He has never once doubted me, even when I doubted myself. He knows when to push me and when I need to take a break. He has been through all of the highs and lows of this grad school process by my side, and I can almost guarantee that I would not have made it through in one piece without him. I consider myself incredibly lucky to have such a wonderful partner and friend, and I am excited for the next step of our adventure together in Philadelphia!

Finally, I would like to dedicate this dissertation to my grandfather, Dr. Russell Getz, who I never had the chance to meet, but whose dissertation "The influence of familiarity through repetition in determining optimism response of seventh grade children to certain types of serious music" has been on my bookshelf inspiring me to become the next Dr. Getz from the start of my grad school career.

Laura Getz
Charlottesville, VA, USA
April 8th, 2016

Table of Contents

Chapter 1: Introduction

One of the main problems of the human perceptual system is deciding when and how multimodal inputs should be combined into a single percept; this problem is commonly referred to as multisensory integration or the binding problem. Traditionally, spatial and temporal coincidence effects have been offered as useful constraints in solving this binding problem (Spence, 2007). For example, the ventriloquism effect shows the importance of temporal coincidence in binding sounds to lip movement: although a sound comes from one location and lip movements from another location, the sound seems to emanate from the location of the mouth when the two are presently simultaneously.

More recently, cross-modal correspondences, i.e., the consistent matching between sensory features across two modalities, have been proposed as a reliable solution to the binding problem (Spence, 2011). For example, auditory pitch and visual size tend to be matched in a way that reflects our experience with the environment; large objects make low-pitched sounds and small objects make high-pitched sounds. Matched endpoints of large/low and small/high tend to result in faster and more accurate processing than when the endpoints are reversed (i.e., large/high and small/low).

This dissertation seeks to address three concerns with the cross-modal correspondence literature. *First*, although it has been proposed that cross-modal correspondences may help solve the binding problem, little work has been done to assess the **replicability** of such audiovisual correspondences. If the matching of features across modalities is not stable across changes in methodology, the correspondence may not be a relevant solution to the problem. This is particularly relevant given recent discussions in the field of psychology on the need for replication (Cumming, 2014; Lakens & Evers, 2014; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011).

Second, a current debate in the field of visual perception is the extent to which perception is modular or "cognitively impenetrable" to the effects of higher-order influence (Firestone & Scholl, in press). There are many proponents for a blurring of the lines between perception and cognition (Clark, 2013; Collins & Olson, 2014; Goldstone, de Leeuw, & Landy, 2015; Vetter & Newen, 2014). Cross-modal correspondences have been left out of this debate because the effects are all assumed to be automatic (i.e., "contained within perception itself, rather than being an effect of more central cognitive processes *on* perception" Firestone & Scholl, in press, p. 11). However, a logical question to ask is whether all cross-modal correspondences are encapsulated from cognitive influence, or alternatively, whether some correspondences are subject to **top-down influence** from higher-order cognitive processes such as motivation, emotion, attention, or memory.

Third, based on differences in neural substrates and different effects on human information processing, it has been proposed that there are qualitatively different **correspondences types**; namely, structural, statistical, and semantic (see Table 1 for the table taken from Spence, 2011). *Structural* correspondences are likely innate: there is evidence that they appear in infants and universally across cultures (e.g., loudness and brightness are redundant coding mechanisms for intensity). *Statistical* correspondences are learned through regularities in the environment (e.g., pitch and size, see above). *Semantic* correspondences are dependent on language knowledge (e.g., pitch and visual elevation each use the words 'low' and 'high' to describe dimension endpoints). However, little has been done to challenge this categorization scheme despite a number of problems (described in detail below).

Replicability

Speeded Classification Paradigm

In speeded classification experiments, participants are asked to rapidly classify a multimodal stimulus according to its value on a single relevant dimensions while ignoring the value on an irrelevant dimension. Figure 1 shows two illustrations of the typical speeded classification task paradigm. For example, Gallace and Spence (2006) completed a speeded *size* classification task. They collected reaction times as participants identified the *size* of the object while ignoring irrelevant *pitch* information present on each trial (Figure 1a).

In a majority of audiovisual speeded classification tasks, it is the *visual* dimension that is relevant while the *auditory* information is irrelevant. However, Evans and Treisman (2010) compared a visual-relevant task to an auditory-relevant task. They collected reaction times on a *visual task* where participants identified the *elevation* of the object while ignoring irrelevant *pitch* information and on an *auditory task* where participants identified the *height* of the pitch while ignoring irrelevant *visual elevation* information (Figure 1b).

Speeded Classification Typical Results

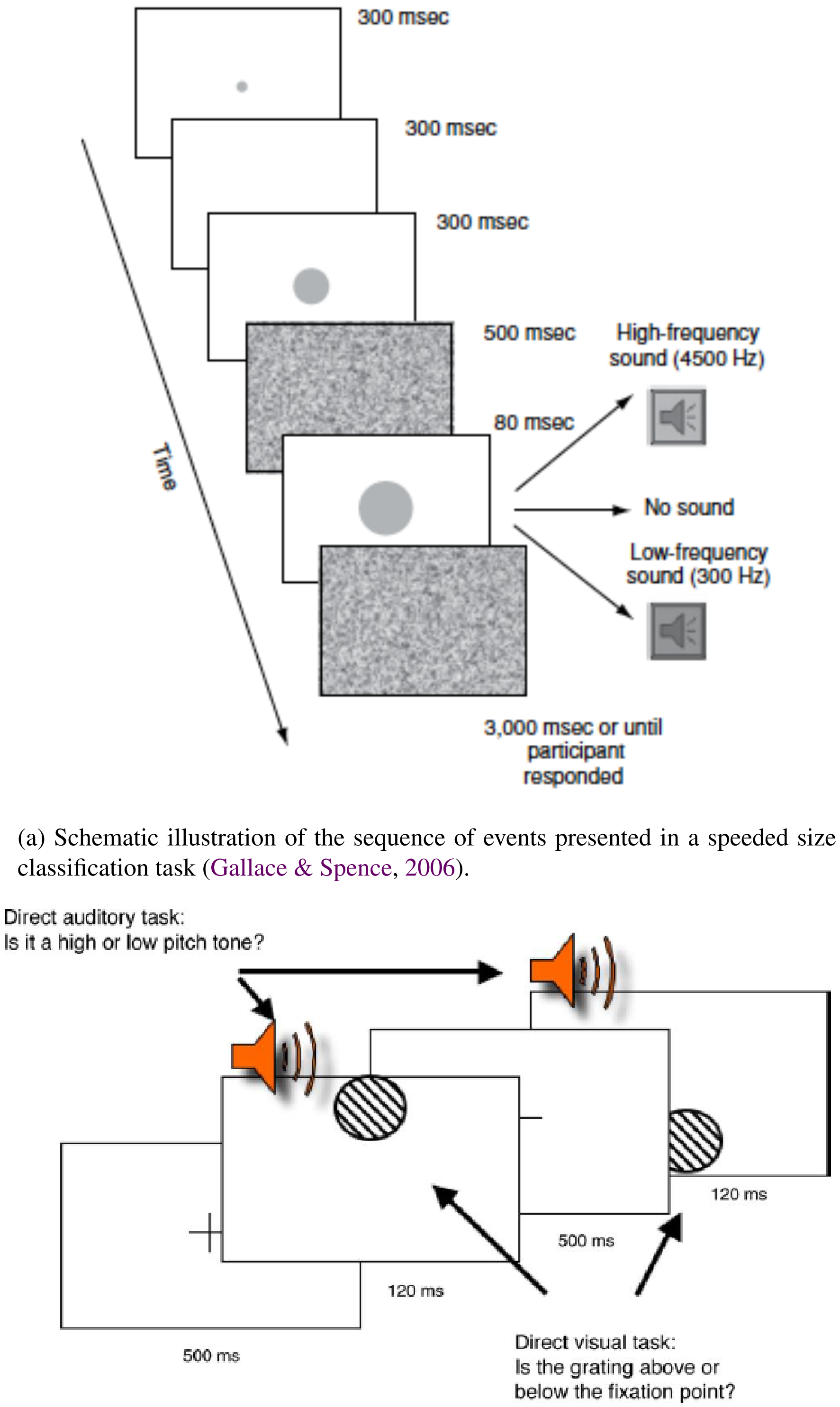
If two modalities (e.g., pitch and size) are automatically associated, then it may be difficult to selectively attend to one dimension while ignoring variation on the irrelevant dimension. Thus participants typically complete three types of trials: (a) *congruent* trials are those in which the value on the irrelevant dimension matches the classification of the relevant dimension (e.g., a *large* object is accompanied by a *low* pitch); (b) *incongruent* trials are those in which the value on the irrelevant dimension is opposite of the classification of the relevant dimension (e.g., a *small* object is accompanied by a *low* pitch); and (c) *unimodal*[1] trials are those where the relevant dimension is presented without a value in the irrelevant dimension (e.g., a visual object is presented *without* any pitch).

The typical finding is that matching congruent endpoints leads to faster and/or more accurate processing. From this researchers have concluded that a variety of audiovisual correspondences rely on bottom-up processing (i.e., they are not under voluntary control), because participants are asked to ignore the pitch but it still affects their ability and speed to complete the visual classification.

Dissertation Replication

Despite finding a congruency effect using the speeded classification paradigm, two things remain unclear. *First*, the relationship between *unimodal* trials and bimodal *congruent* and *incongruent* trials is ambiguous. Some studies have found reaction time facilitation for congruent pairings and reaction time interference for incongruent pairings compared to the unimodal condition (Evans & Treisman, 2010; Patching & Quinlan, 2002), others find facilitation for both bimodal conditions compared to the unimodal condition (Gallace & Spence, 2006), and others leave out the unimodal condition entirely (Ben-Artzi & Marks, 1995; Marks, 1987a). Thus one objective of my replication experiments was to more systematically address the distinction between unimodal and bimodal conditions.

[1] Though previous literature uses ‘neutral’ to describe the unimodal condition, I will use unimodal here to avoid confusion when switching between visual-relevant and auditory-relevant tasks.


300 msec
300 msec
300 msec
500 msec
80 msec
High-frequency sound (4500 Hz)
No sound
Low-frequency sound (300 Hz)
Time
3,000 msec or until participant responded
Direct auditory task:
Is it a high or low pitch tone?
120 ms
500 ms
120 ms
500 ms
Direct visual task:
Is the grating above or below the fixation point?

(a) Schematic illustration of the sequence of events presented in a speeded size classification task (Gallace & Spence, 2006).

(b) Schematic illustration of the sequence of events presented in a speeded pitch/elevation classification task (Evans & Treisman, 2010).

Figure 1. Two sample illustrations of a typical speeded classification task paradigm.

Second, little work has been done to assess the replicability of audiovisual correspondences. In my dissertation, I attempted direct and conceptual replications of a variety of audiovisual correspondences in order to test their stability. If the matching of features across modalities is not reliable across changes in methodology, then cross-modal correspondences may not be a reliable solution to the binding problem. Alternatively, if some correspondences hold up to methodological changes, they are likely to be more relevant solutions. Thus a second objective of my replication experiments was to discover potential consistent properties of cross-modal correspondences that influence their stability.

Top-Down Influence

The correspondence between pitch and size (as well as other audiovisual correspondences) has been assumed to stem from bottom-up processing. However, before my dissertation, no studies have specifically tested for top-down influences on correspondence processing. A plethora of recent research suggests that visual perception may change the way it operates in response to higher-order cognitive functions such as motivation, action, emotion, categorization, and language (Clark, 2013; Collins & Olson, 2014; Goldstone et al., 2015; Vetter & Newen, 2014). My work addresses whether cross-modal perception is "cognitively impenetrable" or whether it is subject to similar top-down influences. My objective is to determine the nature and scope of potential top-down effects.

Improvements to Speeded Classification Paradigm

To investigate the cognitive penetrability of cross-modal correspondences, I created a modified version of the typical classification task (see Figure 2). Specifically, I felt two changes were necessary. *First,* previous studies have assumed that the congruency effect is the result of bottom-up processing without specifically testing for top-down influences. I investigated the top-down influence of task instructions by reversing the mapping to which participants responded; participants completed blocks with with compatible instructions and blocks with incompatible instructions.

- **Compatible instructions:** blocks where "perceptually congruent" endpoints are paired (e.g., participants would select either a large shape/low pitch or small shape/high pitch).
- **Incompatible instructions:** blocks where "perceptually incongruent" endpoints are paired (e.g., participants would select either large shape/high pitch or small shape/lower pitch).

Second, as described earlier, participants in a typical speeded classification task are only required to pay attention to a single modality while the other modality remains irrelevant throughout. I required participants to pay attention to both modalities on each trial as it was not revealed until the end whether they should respond to the pitch or shape dimension. On each trial, the first shape appeared left of center; it was accompanied by a tone of varying pitch. Following, the second shape appeared right of center; it was also accompanied by a tone. After both pitch/shape pairings, either an 's' for shape or 'p' for pitch appeared on the screen to cue the participant to which modality to respond. Once the cue appeared, participants pressed a key to indicate whether the first or second stimulus met the instruction criteria (e.g., for larger/lower instructions, participants would respond to the *lower* pitch if a 'p' appeared or the *larger* shape if an 's' appeared).

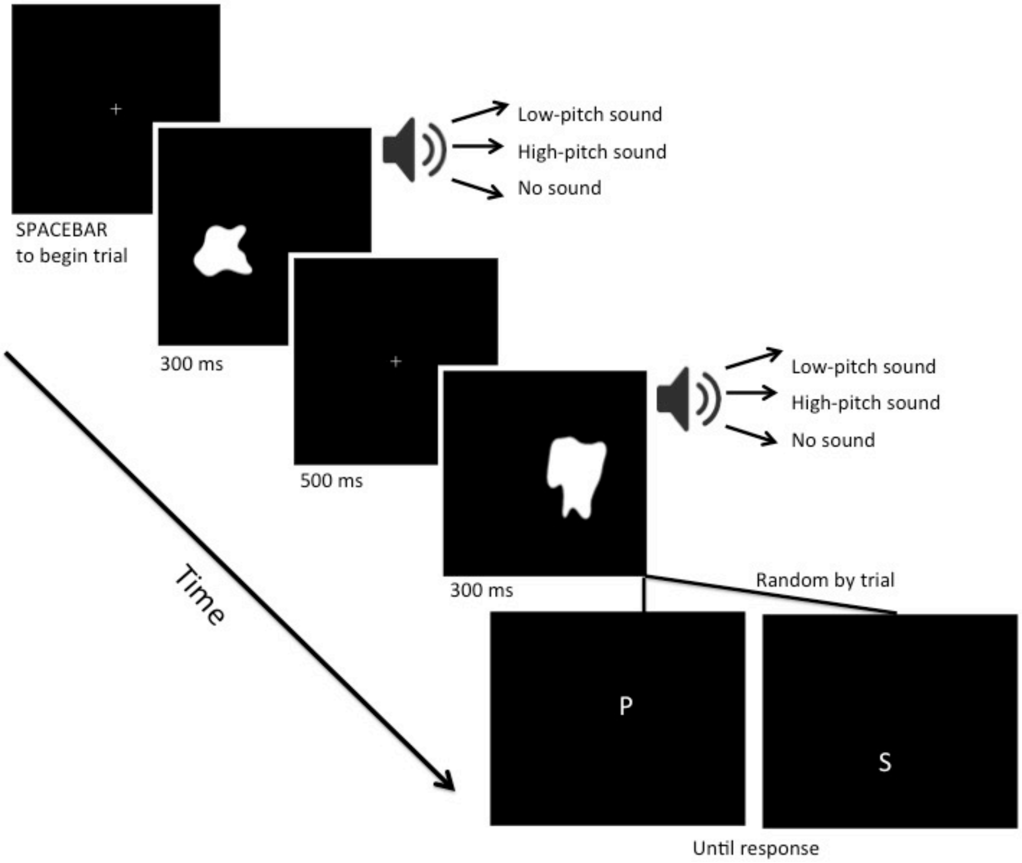


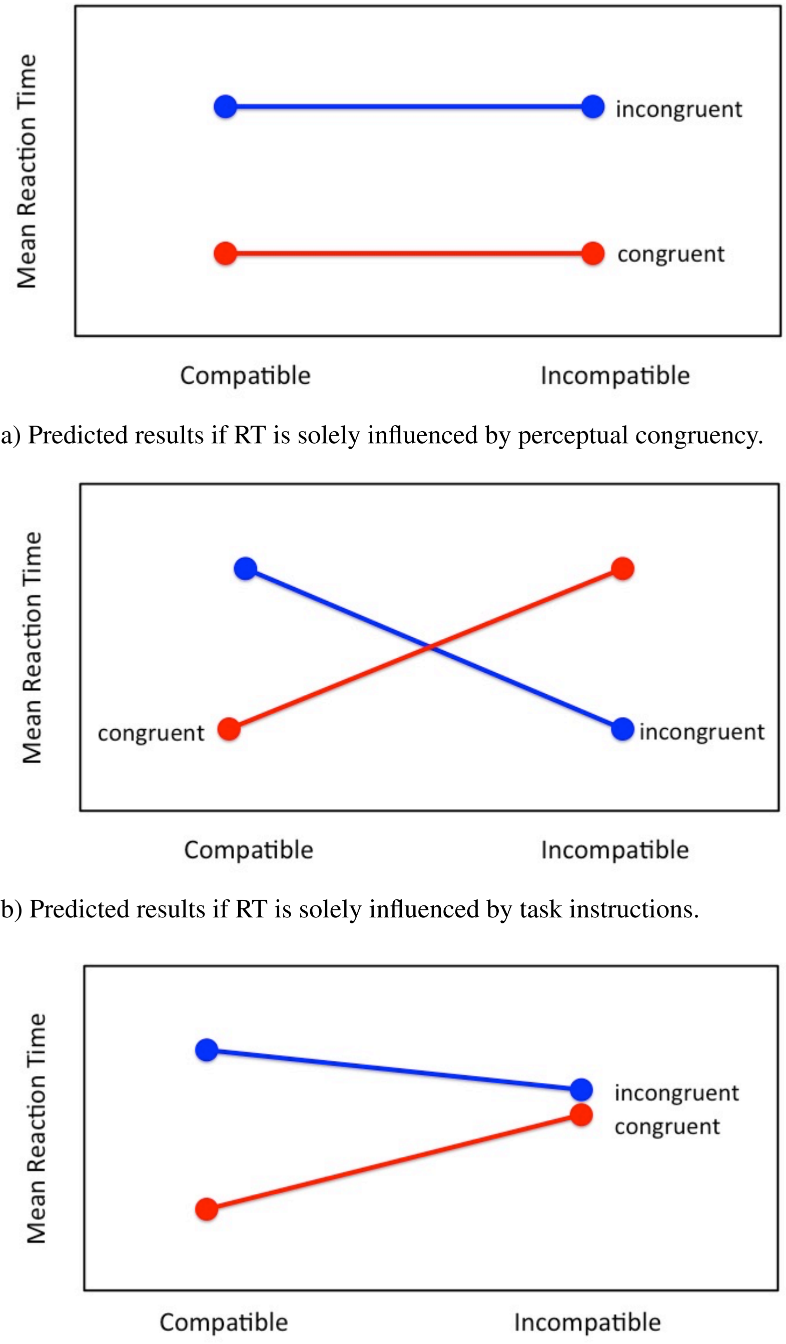
Figure 2. Schematic illustration of the sequence of events presented in each trial of my modified speeded classification task. Specifics of each experiment are included in the respective method section.

Alternative Outcomes

Figure 3 shows three possible outcomes for the top-down influence studies. In Figure 3a, participants respond solely on the basis of perceptual congruency; responding is always faster to the perceptually congruent pairing regardless of task instructions. This alternative would show strong support for a natural correspondence that is impenetrable to top-down influence.

In Figure 3b, participants respond solely to the task instructions; whatever they are asked to pair together in the instructions, they are faster when those dimensions appear together. In this alternative, perceptual congruency does not influence reaction time. This would imply no natural bottom-up correspondence between the dimensions and only evidence for top-down processing.

In Figure 3c, participants are influenced by perceptual congruency and task instructions. They are always faster to respond to the perceptually congruent pairing, showing evidence of some natural correspondence between the dimensions. However, when the instructions ask them to pair incongruent endpoints, those trials gain speed and congruent trials lose some speed, with the effects neutralizing one another in the incompatible condition. This alternative would show evidence both for bottom-up and top-down processing.



(a) Predicted results if RT is solely influenced by perceptual congruency.

(b) Predicted results if RT is solely influenced by task instructions.

(c) Predicted results if RT is affected by congruency and task instructions.

Figure 3. Alternative outcomes for the top-down influence experiments.

Audiovisual Correspondence Types

Current Categorization

Not all cross-modal correspondences arise from the same neural substrates or have the same effects on human information processing. Because of this, Spence (2011) argues that cross-modal correspondences should be divided into qualitatively different types: structural, statistical and semantic (see Table 1).

Table 1
Three principle types of audiovisual correspondences according to Spence (2011).

Type	Examples	Time Course and Explanation	Consequences
Structural	loudness–brightness	Possibly innate, but may also depend on maturation of neural structures for stimulus coding	Perceptual & decisional
Statistical	pitch–elevation pitch–size loudness–size	Learned: Coupling priors established on the basis of experience with regularities of the environment	Perceptual & decisional
Semantic	pitch–elevation pitch–spatial frequency	Learned: Emerge following language development as certain terms come to be associated with more than one perceptual continuum	Primarily decisional

Structural. Structural correspondences result from intrinsic properties of the human neural system organization; in other words, they result from neural connections likely present at birth (Marks, 1978). Although postulated to be innate, there are still several possible mechanisms that could account for such correspondences (Ramachandran & Hubbard, 2001). They could simply be a chance byproduct of the neural architecture required to run the cognitive system more generally. It is also possible that dimensions come to be associated due to an overlap in firing space of adjacent brain regions. Finally, the brain could use redundant mechanisms to process features in different sensory modalities, thus resulting in structures to become associated cross-modally.

Examples of structural correspondences include intensity matching between auditory loudness and visual brightness (Bond & Stevens, 1969; Lewkowicz & Turkewitz, 1980; Marks, Ben-Artzi, & Lakatos, 2003), structural magnitude matching (see Walsh, 2003, A Theory of Magnitude: ATOM) between auditory loudness and visual size (R. Walker, 1985)[2], and magnitude sound symbolism (Ohala, 1997; Sapir, 1929).

Statistical. Statistical correspondences are learned from pairs of stimulus dimensions that are correlated in nature and reflect our brain's ability to respond to regularities in the environment (e.g., R. Walker, 1987). These correspondences are likely to be universal, as they are determined

[2] Note, however, that Spence (2011) lists loudness–size as a statistical correspondence.

by physical properties of objects. For example, associations between the size of an object and its resonant frequency (e.g., larger objects have lower frequency when struck, dropped, sounded, etc.; see Grassi, 2005), between the size of an object and the pitch of its voice (e.g., mice squeak and lions roar; children have higher pitched voices than adults; see Spector & Maurer, 2009), and between auditory frequency and vertical elevation (e.g., human larynx descends when producing lower-pitched sounds; see Atkinson, 1978; Parkinson, Kohler, Sievers, & Wheatley, 2012) appear to be internalized regardless of culture.

Semantic. Semantic correspondences arise following language development as a result of overlap in the terms used to describe stimuli in two dimensions.[3] These correspondences are likely to be contextually determined, not universally present across all cultures, and not present in infancy (Martino & Marks, 1999). Unlike structural and semantic correspondences which may rely on early perceptual stages of processing, semantic correspondences likely rely only on later decisional stages of information processing. The most cited example of a semantic correspondence is between auditory pitch and visual elevation/height (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010; Melara & O'Brien, 1987; Patching & Quinlan, 2002; Roffler & Butler, 1968).

Problems

Innate vs. Learned Distinction. In Spence's (2011) categorization, structural correspondences are innate, whereas statistical or semantic correspondences must be learned. However, there is not a clear distinction between innate vs. learned correspondences. Rather, behavioral research reveals that innate biases facilitate learning of the statistics present in the environment, which then reinforce the strength of the innate associations (Maurer, Gibson, & Spector, 2012; Spector & Maurer, 2009). Similarly, the neural basis of cross-connections between modalities only permits the opportunity for associations to become systematic, and thus learning is necessary to strengthen the correspondences (Esterman, Verstynen, Ivry, & Robertson, 2006; Muggleton, Tsakanikos, Walsh, & Ward, 2007; Ramachandran & Hubbard, 2001).

Perceptual vs. Decisional Distinction. In Spence's (2011) categorization, only semantic correspondences imply the need for language knowledge; structural and statistical correspondences may be perceived at a more basic perceptual level. However, there is not a clear distinction between perceptual vs. decisional correspondences. Rather, semantic mediation seems to be one type of reinforcement used to strengthen cross-modal correspondences which are either natural biases or learned associations based on body experiences or environmental interactions (Dolscheid, Shayan, Majid, & Casasanto, 2013; Eitan & Timmers, 2010; Parkinson et al., 2012).

To take an example of the pitch–height correspondence, the fact that participants can easily be trained to use an unfamiliar pitch mapping (used in a different culture), but *not* a mapping unused by any culture (Dolscheid et al., 2013) suggests that statistical learning or natural biases may play more of a role than language knowledge in forging the pitch–height mapping. There is also evidence that Westerners can understand pitch metaphors not used in their culture (Eitan & Timmers, 2010) and evidence for pitch–height congruency among individuals who do not use such a metaphor in their native language (Parkinson et al., 2012).

[3] P. Walker (2012) argues that *Lexical* may be a more appropriate term because of the emphasis on "features of a word that are used to find the word's entry in a dictionary, *without* reference to any of its meanings" (p. 1794). *Semantic* or *Linguistic* are too broad if the intent is just to classify overlap in the words used to classify a dimension across modalities. I will use *Lexical* to describe this overlap throughout my dissertation.

Incomplete Categorization Scheme. The lack of distinction between correspondences means that there is often evidence to place audiovisual correspondences in more than one category, i.e., the categories are not mutually exclusive. Further, there is often little evidence to place correspondences into any of the given categories, i.e., the categories are not collectively exhaustive. For example, pitch–height could be placed into all three categories, whereas pitch–sharpness and pitch–brightness do not fall neatly into any of the three given categories.

It is therefore possible that there is a missing category that could help explain some of the hard-to-place correspondences. One such example I will put forth is a *metaphorical* correspondence (see Lakoff & Johnson, 1980a, 1980b). Similar to semantic correspondences, metaphorical correspondences would rely on language knowledge. However, instead of using the same word to describe categories from both modalities (e.g., 'high' and 'low' pitch and visual elevation), these would rely on our understanding of how dimensions are described symbolically (e.g., high pitches are described as 'sharp' or 'piercing'; low pitches are described as 'dark' or 'rich').[4]

Dissertation approach

Clearly an updated classification scheme is needed. There is a lot of overlap between Spence's (2011) structural, statistical, and semantic categories; the distinction between the three categories is not clear regarding innate vs. learned time course or regarding perceptual vs. decisional consequences. Further, the classification types are not mutually exclusive or collectively exhaustive, meaning there may be a need for additional categories.

By investigating the replicability of a number of audiovisual correspondences, I hope to better understand the likelihood of a given correspondence being innate. Replicability across changes in methodology may provide better evidence for a structural component to the correspondence mapping. By investigating top-down influences, I hope to better understand whether the correspondence is purely perceptual or whether it is subject to semantic-level consequences. Together, these methods will help offer a principled way to categorize cross-modal correspondences either with new categories or with an update to Spence's (2011) categorization. This will be beneficial to better group known correspondences and to make predictions about potential new correspondences.

Dissertation Experiments

The subsequent chapters will address the concerns about audiovisual correspondence replicability, top-down influences, and categorization by looking at correspondences between auditory pitch and visual dimensions of size, height, spatial frequency, brightness, and sharpness. Figure 4 provides an example of the stimuli used to investigate each of these correspondences.

Chapter 2: Pitch–Size Correspondence

Previous research has shown that auditory pitch and visual object size are perceptually matched in a way that reflects our experience with the environment (i.e., a *statistical* correspondence; Evans & Treisman, 2010; Gallace & Spence, 2006; Mondloch & Maurer, 2004; Spence, 2011).

[4] A similar category was suggested by P. Walker (2012). In four experiments, he found that cross-sensory correspondences can reflect activation among dimensions of connotative meaning (i.e., "what it suggests, implies, or invokes, rather than what it explicitly or directly denotes"; p. 1792).

Chapter 3: Pitch–Height Correspondence

Auditory pitch and visual height/elevation is the most cited audiovisual correspondence (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010; Melara & O'Brien, 1987; Patching & Quinlan, 2002; Spence, 2011). In addition to being a reflection of the *statistics* in the environment, this correspondence is also considered *semantic/lexical* because many languages use the same words—'low' and 'high'—to describe auditory stimuli that vary in pitch and visual stimuli that vary in height.

Chapter 4: Pitch–Spatial Frequency Correspondence

Similar to pitch and height, auditory pitch and visual spatial frequency is a *semantic/lexical* correspondence (Evans & Treisman, 2010; Spence, 2011) because the words 'low' and 'high' are used to describe stimuli that vary in pitch and spatial frequency (i.e., the repetition rate of an object's sinusoidal components per unit of distance, or the wideness/narrowness of the object's striping pattern).

Chapter 5: Pitch–Brightness Correspondence

Auditory pitch and visual brightness may be a *structural* correspondence because the correspondence is unlikely to arise from environmental associations (Maurer et al., 2012), yet endpoints seems to be consistently matched across individuals (Marks, 1974, 1987a; Martino & Marks, 1999; Mondloch & Maurer, 2004). Pitch–brightness is also an example of a *metaphorical* association: we often describe high pitches as 'bright' and low pitches as 'dark'.

Chapter 6: Pitch–Sharpness Correspondence

Auditory pitch and visual sharpness may be another example of a *structural* correspondence. There is no obvious connection between the roundness/sharpness of an object and the frequency of the sound it makes, yet infants and adults alike consistently match endpoints, suggesting an intrinsic mapping between brain areas responsible for auditory and visual perception (Marks, 1987a; Maurer et al., 2012; O'Boyle & Tarte, 1980; Parise & Spence, 2009). Pitch–sharpness is another example of a *metaphorical* association: we describe high pitches as 'sharp', 'piercing' or 'shrill'.

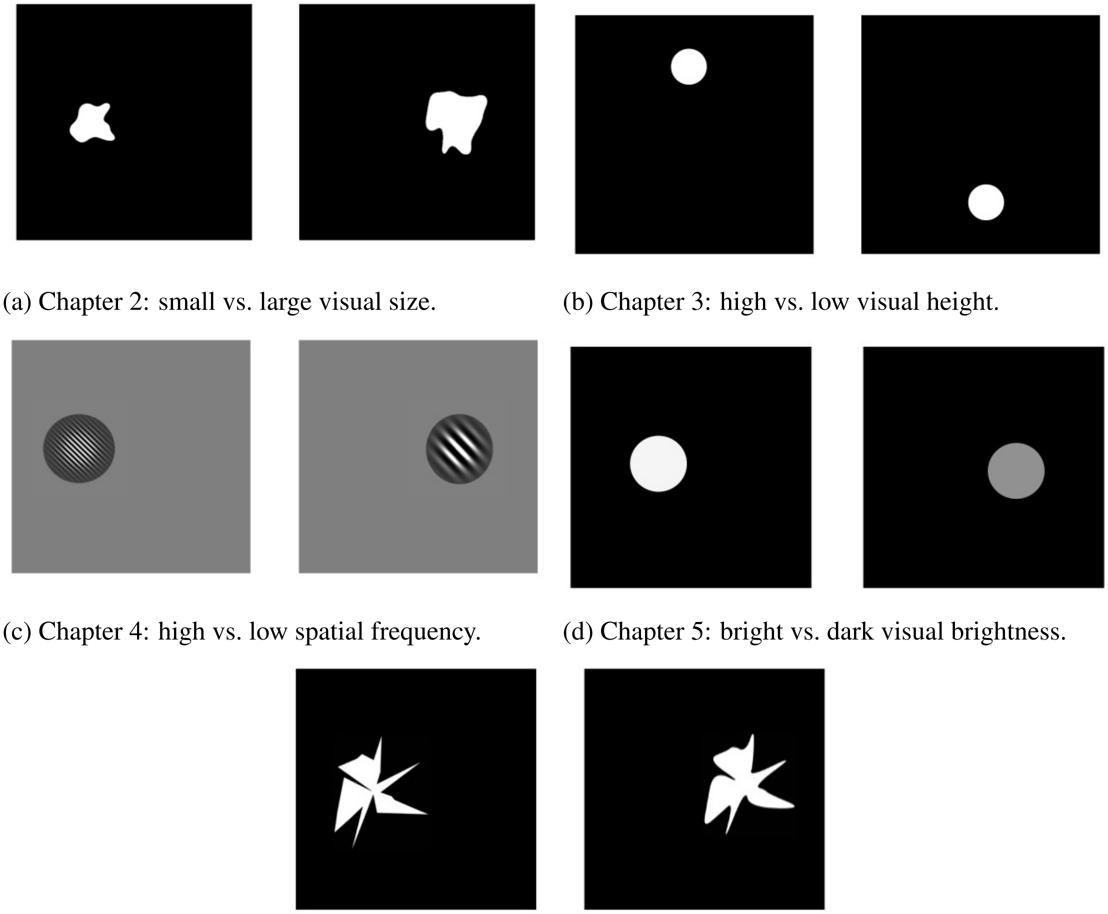
Analysis Plan

I take a more robust analytical approach to the field of cross-modal correspondences than used in previous studies. All of the analyses were performed using `R` (R Development Core Team, 2016).

Data management. In all of the experiments, we excluded participants who failed to meet one or both of the following criteria:

1. At least 60% accuracy on pitch and visual trials;
2. At least 60% completion rate of the total trials.

Our reaction time analysis only included correct responses (error data were analyzed separately). For each experiment, we went through several steps to transform the reaction time data. First, we eliminated responses that were faster than 50ms because they were assumed to be errors of program timing. Next, the reaction time data were subjected to a Box-Cox analysis using the `R` package `car` (Fox & Weisberg, 2011) to determine the appropriate transformation to assure normality (Box & Cox, 1964). Because the transformation suggested by the majority of the experiments was a



(a) Chapter 2: small vs. large visual size.

(b) Chapter 3: high vs. low visual height.

(c) Chapter 4: high vs. low spatial frequency.

(d) Chapter 5: bright vs. dark visual brightness.

(e) Chapter 6: sharp vs. rounded visual shape.

Figure 4. Audiovisual correspondences used in Chapters 2–6.

logarithmic transformation, logRT will be reported throughout.[5] Finally, with the transformed data we removed outliers that were more than three median absolute deviations (MADs) from the median reaction time; MAD is a more robust measure of dispersion than standard deviation (Leys, Ley, Klein, Bernard, & Licata, 2013).

Model comparison. We modeled our reaction time data with linear mixed-effects models (LMMS), computed using the `R` package `nlme` (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2016). We modeled our error data using a binomial response variable (correct vs. incorrect), which allowed us to use logistic mixed-effects regression models[6] (GLMMS), computed using the `R` package `lme4` (Bates, Maechler, & Bolker, 2014).

[5] For studies that suggested a different transformation, I ran the analyses both with the logRT and the alternate transformation. In no case did I find differences in significant results across transformations, which justifies my consistent use of logRT throughout all experiments

[6] For experiments with < 2% errors, we used a poisson rather than binomial distribution.

LMMS and GLMMS, which use maximum-likelihood estimation, have many advantages over traditional repeated-measures analysis of variance (ANOVA), which use ordinary least-squares. In addition to providing estimates of fixed effects, they allow us to predict subject-by-subject variations in model parameters (called random effects). Because we are interested not in effects present only at an individual level but rather in generalizable effects, LMMS/GLMMS allow us to partition out differences between individuals and model them jointly as random effects, thus leaving the variance we care about to be explained by the fixed effects. This provides a clear advantage over traditional ANOVA approaches that require prior averaging across subjects and/or items (Baayen, Davidson, & Bates, 2008).

To determine which fixed and random effects were important in modeling reaction time and errors, we used a method of maximum-likelihood model comparison based on the Akaike information criterion (AIC), which offers a principled balance between goodness-of-fit and parsimony (see Bozdogan, 1987; Burnham, Anderson, & Huyvaert, 2011, for introductory presentations). Because the probability of overfitting can be substantial when using AIC (Claeskens & Hjort, 2008), we used AICc— which penalizes extra parameters more heavily than does AIC— as recommended by Anderson and Burnham (2002).

Model fit. Whereas AICc is an appropriate method for model comparison and selection, it tells us nothing about the absolute model fit of a model. To give us an idea of this fit, we computed two types of R^2 for LMMS using the `MuMIn` package (Bartoń, 2014). The first, called the marginal R^2 ($R^2_{\text{marg.}}$), estimates the proportion of variance accounted for by the fixed effects, whereas the second, called the conditional R^2 ($R^2_{\text{cond.}}$), estimates the proportion of variance accounted for by the fixed and random effects taken together (Johnson, 2014; Nakagawa & Schielzeth, 2013).

Significant predictors. Instead of reporting null-hypothesis tests for our LMMS, we report 95% confidence intervals for the fixed-effect parameters displayed on a coefficient plot. These may be interpreted as tests of significance: if the confidence interval for an estimated parameter does not straddle zero, this estimate may be considered significant at $\alpha < 0.05$.

Model visualization. The traditional method for plotting main effects and interactions for LMMS uses the `effects` package (Fox, 2003). However, the error bars on such graphs take into account the random effects and thus with large individual differences (e.g., when individual participants are faster or slower to respond overall), the effect plot gives us very large error bars that overlap even when the effect is significant (see Size Study 1 for an example).

In order to avoid these large error bars that make inferences regarding significance difficult, we use an alternative method to visualize the model predictions: least significant difference (`LSD`) bars using Tukey's correction for pairwise comparisons (Tukey, 1949). `LSD` analysis is used to determine the minimum difference between means of any two groups before they can be considered significantly different.

This plotting method uses the `R` packages `lmerTest` (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2016) and `predictmeans` (Luo, Ganesh, & Koolaard, 2014). Individual `LSD` values are calculated for each participant; I used the average `LSD` value in the analyses. On each graph, I have included a lower and upper `LSD` bar (situated at the minimum and maximum point estimates) for ease of group comparison. Any difference between groups larger than the height of the bar is statistically significant.

Chapter 2: Pitch–Size Correspondence

Previous research has shown that auditory pitch and visual object size are perceptually matched in a way that reflects our experience with the environment (i.e., a *statistical* correspondence; Evans & Treisman, 2010; Gallace & Spence, 2006). Evidence of this association comes both from the relationship between the size of an object and its resonant frequency (e.g., larger objects have lower frequency when struck, dropped, sounded, etc.; see Grassi, 2005) and from the relationship between the size of an object and the pitch of its voice (e.g., mice squeak and lions roar; children have higher pitched voices than adults; see Spector & Maurer, 2009). Thus we will call large/low and small/high "perceptually congruent" endpoints.

In this series of studies, I investigated the stability of the pitch-size correspondence and the possibility of top-down influence of different task instructions. The procedure for all experiments is based on Figure 2; details/changes to the procedure will be explained in the methods of each study.

In Size Studies 1–4, I used an improved speeded classification paradigm to investigate how easily the endpoint mapping can be reversed under top-down control of attention (see Figure 3 for alternative outcomes). In each experiment, participants completed two blocks with with compatible instructions and two blocks with incompatible instructions.

- **Compatible instructions:** blocks where "perceptually congruent" endpoints are paired. In *Block 1* listeners selected either large shape or low pitch; in *Block 2* listeners selected either small shape or high pitch.
- **Incompatible instructions:** blocks where "perceptually incongruent" endpoints are paired. In *Block 3* listeners selected either large shape or high pitch; in *Block 4* listeners selected either small shape or low pitch.

Secondarily, I sought to address whether the pitch–size correspondence would be present with smaller sine-tone pitch differences than the extreme pitch difference used in many previous studies (e.g., 300 Hz vs. 4500 Hz as the low and high pitches, respectively; Gallace & Spence, 2006; Parise & Spence, 2009). In addition to the pitch difference from previous studies, Size Studies 1–2 used two smaller pitch ranges: 440 Hz vs. 880 Hz and 500 Hz vs. 630 Hz.[7] Additionally, given that the resonant properties of percussion instruments give more indication of their relative size than pure tones (Carello, Anderson, & Kunkler-Peck, 1998; Coward & Stevens, 2004; Grassi, 2005), Size Studies 3–4 used pitched percussion tones of varying pitch differences: 2468 Hz vs. 142 Hz; 752 Hz vs. 183 Hz; and 285 Hz vs. 143 Hz.

In Size Studies 5–7, I conducted several replications of Gallace and Spence (2006) to establish the stability (i.e., by direct replication) and the strength (i.e., by conceptual replication) of the correspondence. If cross-modal correspondences are to be used as a reliable source of information in solving the binding problem, then the benefits of congruent mapping should be easily replicable under a variety of conditions.

[7] Smaller pitch differences (600-680 Hz, 460-820 Hz, 320-960 Hz, and 180-1100 Hz) have been used to investigate the pitch–height correspondence (Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002).

Size Study 1: Top-down Influence [Pure tones, within-participants]

For Size Study 1, auditory stimuli were sine tones of varying frequencies. We used a within-participant design in which each participant completed a block of each of the four types of instructions (described above).

Method

Participants

Thirty-one U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

Visual shapes. Visual stimuli were generated randomly using the imagemagick command-line tools for Unix with Fred Weinhaus's extensions (`www.fmwconcepts.com/imagemagick/randomblob/index.php`). All shapes were set to be a white stimulus on a black background. Large shapes were 325 pixels in area and small shapes were 200 pixels in area (see Figure 4a). The outer image was set to be 600×600 pixels. Each image included 16 randomly generated points drawn from a uniform distribution. The points were successively connected with a spline curve and a guassian blur was added to the lines of the image.

Auditory pitches. We used three different pitch pairings: 'large' (300 Hz vs. 4500 Hz), 'octave' (440 Hz vs. 880 Hz), and major third ['M3'] (500 Hz vs. 630 Hz). All tones were pure sine tones and were 300 ms in duration. LG determined equal loudness for the tones beforehand (rather than having them set individually by participant); all tones were between 95-99 dB.

Design

We used a within-participant design in which each participant completed four blocks of 96 trials in a random order. The experiment instructions changed by block so that two blocks had 'compatible' instructions and two blocks had 'incompatible' instructions.

Procedure

The experiment was programmed using Matlab (MATLAB, 2013). Figure 2 shows an illustration of the sequence of events presented in each trial. At the start of each trial, the instructions were presented on the screen. In two of the four blocks, the pairings were 'compatible' (i.e., larger/lower and smaller/higher), and in the other two blocks, the pairings were incompatible (i.e., larger/higher and smaller/lower). The instructions were presented on every trial to remind participants of the pairing; participants pressed the `SPACEBAR` to proceed with the trial. The first shape appeared for 300ms left of center; it was accompanied by a 300ms high or low tone. This was followed by a 500ms blank screen. Then the second shape appeared for 300ms right of center; it was also accompanied by a 300ms tone. After both pitch/size pairings, either an 's' for shape or 'p' for pitch appeared on the screen to cue the participant to which modality to respond. Participants indicated whether the first or second stimulus presented met the instruction criteria (e.g., for larger/lower instructions, participants would respond to the *lower* pitch if a 'p' appeared or the *larger* shape if an 's' appeared). The cue remained on the screen until the participant responded, at which point the instruction screen reappeared to begin the next trial.

Data management

Participants. In Size Study 1, we removed three participants with less than 60% accuracy on one or both stimulus dimensions. The average overall accuracy of the remaining 28 participants was 92.9%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (1.2% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the four fixed effect predictors; the result suggested a logarithmic transformation (logRT; see Figure 5). Finally, using the logRT data we removed outliers that were more than three median absolute deviations (MADs) from the median logRT, which removed an additional 6% of responses.

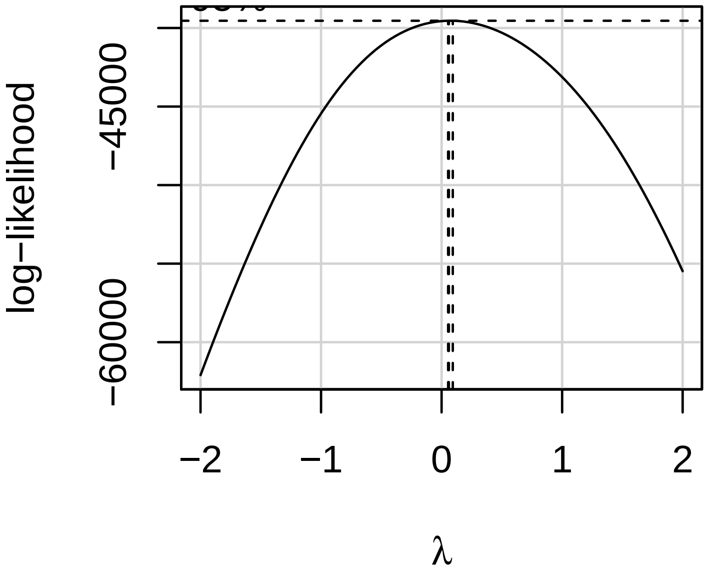


Figure 5. Size Study 1: Box-Cox analysis using an additive model of the four predictors (described in text). The resulting analysis suggests that a logarithmic transformation would correct the skew of the data.

Results and Discussion

Reaction Time

Model comparison. Our LMMS (here and in Size Studies 2, 3, and 4) included four categorical fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch vs. responding to shape), and `pitch difference` (M3 vs. octave vs. large). We also included the subject-by-subject variation in the intercept as a random effect[8].

[8] Because of the large number of fixed effects and their interactions, we did not have enough power to include differences in predictor slopes as additional random effects.

Table 2

Size Study 1*: Comparison of two models predicting reaction time, ordered by ΔAICc relative to the model with the lowest (best) AICc. K is the number of parameters in the model.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	18	0.0	1.00	0.036	0.288
`Additive`	`Intercept`	7	59.4	0.00	0.024	0.279

We compared an additive model including just the four fixed effects to a non-additive model including the four fixed effects and all of their two, three, and four-way interactions. Table 2 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 59.4, which (using the terminology introduced by Jeffreys, 1961) is "very strong evidence" in favor of the better model.

The marginal R^2 for the best model, which reflects the proportion of the variance accounted for by the fixed effects only, is rather small ($R^2_{\text{marg.}} \approx 0.04$). The conditional R^2, which reflects the proportion of the variance accounted for by the model as a whole (fixed and random effects together), is not insubstantial: $R^2_{\text{cond.}} \approx 0.29$.

Significant findings. Figure 6 shows the predictors of logRT and their 95% confidence intervals. From the figure, we can see that there are only two significant interactions (whose 95% CI does not straddle zero): `congruency x compatibility` and `modality x pitch difference`.

The `congruency` × `compatibility` interaction is the most important. By analyzing the compatible and incompatible trials separately, we see that the congruency advantage completely reverses with the change in instructions. Although there is a marginal *congruency* advantage on compatible trials (0.04, [0, 0.07]), there is an even greater *incongruency* advantage on incompatible trials (−0.13, [−0.16, −0.09]).

Despite these significant effects, plotting the results of the LMM in Figure 7 with ±1 standard error bars showcases large individual differences. The effect is hard to visualize because the standard error bars take into account the random effect of Intercept (see Figure 8). Figure 9 shows `LSD` bars instead of standard error bars. This figure clearly shows the top-down influence of instructions on performance: whatever dimensions are paired in the instructions for a given block, participants respond faster when those attributes are paired together (i.e., on the larger/lower block, participants are faster to respond to a large object paired with a low pitch; on the larger/higher block, participants are just as fast to respond to a large object paired with a high pitch).

The other significant two-way interaction is `modality` × `pitch difference` (Figure 10). The interaction arises from the fact that pitch difference matters to reaction time when participants respond to the pitch, but not when they respond to shape. In other words, participants are faster to identify the lower/higher pitch when there is a large pitch difference compared to an octave or M3 pitch difference, but they are equally fast at identifying the larger/smaller shape with a large, octave, and M3 pitch pairing. This replicates previous reaction time results using position judgments and pitch judgments to investigate the pitch–height correspondence (Ben-Artzi & Marks, 1995).

Because the three-way interactions of `congruency` × `compatibility` with `modality` and `pitch difference` are not significant, we can conclude that the top-down influence of instructions generalizes when participants respond to shape and pitch and across all three pitch differences.

Another noteworthy finding comes from a three-way interaction model with congruency, compatibility, and trial number as fixed effects (Figure 11). The three way-interaction is not significant (-1.004, $[-4.61, 2.61]$), which means the interaction strength does not change across trials. In other words, participants respond faster on later trials within a block, but the increased speed is not different between the congruent and incongruent trials or the compatible and incompatible blocks. Therefore, we can rule out a learning explanation for our top-down influence.

Errors

Model comparison. Our binomial GLMMS (here and in Size Studies 2, 3, and 4) included four categorical fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch vs. responding to shape), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

Because of the low error probability overall, a non-additive model including the four fixed effects and all of their two-, three-, and four-way interactions failed to converge. Instead, we compared a non-additive model including the four fixed effects and the `congruency × compatibility` two-way interaction (as it was the most important in predicting reaction times) to an additive model including just the four fixed effects. The AICc difference (ΔAICc) between the two models is 25.3, which is "strong evidence" in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability (p-value) that there is no difference between the log-odds of correct responses for compatible and incompatible trials is $p = 0.24$ ($z = 1.19$), meaning listeners make approximately the same number of errors on congruent trials regardless of compatibility condition. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p = 0.03$ ($z = 2.11$), meaning listeners make a few more errors on incongruent trials. These effects are converted to proportions and shown in Table 3. The more accurate conditions were also faster (see Figure 9), meaning there is no evidence of a speed-accuracy tradeoff.

Table 3
Size Study 1*: Proportion of errors across compatibility and congruency conditions.*

	`Compatible blocks`	`Incompatible blocks`
`Congruent trials`	0.06	0.07
`Incongruent trials`	0.10	0.04

Conclusions

We found strong evidence for top-down influence of instructions on performance across modalities and pitch differences: whatever dimensions are paired in the instructions for a given block, participants respond faster when those attributes are paired together. Further, this interaction was not a function of simply learning the new pairing throughout the block of trials, as the interaction strength did not change from the first to last trial in any block.

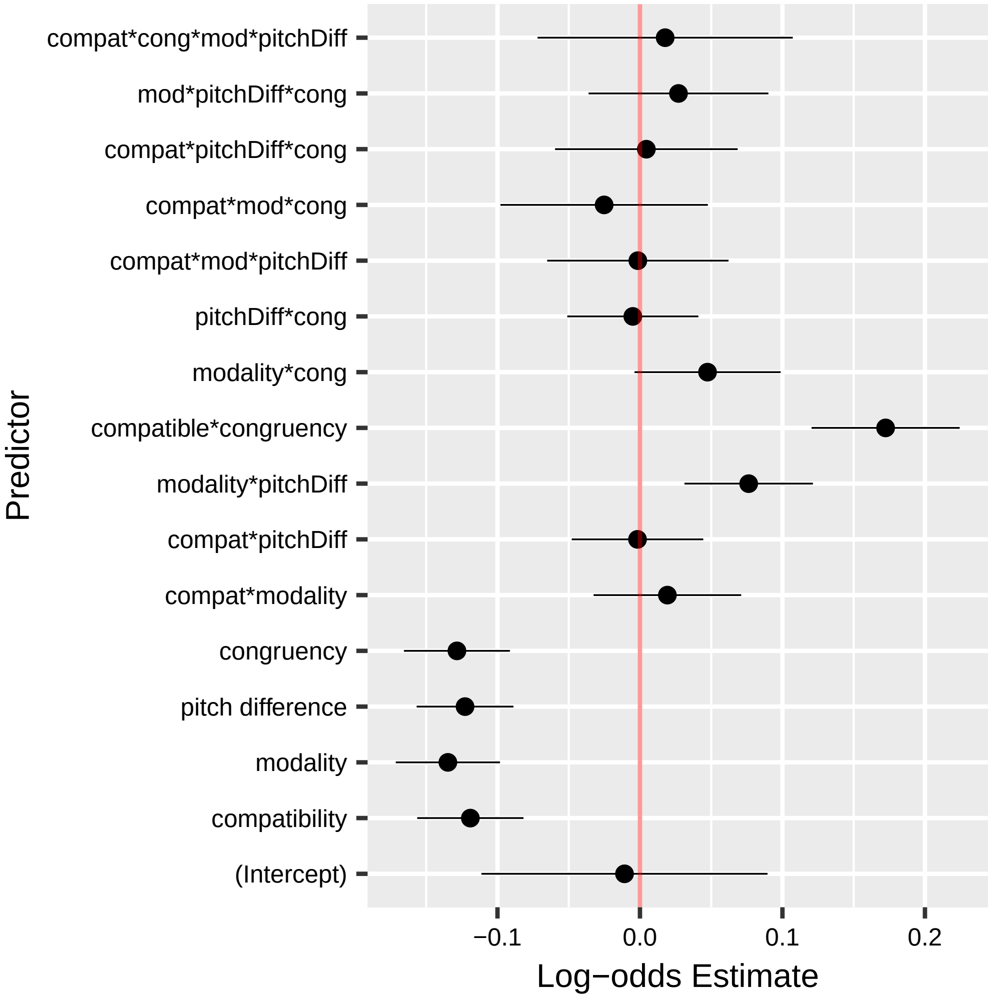


Figure 6. *Size Study 1*: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals) Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference.

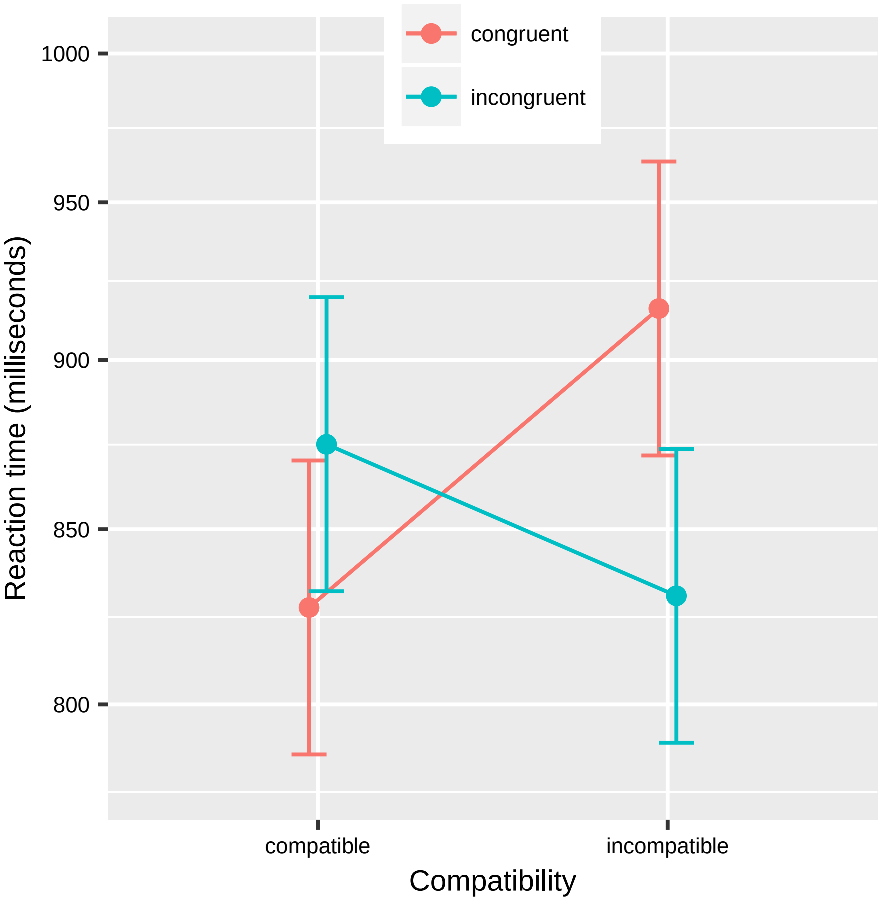


Figure 7. *Size Study 1*: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with standard error bars). There is a significant congruency effect for compatible trials and significant incongruency effect for incompatible trials. The large error bars are a function of large individual differences (Figure 8).

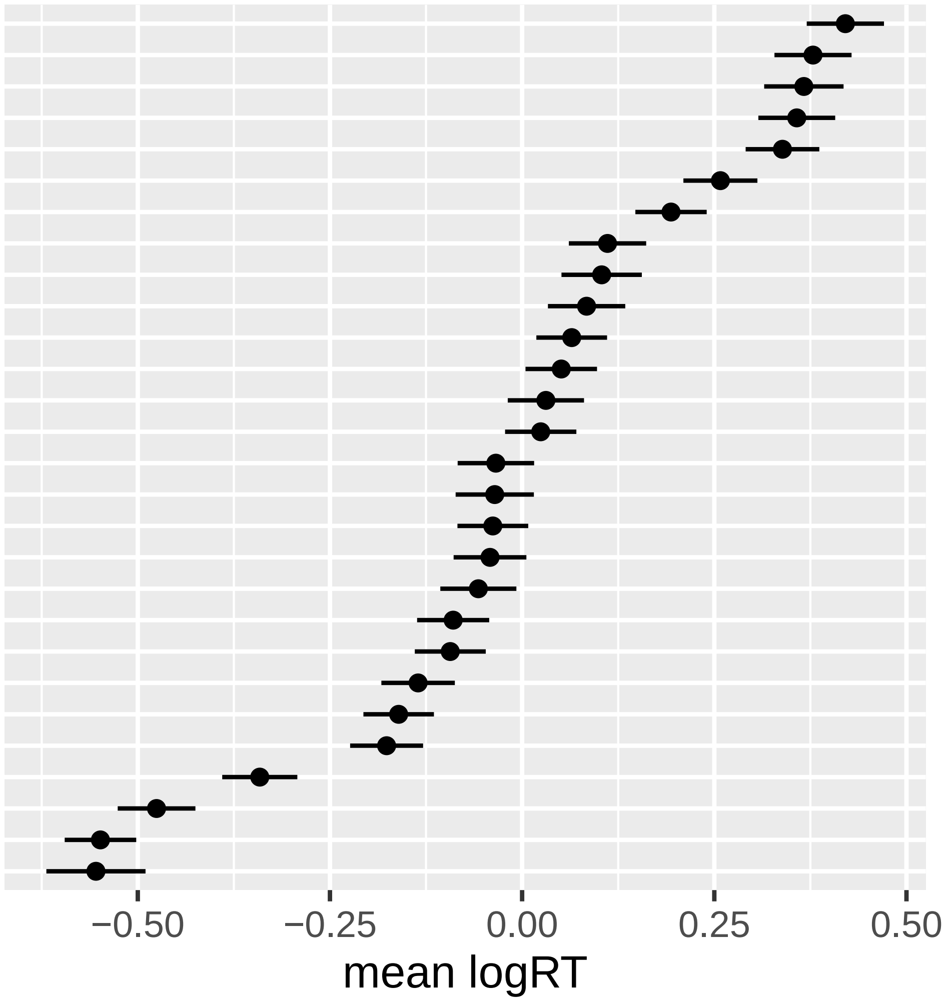


Figure 8. Size Study 1: Effect plot showing the mean log reaction time of the intercept (with 95% confidence intervals) for each participant.

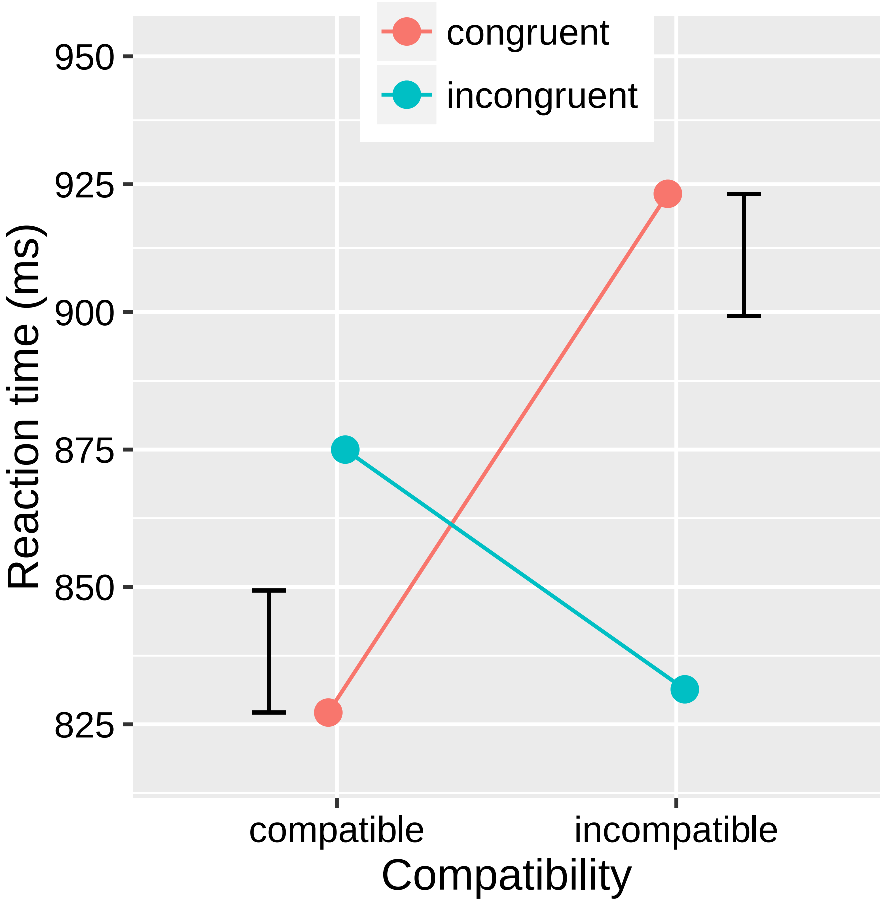


Figure 9. *Size Study 1*: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower LSD bars). Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

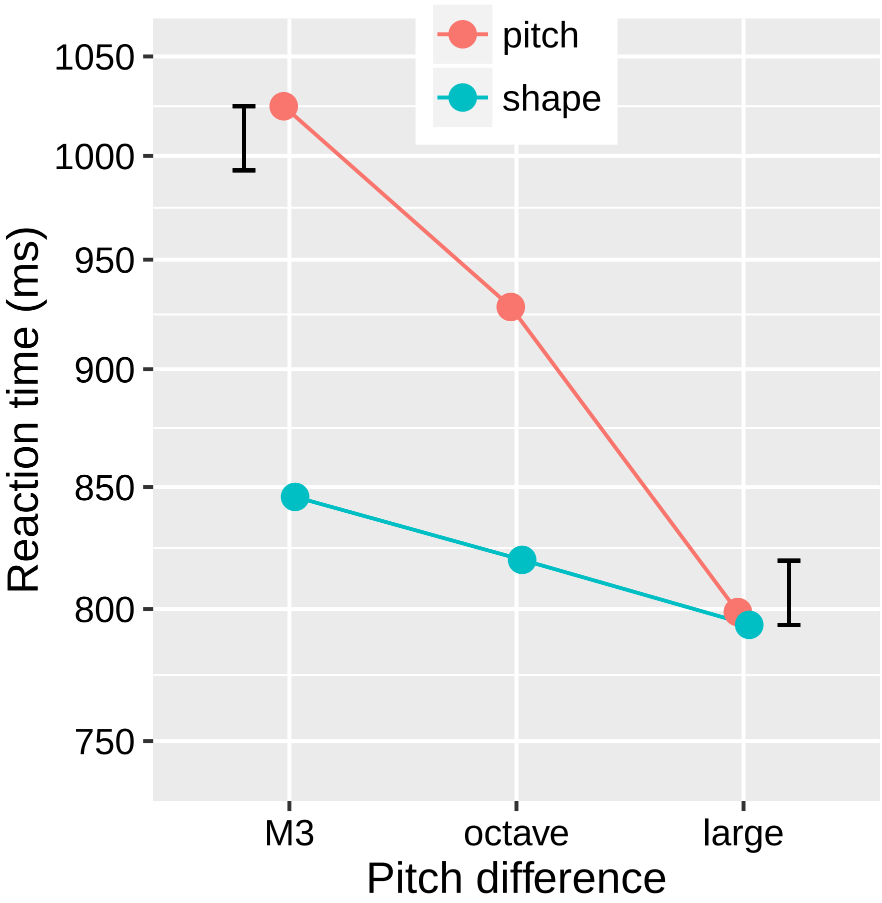


Figure 10. Size Study 1: Effect plot showing the interaction between response modality and pitch difference (with upper and lower `LSD` bars). Reaction time differed more by pitch difference when participants responded to the pitch compared to when they responded to the size of the shape.

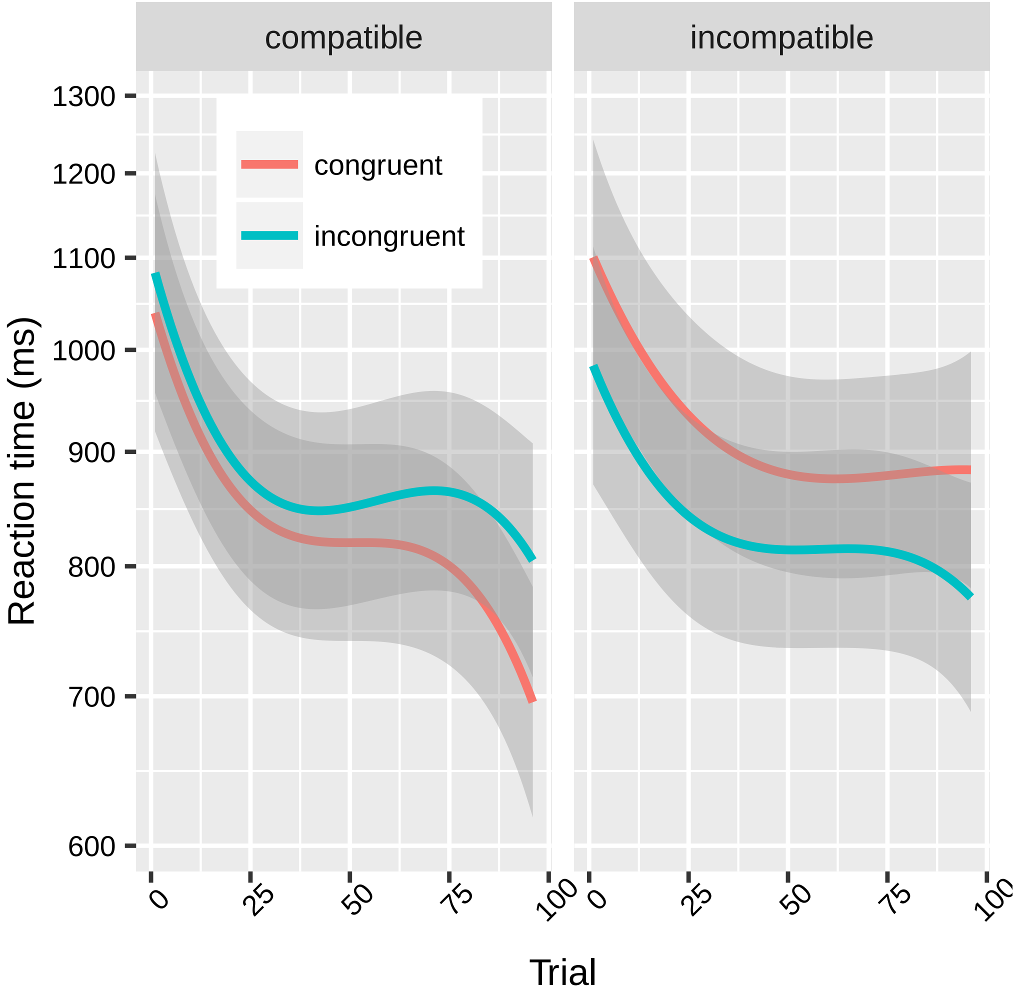


Figure 11. Size Study 1: Effect plot showing the interaction between congruency, compatibility, and trial number (with 95% confidence intervals). The difference between congruent and incongruent trials does not change across trials, running out a learning explanation.

Size Study 2: Top-down Influence [Pure tones, between-participants]

The within-subjects design of Size Study 1 may have caused too much focus on the instructions for individual participants to show an effect of perceptual congruency. Thus in Size Study 2, we used a between-subjects design where each participants heard sine tones of varying frequencies, but only received one set of instructions for the entirety of the experiment.

Method

Participants

Twenty-eight U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The visual and auditory stimuli were the same as Size Study 1.

Design and Procedure

We used a between-participant design in which each participant completed four blocks of 96 trials. Participants were randomly assigned to one of four sets of experiment instructions: larger/lower, smaller/higher, larger/higher, or smaller/lower.

The procedure was the same as Size Study 1 except the instructions appeared only on trial 1 of each of the four blocks rather than on each fixation screen.

Data management

Participants. We removed two participants with less than 60% accuracy on one or both stimulus dimensions. The average overall accuracy of the remaining 26 participants was 91.4%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (3.7% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the four fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 5.7% of responses.

Results and Discussion

Reaction Time

Model comparison. We compared an additive model including just the four fixed effects (see Size Study 1) to a non-additive model including the four fixed effects and all of their two, three, and four-way interactions. Table 4 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 378.2, which is "decisive evidence" in favor of the better model (Jeffreys, 1961).

Table 4

Size Study 2*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	18	0.0	1.00	0.071	0.326
`Additive`	`Intercept`	7	378.2	0.00	0.032	0.278

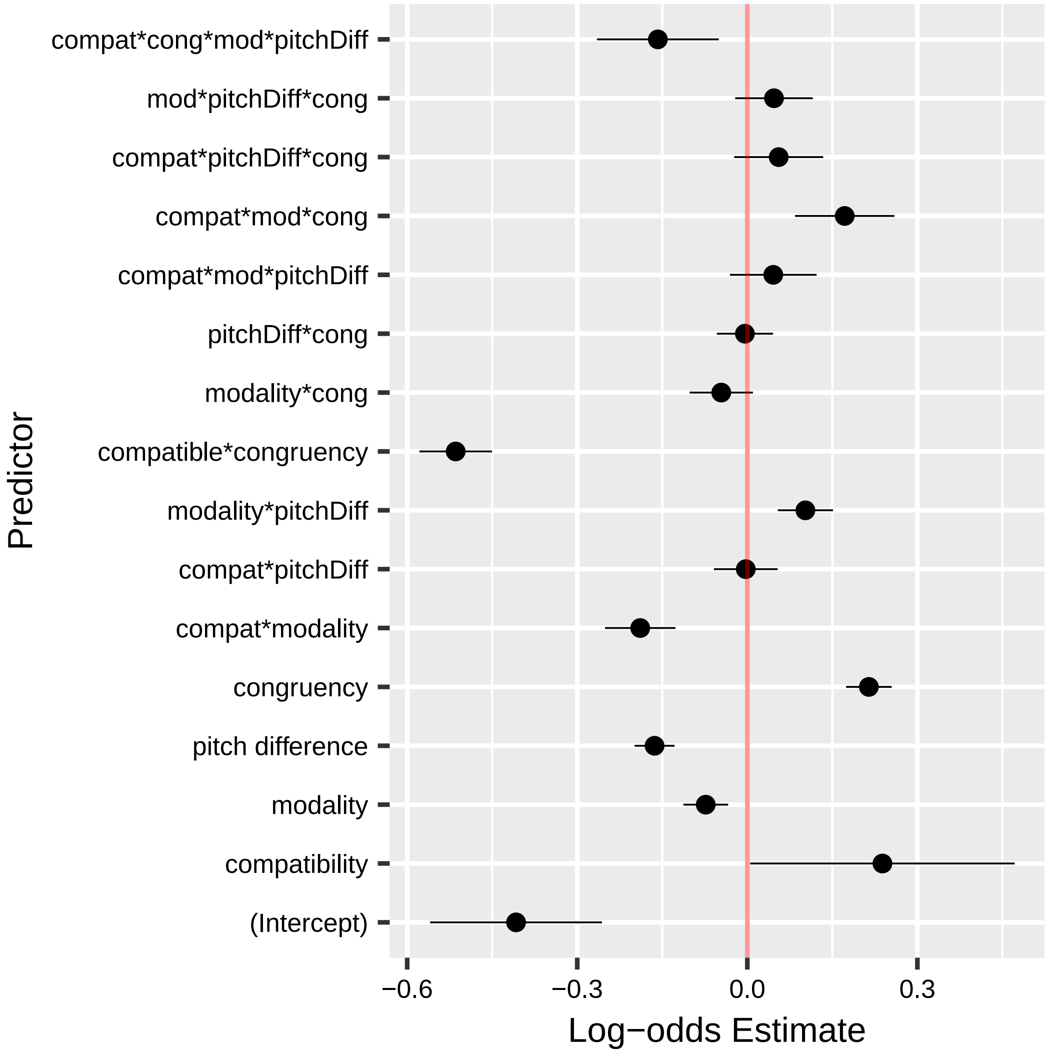


Figure 12. Size Study 2: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals). Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference.

Significant findings. Figure 12 shows the predictors of logRT and their 95% confidence intervals. The `congruency x compatibility` interaction is again the strongest and most important. Figure 13 again shows the top-down influence of instructions on performance: whatever dimensions

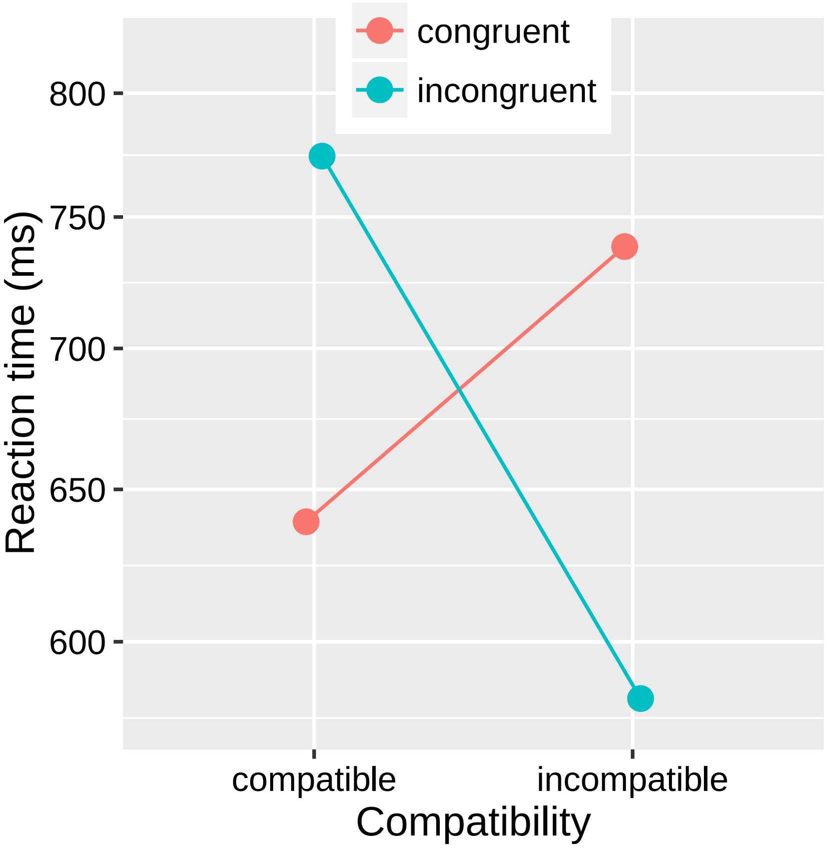


Figure 13. Size Study 2: Effect plot showing the interaction between perceptual congruency and instructions compatibility (`LSD` bars not shown because they are not easily interpreted in a between-participant design. See text for significance.). Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

are paired in the instructions for a given block, participants respond faster when those attributes are paired together. There is a significant *congruency* advantage on compatible trials (0.21, [0.18, 0.25]), and an equal *incongruency* advantage on incompatible trials (−0.3, [−0.35, −0.25]).

In addition, the interaction strength does not change as a function of trial number (−0.36, [−4.33, 3.61]), meaning reaction time decreases as a function of trial similarly across congruency and incongruent trials and compatible and incompatible blocks.

Errors

Model comparison. Because of the low error probability overall, a non-additive model including the four fixed effects (see Size Study 1) and all of their two-, three-, and four-way interactions again failed to converge. Instead, we compared a non-additive model including the four fixed effects and the `congruency × compatibility` two-way interaction to an additive model including just the four fixed effects. The AICc difference (ΔAICc) between the two models is 81.7, which is "very strong evidence" in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible trials is $p = 0.08$ ($z = 1.75$), meaning listeners make approximately the same number of errors on congruent trials regardless of compatibility condition. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 4.92$), meaning listeners make more errors on incongruent trials. These effects are converted to proportions and shown in Table 5. The more accurate conditions were also faster, meaning there is no evidence of a speed-accuracy tradeoff.

Table 5
Size Study 2*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.07	0.12
Incongruent trials	0.11	0.05

Conclusions

The main finding of Size Study 2 is that study design does not matter; a similar interaction between congruency and compatibility exists regardless of using a within-participants or between participants design. The results replicate the importance of top-down influence of instructions on performance and show that the interaction is not a function of learning an unfamiliar pairing throughout the trials.

Size Study 3: Top-down Influence [Percussion tones, within-participants]

Having established the top-down influence of instructions using pure sine tones, we next attempted to extend the effect using pitched percussion tones. Given that the resonant properties of percussion instruments give more indication of their relative size than pure tones (Carello et al., 1998; Coward & Stevens, 2004; Grassi, 2005), it is possible that the natural pitch-size mapping will be harder to modify with percussion tones than sine tones. Alternatively, if the `congruency x compatibility` interaction remains in Size Studies 3 (within-participants) and 4 (between-participants), we will have stronger evidence that the pitch-size correspondence relies more on top-down than bottom-up processing.

Method

Participants

Twenty-eight U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The visual stimuli were the same as Size Studies 1–2.

Percussion tones. We used six different percussion instrument samples downloaded from `http://www.freewavesamples.com`: woodblock, high bongo, high conga, low conga, low-mid tom, and low tom. We analyzed the frequency spectrum of each wave file using Audacity software (`http://audacityteam.org`) in order to determine the fundamental frequency of each sample. We then created three pairings of varying pitch differences: ‘large’ (woodblock: 2468 Hz vs. low tom: 142 Hz), ‘mid’ (high conga: 752 Hz vs. low conga: 183 Hz), and ‘small’ (high bongo: 285 Hz vs. low-mid tom: 143 Hz). All samples were played between 66-72 dB.

Design and Procedure

We used a within-participant design identical to Size Study 1. The procedure was the same as Size Study 1.

Data management

Participants. We removed three participants with less than 60% accuracy on one or both stimulus dimensions and two participants who completed fewer than 60% of the total trials. The average overall accuracy of the remaining 23 participants was 91.4%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (5.6% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the four fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 5% of responses.

Table 6
Size Study 3*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	18	0.0	1.00	0.053	0.238
`Additive`	`Intercept`	7	288.0	0.00	0.013	0.195

Results and Discussion

Reaction Time

Model comparison. We compared an additive model including just the four fixed effects (see Size Study 1) to a non-additive model including the four fixed effects and all of their two, three, and four-way interactions. Table 6 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 288; "decisive evidence" in favor of the better model (Jeffreys, 1961).

Significant findings. Figure 14 shows the predictors of logRT and their 95% confidence intervals. From the figure, we can see that there is only one significant interaction: `congruency x compatibility`. This interaction is a replication of Size Studies 1–2: whatever dimensions are paired in the instructions for a given block, participants respond faster when those attributes are paired together (Figure 15). There is a significant *congruency* advantage on compatible trials (0.27, [0.22, 0.33]), and an equal *incongruency* advantage on incompatible trials (−0.29, [−0.34, −0.23]).

In addition, the interaction strength does not change as a function of trial number (2.3, [−1.16, 5.75]), meaning reaction time decreases as a function of trial similarly across congruency and incongruent trials and compatible and incompatible blocks.

A final finding of interest is that the effect of pitch difference is *not* significant (−0.05, [−0.1, 0]) when the tones were pitched percussion timbres, whereas it played a much larger role when the tones were sine tones of varying frequencies in Size Studies 1–2.

Errors

Model comparison. Because of the low error probability overall, a non-additive model including the four fixed effects (see Size Study 1) and all of their two-, three-, and four-way interactions again failed to converge. Instead, we compared a non-additive model including the four fixed effects and the `congruency × compatibility` two-way interaction to an additive model including just the four fixed effects. The AICc difference (ΔAICc) between the two models is 23.6, which is "strong evidence" in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible trials is $p \approx 0$ ($z = 5.08$), meaning listeners make more errors on the incompatible blocks. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 4.15$), meaning listeners make more errors on incongruent trials. These effects are converted to proportions and shown in Table 7. The more accurate conditions were also faster, meaning there is no evidence of a speed-accuracy tradeoff.

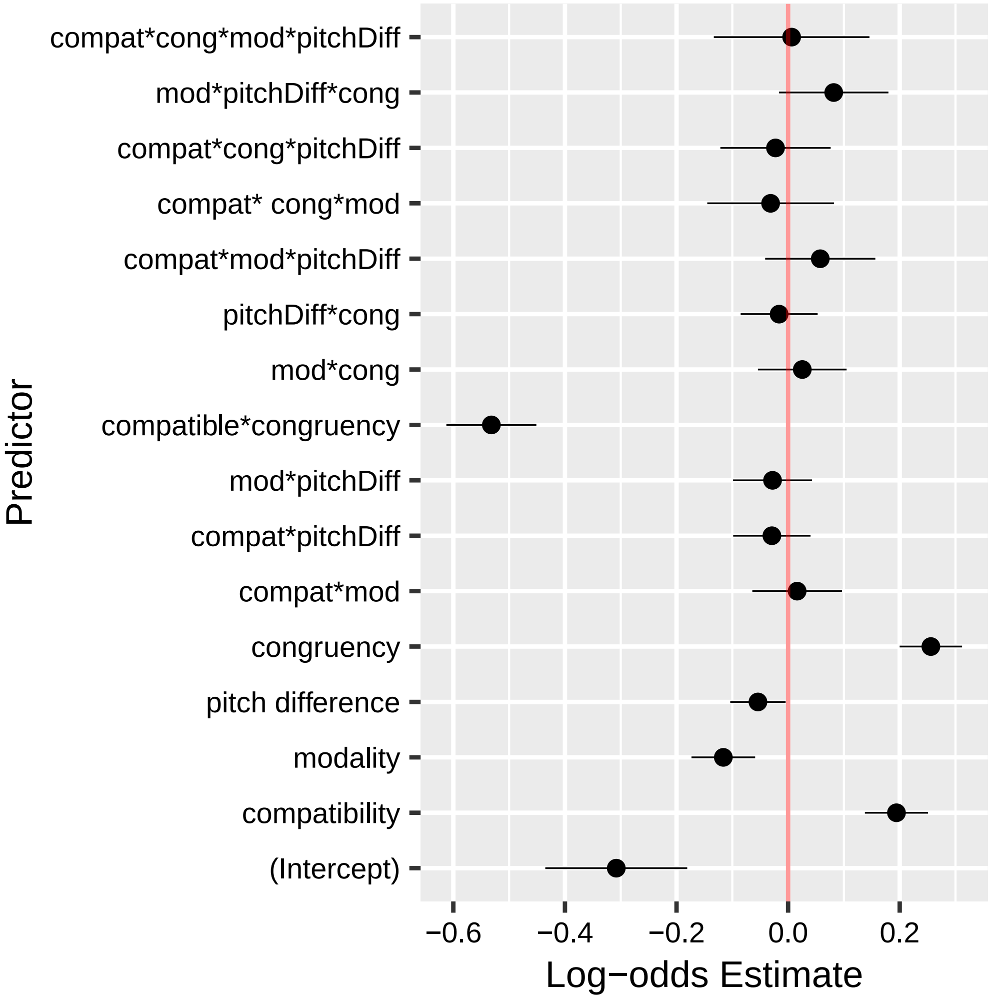


Figure 14. Size Study 3: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals). Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference.

Table 7
Size Study 3*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.05	0.09
Incongruent trials	0.08	0.07

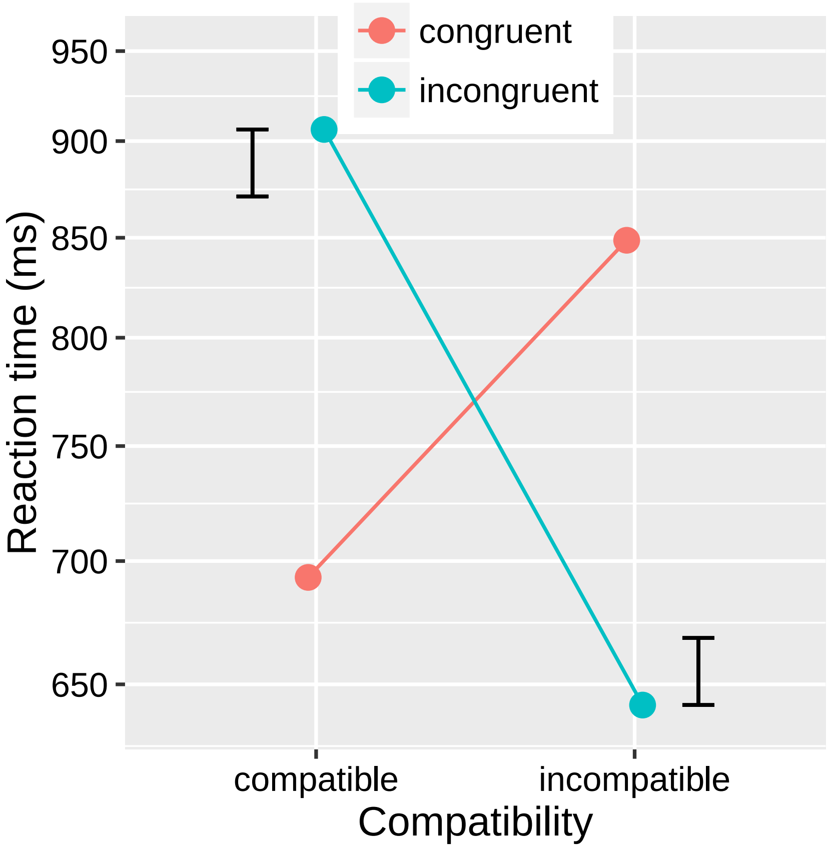


Figure 15. Size Study 3: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower LSD bars). Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

Conclusions

The main finding of Size Study 3 is that timbre does not matter; a similar interaction between congruency and compatibility exists regardless of using sine tones or pitched percussion tones. The results replicate the importance of top-down influence of instructions on performance and show that the interaction is not a function of learning an unfamiliar pairing throughout the trials.

Size Study 4: Top-down Influence [Percussion tones, between-participants]

Given that Size Studies 1 and 2 looked at differences between within-participants and between-participants design using sine tones, here we sought to establish a similar comparison between within-participants (Size Study 3) and between-participants design using pitched percussion tones.

Method

Participants

Forty-one U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The visual stimuli were the same as Size Studies 1–3. The auditory stimuli were the same as Size Study 3.

Design and Procedure

We used a between-participant design identical to Size Study 2. The procedure was the same as Size Studies 1–3.

Data management

Participants. We removed fifteen participants with less than 60% accuracy on one or both stimulus dimensions.[9] The average overall accuracy of the remaining 26 participants was 91.3%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (1.7% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the four fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, removing an additional 5.3% of responses.

Results and Discussion

Reaction Time

Model comparison. We compared an additive model including just the four fixed effects (see Size Study 1) to a non-additive model including the four fixed effects and all of their two, three, and four-way interactions. Table 8 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the best model and the next is 117.1; "decisive evidence" in favor of the better model (Jeffreys, 1961).

Significant findings. In every sense, Size Study 4 replicates Size Study 3. Figure 16 shows the predictors of logRT and their 95% confidence intervals; the only significant interaction is `congruency x compatibility`. Again, this interaction shows that whatever dimensions are paired in the instructions for a given block, participants respond faster when those attributes are paired together (Figure 17). There is a significant *congruency* advantage on compatible trials (0.13,

[9] It is unclear why so many participants did poorly on this version of the experiment, other than that the experiment was conducted at the end of the semester.

Table 8
Size Study 4*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	18	0.0	1.00	0.026	0.251
`Additive`	`Intercept`	7	117.1	0.00	0.007	0.233

[0.09, 0.17]), and an equal *incongruency* advantage on incompatible trials (−0.15, [−0.2, −0.1]). Further, the interaction strength does not change as a function of trial number (−1.48, [−5.3, 2.33]), meaning reaction time decreases as a function of trial similarly across congruency and incongruent trials and compatible and incompatible blocks. Finally, the effect of pitch difference was not significant (−0.03, [−0.07, 0]).

Errors

Model comparison. Because of the low error probability overall, a non-additive model including the four fixed effects (see Size Study 1) and all of their two-, three-, and four-way interactions again failed to converge. Instead, we compared a non-additive model including the four fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the four fixed effects. The AICc difference (ΔAICc) between the two models is 73.3, which is "very strong evidence" in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible trials is $p = 0.07$ ($z = 1.83$), meaning listeners make approximately the same number of errors on congruent trials regardless of compatibility condition. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 4.92$), meaning listeners make more errors on incongruent trials. These effects are converted to proportions and shown in Table 5. The more accurate conditions were also faster, meaning there is no evidence of a speed-accuracy tradeoff.

Table 9
Size Study 4*: Proportion of errors across compatibility and congruency conditions.*

	`Compatible blocks`	`Incompatible blocks`
`Congruent trials`	0.06	0.11
`Incongruent trials`	0.10	0.06

Conclusions

Size Study 4 completes the series of studies looking at the top-down influence of instructions on the pitch–size mapping. This study replicates Size Studies 1–3, thus providing strong evidence that instructions matter more than perceptual congruency. The natural pitch–size correspondence found in previous studies is thus not the result of solely bottom-up processing, but instead can be overridden easily with top-down control of attention.

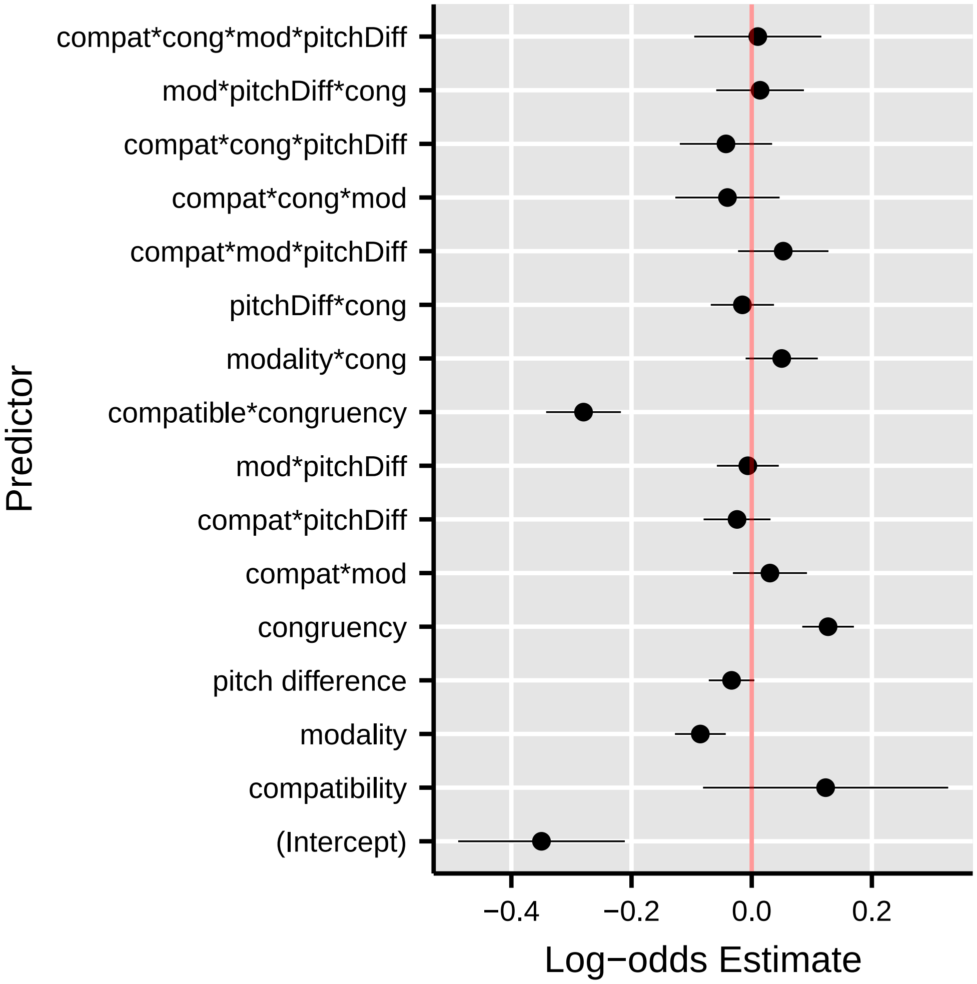


Figure 16. *Size Study 4*: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals). Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference.

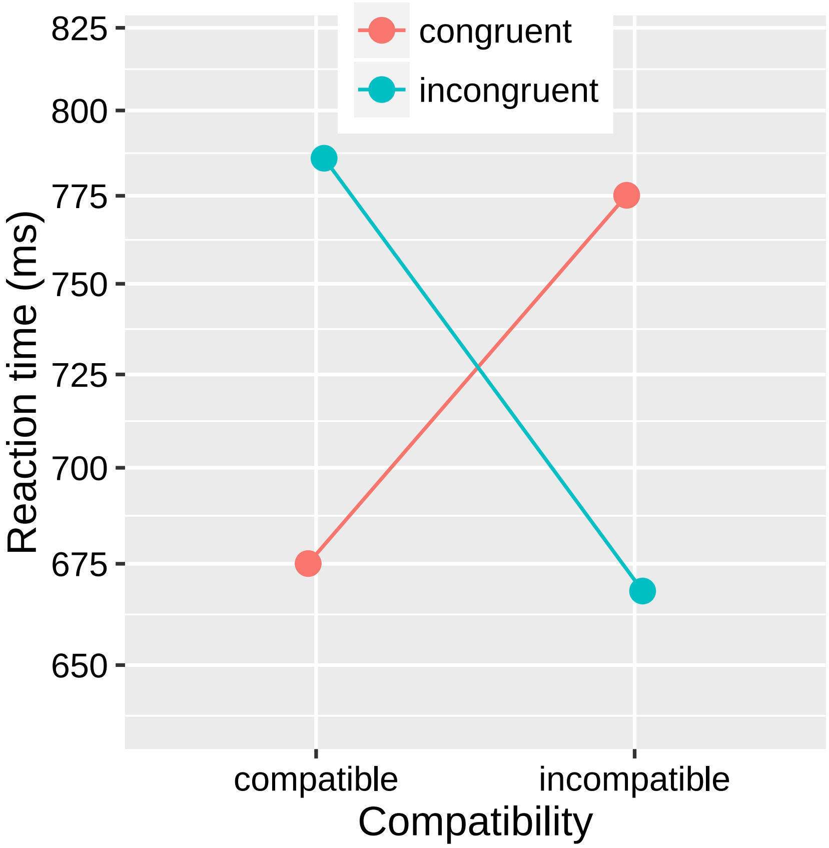


Figure 17. *Size Study 4*: Effect plot showing the interaction between perceptual congruency and instructions compatibility. (LSD bars not shown because they are not easily interpreted in a between-participant design. See text for significance.) Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

Size Study 5: Size–Relevant Conceptual Replication

Because the new methodology used in Size Studies 1–4 showed no evidence for an automatic correspondence between pitch and size, but rather that participants responded faster to whatever they were told to pay attention to, we wanted to verify that the effect replicates with one relevant dimension. It is possible that the modified task changed the task too much or made it too difficult. However, we would expect the perceptual congruency effect to remain in a simple speeded classification task. Size Study 5 was a conceptual replication of the pitch–size correspondence similar to the procedure used in Size Studies 1–4 (Figure 2) except pitch was always the irrelevant dimension and object size was always the relevant dimension.

Method

Participants

Eleven U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory stimuli were the same as Size Studies 1–2. The visual stimuli included one randomly-shaped blob of 200, 250, 300, and 350 pixels in area, creating 6 total size comparisons. The comparisons were recoded to include differences of 50, 100, and 150 pixels in area.

Design and Procedure

The procedure followed the sequence of events presented in Figure 2 with one exception: instead of waiting until the end of the trial to discover whether they should respond to the pitch or the shape, participants always responded to the size of the shape and pitch was always the irrelevant dimension. Each participant completed two blocks of 108 trials (in a random order): (a) Participants were asked to select whether the first or second shape was *larger*; (b) participants were asked to select whether the first or second shape was *smaller*. After the second stimulus appeared for 300 ms, it was replaced by a blank screen until the participant responded. Participants were told they could respond as soon as the second shape appeared. They were also informed that the task-irrelevant pitch would only appear on some trials and others would have no sound.

Data management

Participants. We did not need to remove any participants with less than 60% accuracy. The overall accuracy of the 11 participants was 95.5%.

Reaction Time. There were no responses that were faster than 50ms. We performed a Box-Cox analysis on the reaction time (RT) data; the result suggested an inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (5.8% removed).

Table 10
Size Study 5*: Comparison of four models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive: sizeDiff factor`	`Intercept`	8	0.0	0.89	0.048	0.293
`Additive: sizeDiff continuous`	`Intercept`	7	4.1	0.11	0.046	0.292
`Non-additive: sizeDiff factor`	`Intercept`	20	75.3	0.00	0.049	0.293
`Non-additive: sizeDiff continuous`	`Intercept`	14	93.7	0.00	0.047	0.292

Results and Discussion

Reaction Time

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `task` (respond to larger shape vs. smaller shape), and `size difference` (continuous or factored [50, 100, 150]). We also included the subject-by-subject variation in the intercept as a random effect.

We compared four LMM models: two additive models including just the three fixed effects (with size difference as a factor or continuous predictor) and two non-additive models (with size difference factored or continuous) including the three fixed effects and all of their two and three-way interactions. Table 10 shows that the additive models are clearly superior to the non-additive models; further the additive model with size difference as a factored predictor is the best model. The AICc difference (ΔAICc) between the best model and the next is 4.1; a "substantial evidence" in favor of the better model (Jeffreys, 1961).

Significant findings. We found no evidence for a congruency effect collapsing across size differences (Figure 18). Instead we found that participants were significantly *faster* to respond in the no sound condition (−0.08, [−0.11, −0.05]) than to either the congruent or incongruent sound conditions, which did not differ from each other (−0.02, [−0.05, 0.01]).

We also found a significant effect of size difference between the shapes (−0.002, [−0.002, −0.001]), such that participants were slowest to respond to a 50-pixel size difference and fastest to respond to a 150-pixel size difference, showing that participants were in fact completing the experiment as instructed (Figure 19).

Errors

Our binomial GLMM included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `task` (respond to larger shape vs. smaller shape), and `size difference` (50 vs. 100 vs. 150). We also included the subject-by-subject variation in the intercept as a random effect. Because of the low error probability overall, a non-additive model failed to converge. We chose to fit an additive model with the three categorical fixed effects only because this model was clearly superior for the RT analysis as well.

There was no difference between the log-odds of correct responses for the congruent and incongruent conditions ($p = 0.6$; $z = 0.52$) or for the congruent and unimodal conditions ($p = 0.26$; $z = 1.13$). The proportion of errors in the congruent, incongruent, and unimodal conditions (responding to larger shapes with a 50-pixel size difference) was 0.048, 0.054, and 0.062, respectively.

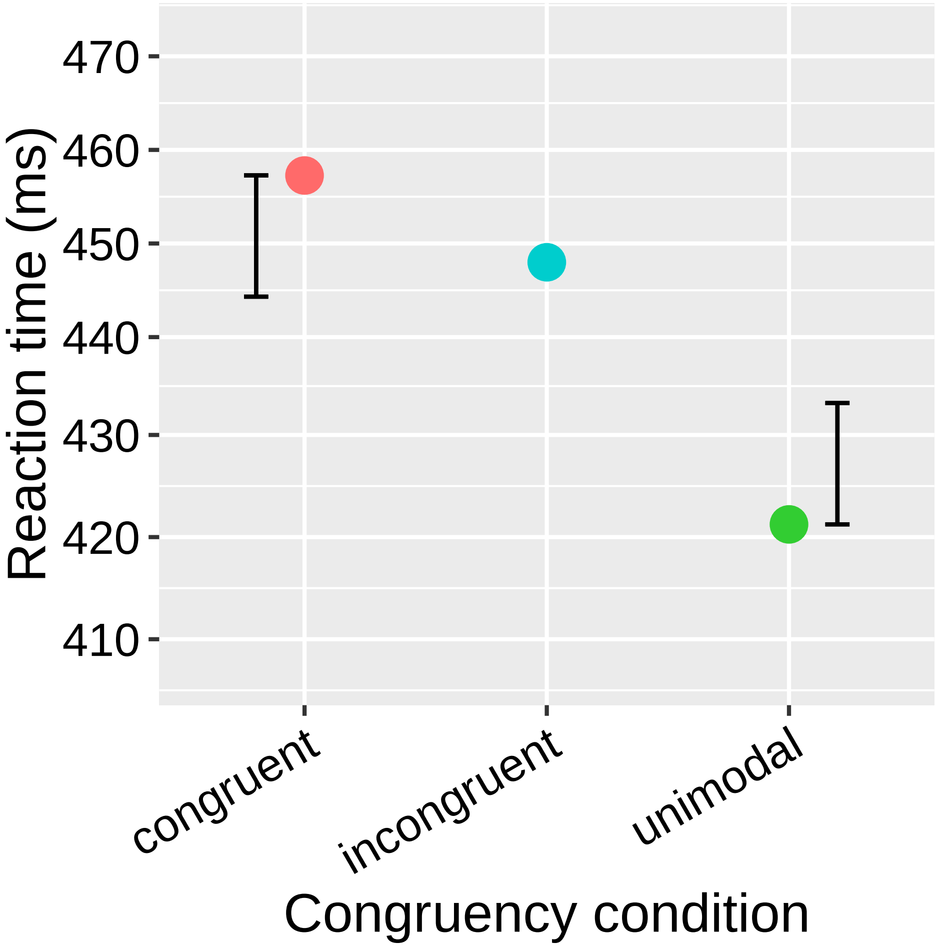


Figure 18. *Size Study 5*: Effect plot showing the main effect of perceptual congruency (with upper and lower `LSD` bars). Participants respond slower with any sound (congruent or incongruent) compared to no sound.

Conclusions

We were unable to replicate previous results using a slightly different methodology; we found no evidence that a congruent pairing between object size and auditory pitch resulted in faster processing. The results here therefore support the results of Size Studies 1–4 in showing that there is no automatic association between pitch and size.

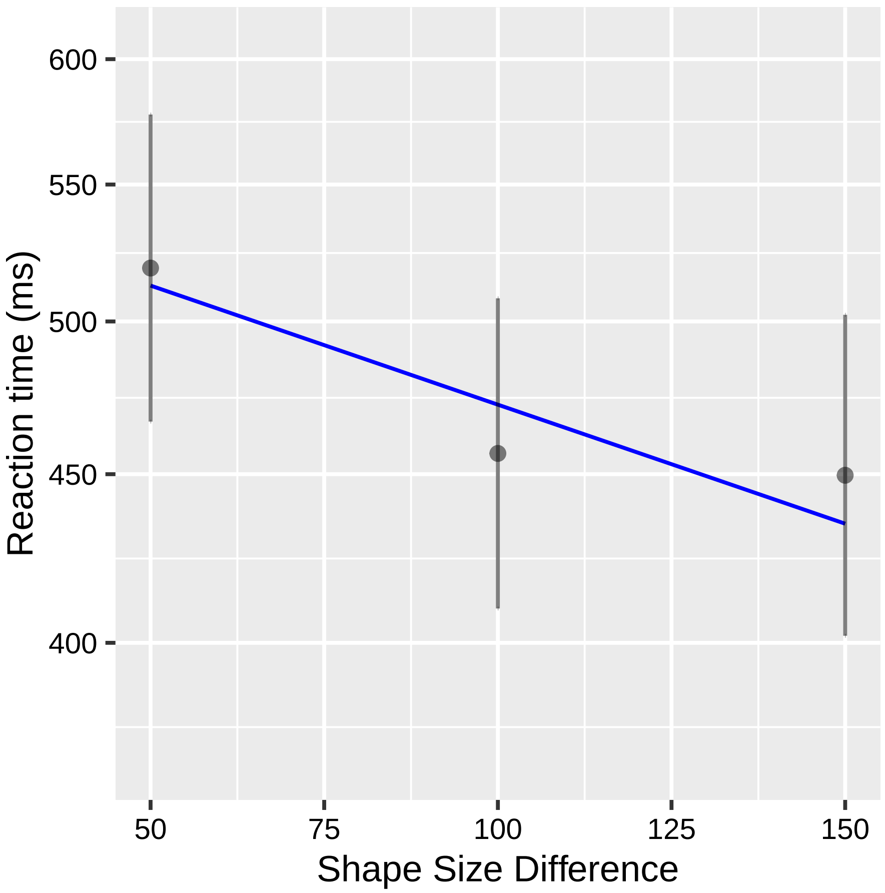


Figure 19. Size Study 5: Effect plot showing the main effect of size difference (factored) between the two visual stimuli (with 95% confidence intervals). The blue line shows the line of best-fit showing the effect of size difference (continuous). Participants respond fastest to the largest size difference and slowest to the smallest size difference between shapes.

Size Study 6: Size–Relevant Direct Replication

We made a number of small changes from the original Gallace and Spence (2006) study that may have led to our different results:

- They used one (exceedingly large) pitch difference; we used three pitch differences.
- They used a comparison circle that was the same on each trial; we used four differently-sized shapes, the order of which was randomized on each trial.
- They only included a pitch with the second shape; we included a pitch with both shapes.
- They used a masking screen between shapes because both shapes were presented in the center of the screen; we used a blank screen and shapes were presented left and right of screen center.
- They only gave participants 3,000 ms to respond to each trial and had no break between trials; we gave participants unlimited time to respond, and they had to press the spacebar to continue with each new trial.
- Their participants heard the sounds from speakers; our participants heard the sounds over headphones.

This necessitated we go back to complete a direct replication to verify if the effect exists under very limited conditions. Size Study 6 was thus a direct replication of Gallace and Spence (2006), except that some participants heard the sounds through speakers (as in the original) and some heard the sounds through headphones. This was done to assure potential binding of pitch and size dimensions was not affected by changing the apparent location of the sound (and thus changing potential causality; Schutz & Kubovy, 2009).

Method

Participants

Forty-one U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

Visual shapes. We used a gray standard circle of 185 pixels (5.5 cm) in diameter. In addition, we used variable-sized circles that were ± 5%, ± 10%, ± 20%, and ± 40% of the diameter of the standard.

Auditory pitches. On each trial, the variable-sized circle was presented together with a low 300 Hz tone, a high (4500 Hz) tone, or no tone. All tones were pure sine tones and were 300 ms in duration.

Design and Procedure

Figure 1a shows an illustration of the sequence of events presented in each trial. At the start of a trial, a red fixation dot 10 pixels (0.3 cm in diameter) was presented in the center of the screen for 300 ms. This was followed by a 300 ms blank white screen. The first shape (the gray 'standard' circle) was presented at the screen's center for 300 ms, followed by a random dot mask that filled the entire screen for 500 ms. The second shape (the gray variable-sized 'comparison' circle) was next presented for 80 ms at a random position (ranging ±10 pixels vertically and horizontally from the center of the screen to prevent the participant from being able to superimpose the two circles to determine the size difference), accompanied by a high tone, low tone, or no tone. A second random dot mask then remained on the screen until a response was made or until 3,000 ms had elapsed, at which point a new trial began.

Participants were instructed to press one key if the comparison circle appeared to be larger than the standard, and a different key if the comparison circle appeared to be smaller than the standard. They were told that on some trials a task-irrelevant sound would accompany the comparison circle, but they were asked to ignore it as much as possible. They were also told that the first disk would always been a standard size.

Each of the eight different variable-sized disks was presented 10 times in each of the three sound conditions (high frequency, low frequency, and no sound), giving rise to a total of 240 trials, which were presented in 2 blocks of 120 trials or 10 blocks of 24 trials.

Data management

Participants. We removed four participants with less than 60% accuracy (including no responses and incorrect responses). The average overall accuracy of the remaining 37 participants was 83.5%.

Reaction Time. First, we eliminated incorrect trials (13.7% of responses removed) and trials where participants failed to respond in the allotted 3,000 ms (2.9% removed). We then eliminated responses that were faster than 50ms (0.08% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested an inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (6.6% removed).

Results and Discussion

Reaction Time

Table 11
Size Study 6*: Comparison of four models predicting reaction time, ordered by* Δ*AICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive: sizeDiff factor`	`Intercept`	13	0.0	1.00	0.071	0.391
`Additive: sizeDiff quadratic`	`Intercept`	8	13.4	0.00	0.064	0.386
`Non-additive: sizeDiff quadratic`	`Intercept`	20	29.4	0.00	0.065	0.386
`Non-additive: sizeDiff factor`	`Intercept`	50	232.4	0.00	0.074	0.393

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `audio` (headphones vs. speakers), and `size difference` (factored or quadratic). We also included the subject-by-subject variation in the intercept as a random effect.

We compared four models: two models were additive, including just the three fixed effects (with size difference as a factored or quadratic predictor), and two models were non-additive, including the three fixed effects and all of their two and three-way interactions (with size difference factored or quadratic). Table 11 shows that the additive model with factored size difference is clearly superior to the other models. The AICc difference (ΔAICc) between the best model and the next model is 13.4; "strong evidence" in favor of the better model (Jeffreys, 1961).

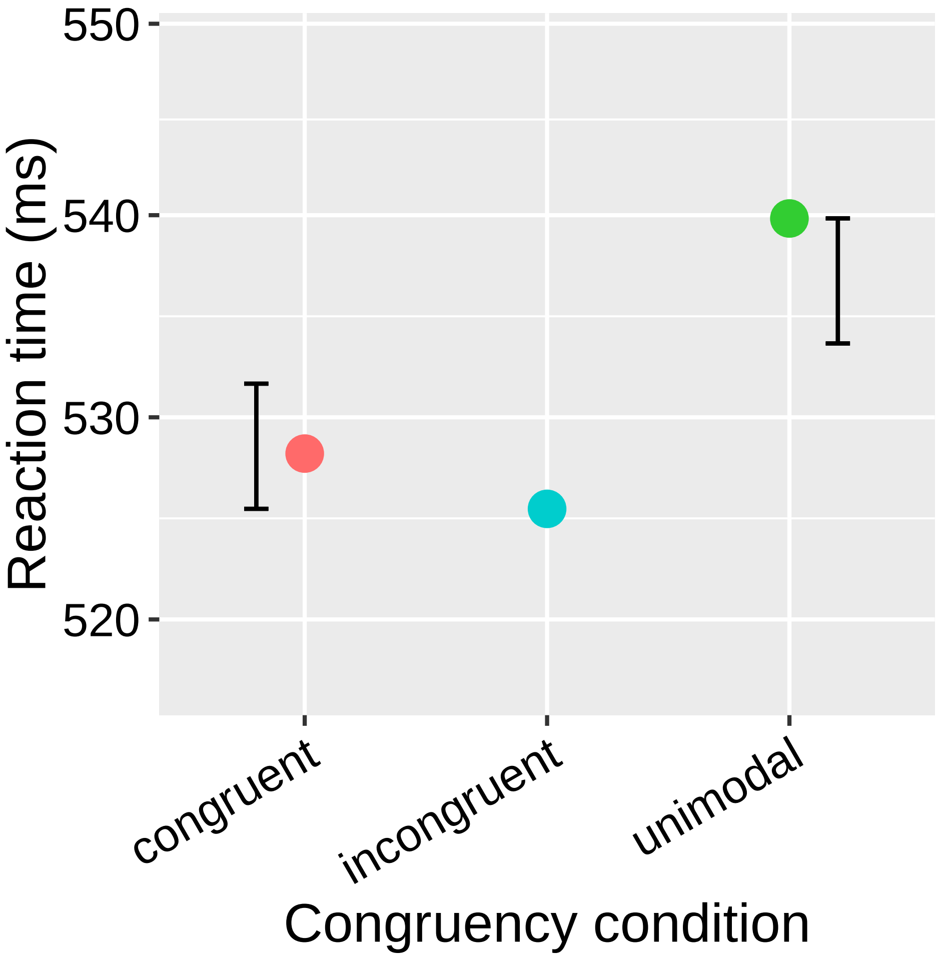


Figure 20. Size Study 6: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars). Participants respond faster with any sound (congruent or incongruent) compared to no sound.

Significant findings. We found no evidence for a congruency effect collapsing across size differences (Figure 20). Instead we found that participants were significantly *slower* to respond in the no sound condition (0.02, [0.01, 0.03]) than to either the congruent or incongruent sound conditions, which did not differ from each other (−0.01, [−0.02, 0.01]).

We also performed two manipulation checks. *First,* there was a significant effect of size difference between the shapes, such that participants were slowest to respond to a ±5% size difference and fastest to respond to a ±40% size difference, showing that participants were in fact completing the experiment as instructed (Figure 21). *Second,* there was no difference between participants who heard the sounds over headphones and those who heard the sounds through speakers (−0.05, [−0.16, 0.06]), validating our use of headphones.

Repeated-measures ANOVA. To confirm that the differences between my results and Gallace and Spence (2006) were not solely due to analysis differences, I completed a replication of their analysis. The (non-transformed) reaction time data (including correct and incorrect responses) were

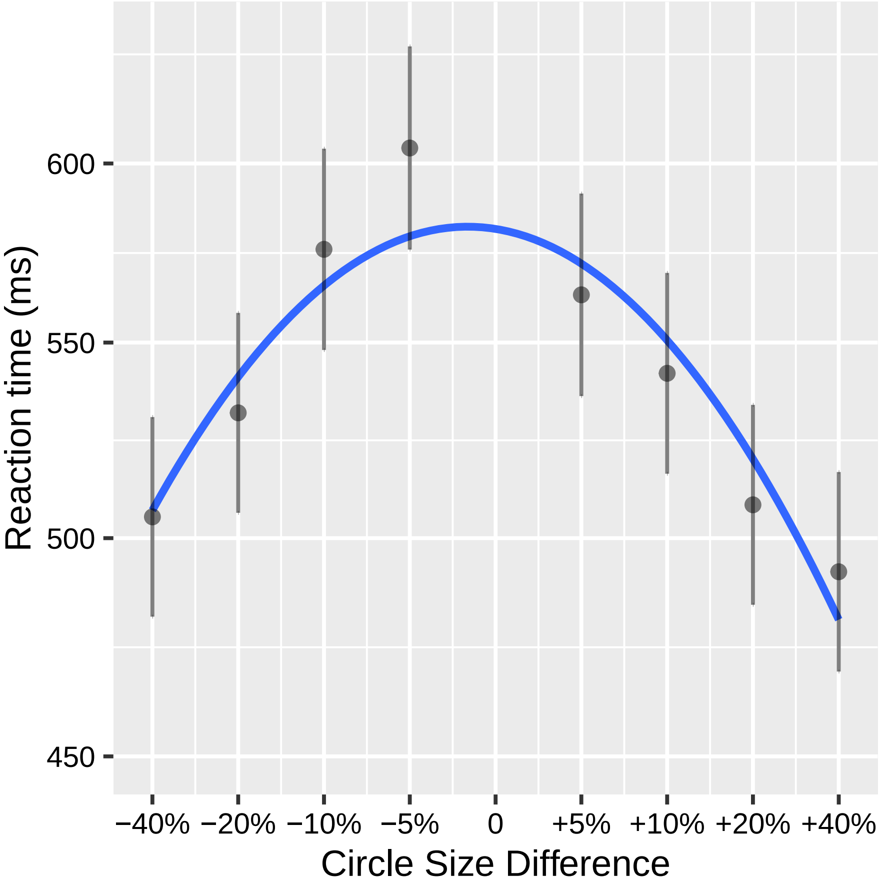


Figure 21. Size Study 6: Effect plot showing the main effect of size difference between the standard and comparison circles (with 95% confidence intervals). The blue line is a quadratic line of best-fit showing the effect of size difference. Participants respond fastest to the largest size difference and slowest to the smallest size difference between circles.

submitted to an 8 × 3 repeated-measures ANOVA with stimulus size difference and congruency as factors.

The analysis of the RT data revealed a significant main effect of congruency [$F(2,72) = 5.57$, $p = 0.006$]. Post-hoc comparison tests using the Benjamini and Hochberg (1995) method of adjustment revealed significant differences between the congruent condition and the unimodal condition ($p = 0.01$) and between the incongruent and unimodal conditions ($p = 0.001$), but *not* between the congruent and incongruent conditions ($p = 0.335$).

The analysis also revealed a significant man effect of size difference [$F(7,252) = 24.51$, $p < 0.001$], with reaction times decreasing with increasing difference between the standard and variable circles. Additionally, there was no interaction between size difference and congruency [$F(14,504) = 0.99$, $p = 0.459$].

These results match the results of the LMM analysis with transformed reaction times, which means the non-significant congruency effect is not solely a function of our analysis method.

Errors

Our binomial GLMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `audio` (headphones vs. speakers), and `size difference` (factored or quadratic). We also included the subject-by-subject variation in the intercept as a random effect. Because of the low error probability overall, a non-additive model including factored size difference failed to converge. Instead, we compared a non-additive model including size difference as a quadratic predictor with the two- and three-way interactions between congruency and audio to additive models with size difference factored or quadratic. The additive factored size difference model was clearly superior; the AICc difference (ΔAICc) between the top two models was 6.7, which is "substantial evidence" in favor of the better model (Jeffreys, 1961).

There was no difference between the log-odds of correct responses for the congruent and incongruent conditions ($p = 0.1$; $z = -1.64$) or for the congruent and unimodal conditions ($p = 0.13$; $z = -1.5$). The proportion of errors in the congruent, incongruent, and unimodal conditions (in the headphones condition at -40% size difference) was 0.041, 0.036, and 0.037, respectively.

Conclusions

We were again unable to replicate previous findings, this time using a direct replication. Instead of a congruency advantage, we found an advantage for size judgments in the presence of any sound compared to no sound. Thus it appears that the pitch–size correspondence is either weak or nonexistent as a perceptual phenomenon and instead relies only on top-down processing. It is also an open question why the `no sound` condition switched from being fastest in the conceptual replication to slowest in the direct replication.

Size Study 7: Pitch–Relevant Replication

In Size Studies 5–6, I found no evidence for a correspondence between pitch and size when pitch was the irrelevant dimension. Size Study 7 used Gallace and Spence's (2006) methodology, but with size always the irrelevant dimension and pitch always the relevant dimension. Participants were asked to make pitch height judgments while simultaneously viewing irrelevant circles of varying sizes.

Method

Participants

Twenty-nine U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The visual stimuli were the ±40% stimuli (111 pixels vs. 259 pixels) from Size Study 6. For the auditory stimuli, we used a standard pitch of 600 Hz. In addition, we used pitches that were ±1, ±3, ±6, and ±12 semitones from 600 Hz. All tones were pure sine tones and were 300 ms in duration.

Design and Procedure

The procedure was adapted from Figure 1a; in this experiment, instead of a standard circle, there was a standard pitch (600 Hz). During the presentation of the comparison pitch, a small circle, large circle, or no circle appeared simultaneously. Participants were instructed to press one key if the comparison pitch appeared to be lower than the standard, and a different key if the comparison pitch appeared to be higher than the standard.

Each of the eight different variable-frequency pitches was presented 10 times in each of the three shape conditions (large circle, small circle, and no circle), giving rise to a total of 240 trials, which were presented in 10 blocks of 24 trials.

Data management

Participants. We removed ten participants who completed less than 60% of the trials.[10] The average overall accuracy of the remaining 19 participants was 83.7%.

Reaction Time. First, we eliminated incorrect trials (9.2% of responses removed) and trials where participants failed to respond in the allotted 3,000 ms (7.1% removed). We then eliminated responses that were faster than 50ms (0.1% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested a squared inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (12.2% removed).

[10] It is unclear why so many participants did poorly on this version of the experiment, other than the experiment moved on after 3 seconds whether or not the participant had responded, whereas in Size Studies 1–5, they had unlimited time to respond).

Results and Discussion

Reaction Time

Table 12
Size Study 7*: Comparison of four models predicting reaction time, ordered by* Δ*AICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive pitchDiff: quadratic`	`Intercept`	7	0.0	0.98	0.039	0.247
`Additive pitchDiff: factor`	`Intercept`	11	7.7	0.02	0.039	0.247
`Non-additive pitchDiff: quadratic`	`Intercept`	12	49.4	0.00	0.041	0.249
`Non-additive: pitchDiff factor`	`Intercept`	26	149.0	0.00	0.043	0.250

Model comparison. Our LMMS included two fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal) and `pitch difference` (factored or quadratic). We also included the subject-by-subject variation in the intercept as a random effect.

We compared four models: two models were additive, including just the two fixed effects (with pitch difference as a factored or quadratic predictor), and two models were non-additive, including the two fixed effects and their interaction (with pitch difference factored or quadratic). Table 12 shows that the additive model with quadratic pitch difference is clearly superior to the other models. The AICc difference (ΔAICc) between the best model and the next model is 7.7; "substantial evidence" in favor of the better model (Jeffreys, 1961).

Significant findings. We found no evidence for a congruency effect collapsing across pitch differences (Figure 22). Participants responded to the incongruent (-0.01, $[-0.02, 0.01]$) and unimodal (0.01, [0, 0.02]) conditions with equal reaction time compared to the congruent condition.

We found a significant quadratic effect of pitch difference (-2.14, $[-2.47, -1.81]$) such that participants were slowest to respond to a ± 1-semitone pitch difference and fastest to respond to a ± 12-semitone pitch difference, showing that participants were in fact completing the experiment as instructed (Figure 23).

Errors

Our binomial GLMM included two fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal) and `pitch difference` (quadratic). We also included the subject-by-subject variation in the intercept as a random effect. Because of the low error probability overall, a non-additive model failed to converge. We chose to fit an additive model with the two fixed effects only since this model was clearly superior for the RT analysis as well.

There was no difference between the log-odds of correct responses for the congruent and incongruent conditions ($p = 0.26$; $z = -1.12$) or for the congruent and unimodal conditions ($p = 0.07$; $z = -1.79$). The proportion of errors in the congruent, incongruent, and unimodal conditions was 0.06, 0.052, and 0.048, respectively.

Conclusions

Despite a significant effect of pitch difference, we again found no evidence for a congruency effect. This study complements Size Studies 5–6 where size was the relevant dimension; here pitch was the relevant dimension, but the results were similar.

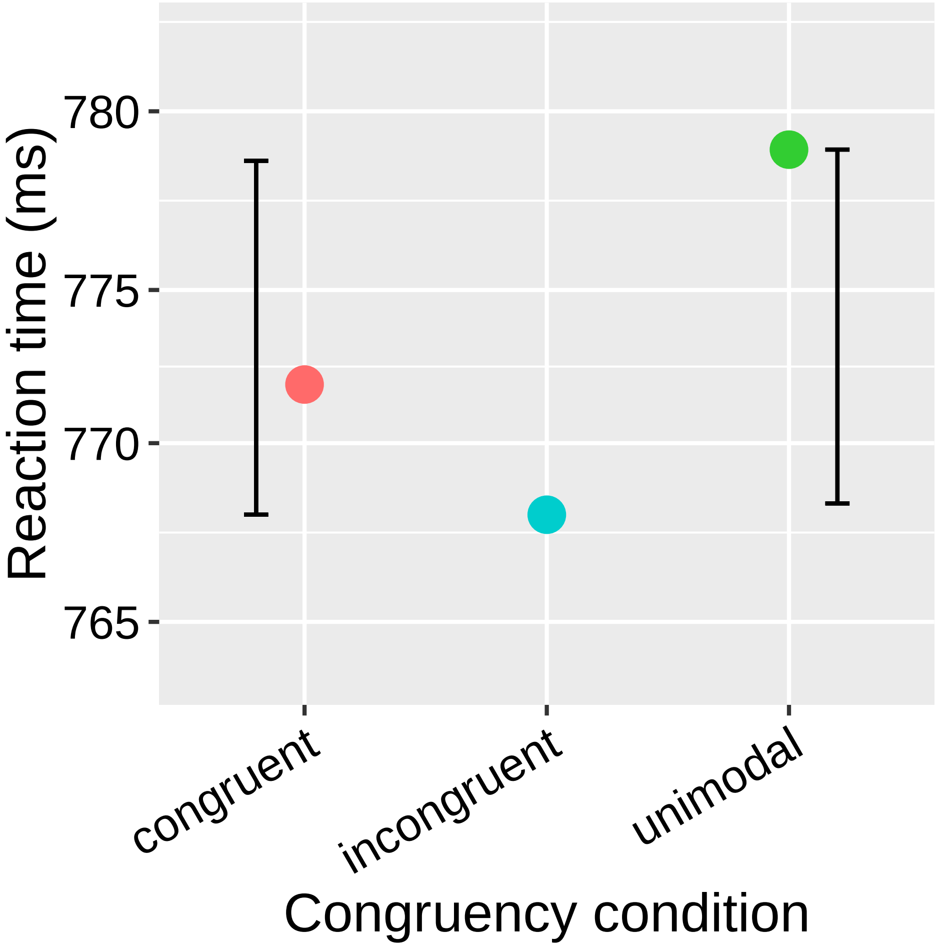


Figure 22. *Size Study 7*: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars). There was no difference in reaction time between any of the congruency conditions.

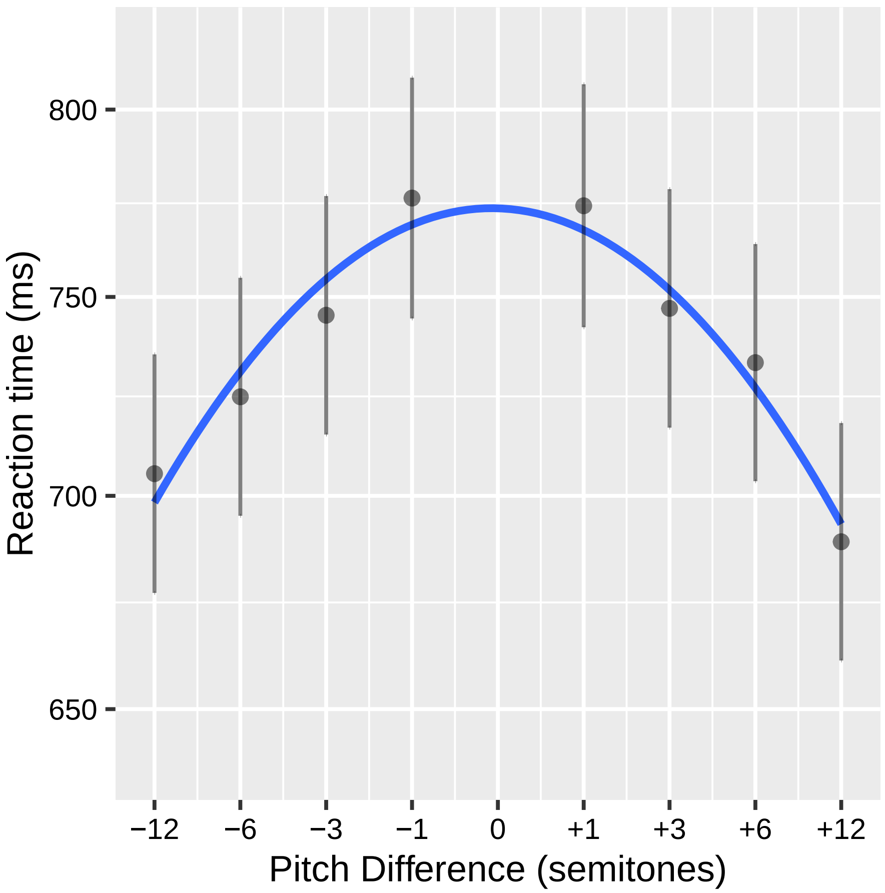


Figure 23. Size Study 7: Effect plot showing the main effect of pitch difference between the standard and comparison pitches (with 95% confidence intervals). The blue line is a quadratic line of best-fit showing the effect of pitch difference. Participants respond fastest to the largest pitch difference and slowest to the smallest pitch difference.

Pitch–Size Discussion

We found a robust `congruency x compatibility` interaction across the first four experiments (as predicted in Figure 3b), clearly showing the top-down influence of instructions on performance. Whatever dimensions are paired in the instructions for a given block, participants respond faster when those attributes are paired together. Further, the response speed for incongruent trials in the incompatible blocks did not differ from the response speed for congruent trials in the compatible blocks.

This interaction generalized across all pitch differences, timbres, and study designs used. The interaction was also not a function of learning the new pairing throughout the block of trials, as the interaction strength did not change from the first to last trial in any block. From this, we conclude that the pitch-size correspondence can be overridden with top-down control of attention to different task instructions. In other words, there is no evidence for an automatic congruency relationship between pitch and object size at a perceptual level.

The failure to replicate with a conceptual and direct replication of the single relevant dimension speeded classification task provides further evidence that the pitch-size correspondence is a fragile, decision-level effect rather than a robust perceptual effect. Although the unimodal condition reversed from being slower or faster than the bimodal conditions, the congruent and incongruent conditions never differed in reaction time.

Chapter 3: Pitch–Height Correspondence

This chapter extends the methodologies used for pitch–size to pitch–height, which is the most cited audiovisual correspondence (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010; Melara & O'Brien, 1987; Patching & Quinlan, 2002). Similar to pitch–size, pitch and height tend to be matched in a way that reflects our experience with the environment (i.e., a *statistical* correspondence): large objects are more likely to be lower in the environment (as they tend to be heavier). This correspondence is also considered *semantic/lexical* because many languages use the same words—'low' and 'high'—to describe stimuli that vary in pitch and stimuli that vary in visual height/elevation (Martino & Marks, 1999; Spence, 2011). The congruent endpoint mapping is clear in this case: low pitch/low elevation and high pitch/high elevation are perceptually congruent.

For this and subsequent chapters, I first completed a conceptual replication of the correspondence to establish the generalizability of the pitch–height congruency effect. Height Studies 1a and 1b adapted the procedure used by Gallace and Spence (2006) for the pitch–size correspondence to the pitch–height correspondence. This was done in order to change as little as possible from the previous methodology in order to provide a manipulation check of my program and the speeded classification paradigm. In Height Study 1a, height was the relevant dimension, and in Height Study 1b, pitch was the relevant dimension.

Following, I investigated the influence of task instructions on the pitch–height endpoint mapping. Specifically, Height Studies 2–4 included four types of trials: two with compatible instructions and two with incompatible instructions.

- **Compatible instructions:** blocks where "perceptually congruent" endpoints are paired. In *Block 1* listeners selected either the low shape or low pitch; in *Block 2* listeners selected either the high shape or high pitch.

- **Incompatible instructions:** blocks where "perceptually incongruent" endpoints are paired. In *Block 3* listeners selected either the low shape or high pitch; in *Block 4* listeners selected either the high shape or low pitch.

Additionally, I investigated the effect of lexical overlap. In Height Studies 2–3, participants were asked to choose the shape *below the midpoint* or *above the midpoint* and the *low* and *high* pitch. In Height Study 4, the words *low* and *high* were used to describe both the shapes' elevation and the pitches' heights.

The goal of this series of studies was to validate the methods used to investigate the pitch–size correspondence by showing that when similar methods are used for the pitch–height correspondence, the congruency effect robustly replicates. This would show that it is nothing about my methods that is suspect, but instead that my findings are the result of the fragility of the pitch–size correspondence. Therefore it was predicted that the pitch–height correspondence will show an effect of congruency in addition to the influence of instructions; in other words, the predicted results would look similar to Figure 3c (rather than Figure 3b). Further, it was predicted that there may be an even stronger congruency effect when the words to describe the dimensions overlap. Establishing which methods are capable of detecting an automatic pitch–height association will also direct future investigations with other less stable correspondences.

Height Study 1a: Height-Relevant Replication

Height Study 1 was a conceptual replication of Gallace and Spence (2006) using the pitch–height correspondence. Participants were randomly assigned to complete Study 1a or Study 1b first. In both versions, there was one relevant dimension for the entire experiment while the other dimension remained irrelevant for all of the trials. In Height Study 1a, height was the relevant dimension.

Method

Participants

The same seventy U.Va. undergraduate students participated in Height Study 1a and in exchange for credit in an introductory psychology course.

Stimuli

Visual shapes. A standard circle appeared horizontally and vertically centered on the screen. In addition, we used comparison circles that were ± 5%, ± 10%, ± 20%, and ± 40% vertically displaced from the screen's center (see Figure 4b). All circles were pixels 50 pixels (1.5 cm) in diameter.

Auditory pitches. We used the same three pitch pairings as Size Studies 1–2.

Design

Participants were randomly assigned to one of three pitch difference conditions [large ($N = 28$), octave ($N = 20$), or M3 ($N = 22$)] and one of two experiment orders [height-relevant first (Study 1a) or pitch-relevant first (Study 1b)].

For Study 1a, each of the eight different variable-height circles was presented 20 times in each of the three sound conditions (high frequency, low frequency, and no sound), giving rise to a total of 480 trials.

Procedure

At the start of a trial, a red fixation dot 10 pixels (0.3 cm in diameter) was presented in the center of the screen for 300 ms. This was followed by a 300 ms blank white screen. The first shape (the gray 'standard' circle) was presented at the screen's center for 300 ms, followed by a random dot mask that filled the entire screen for 500 ms. The second shape (the gray variable-height 'comparison' circle) was next presented for 300 ms at vertically displaced from the screen's center, accompanied by a high tone, low tone, or no tone. A second random dot mask then appeared that remained on the screen until a response was made or until 3,000 ms had elapsed, at which point a new trial began.

Participants were instructed to press one key if the comparison circle appeared to be lower than the standard, and a different key if the comparison circle appeared to be higher than the standard. They were told that on some trials a task-irrelevant sound would accompany the comparison circle, but they were asked to ignore it as much as possible.

Data management

Participants. We did not need to remove any participants from this study based on accuracy. The average overall accuracy was 89.8%.

Reaction Time. First, we eliminated incorrect trials (7% of responses removed) and trials where participants failed to respond in the allotted 3,000 ms (3.2% removed). We then eliminated responses that were faster than 50ms (0.09% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested a squared inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (10.5% removed).

Results and Discussion

Reaction Time

Table 13
Height Study 1a*: Comparison of four models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive: heightDiff factor`	`Intercept`	14	0.0	0.98	0.066	0.299
`Additive: heightDiff quadratic`	`Intercept`	9	8.2	0.02	0.064	0.296
`Non-additive: heightDiff quadratic`	`Intercept`	29	40.7	0.00	0.065	0.297
`Non-additive: heightDiff factor`	`Intercept`	74	466.3	0.00	0.068	0.301

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `height difference` (factored or quadratic), and `pitch difference` (large vs. octave vs. M3). We also included the subject-by-subject variation in the intercept as a random effect.

We compared four models: two models were additive, including just the three fixed effects (with height difference as a factored predictor or a continuous, quadratic predictor), and two models were non-additive, including the three fixed effects and all of their two and three-way interactions (with height difference a factored or quadratic predictor).

Table 13 shows that the additive model with factored height difference is clearly superior to the other models. The AICc difference (ΔAICc) between the best model and the next model is 8.2; "substantial evidence" in favor of the better model (Jeffreys, 1961).

Significant findings

We found a significant congruency effect collapsing across height differences (Figure 24). Participants were significantly *slower* to respond in the no sound condition (0.02, [0.012, 0.021]) and the incongruent condition (0.01, [0.003, 0.013]) than the congruent condition.

We also performed two manipulation checks. *First,* there was a significant effect of pitch difference, such that participants were significantly slower to respond to a `M3` pitch difference (0.09, [0.03, 0.14]). This is particularly interesting given that pitch was the irrelevant dimension.

Second, there was a significant effect of height difference between the shapes, such that participants were slowest to respond to a ±5% height difference and fastest to respond to a ±40% height difference, showing that participants were in fact completing the experiment as instructed.

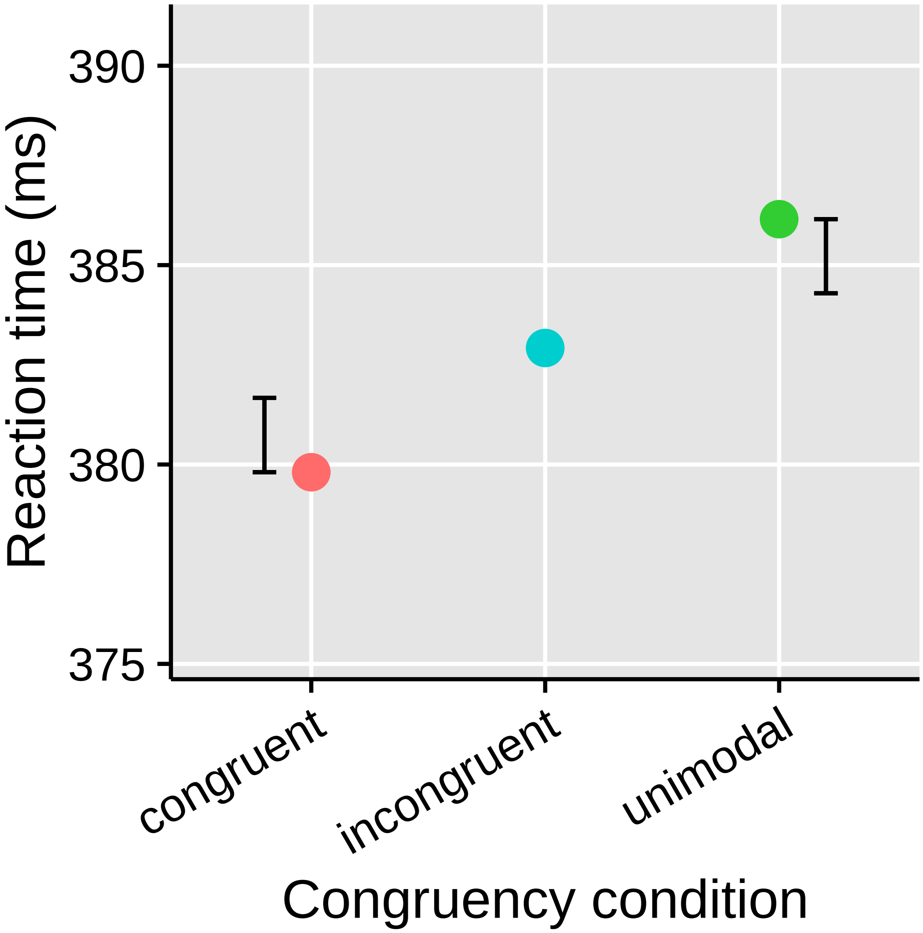


Figure 24. Height Study 1a: Effect plot showing the main effect of perceptual congruency (with upper and lower `LSD` bars). Participants showed a congruency advantage in that they responded significant faster in the congruent condition than the incongruent and unimodal conditions.

Errors

Our binomial GLMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `pitch difference` (M3 vs. octave vs. large), and `height difference` (factored or quadratic). We also included the subject-by-subject variation in the intercept as a random effect. Because of the low error probability overall, an additive model including factored height difference failed to converge, as did non-additive models with height difference as quadratic or factored. Thus we used an additive model with a quadratic height difference, a within-participant congruency factor, and a between-participant pitch difference factor.

There was a significant difference between the log-odds of correct responses for the congruent and unimodal conditions ($p \approx 0$; $z = -4.96$) but no difference between the congruent and incongruent conditions ($p = 0.35$; $z = 0.93$). The proportion of errors in the congruent, incongruent, and unimodal conditions (with a large pitch difference) was 0.077, 0.08, and 0.059, respectively.

Because there were significantly fewer errors in the unimodal condition, which was also the slowest (Figure 24), this may be evidence of a speed-accuracy tradeoff.

Conclusions

We replicated previous studies by finding a congruency effect between low pitches and low shapes (and high pitches/high shapes). Despite the successful replication, the difference between congruent and incongruent trials was small, especially at smaller pitch differences than used in previous studies, and the results may potentially be due to a speed-accuracy tradeoff.

Height Study 1b: Pitch-Relevant Replication

Height Study 1b repeats Study 1a with pitch rather than height as the relevant dimension.

Method

Participants

The same seventy U.Va. undergraduate students participated in Height Study 1a and in exchange for credit in an introductory psychology course.

Stimuli

Auditory and visual stimuli were the same as Height Study 1a.

Design

For Study 1b, the two pitches (high vs. low) were presented 20 times in each of the three height conditions (below center ×4 heights, above center ×4 heights, and no shape ×4 trials), giving rise to a total of 480 trials.

Procedure

The sequence of trials was the same as Height Study 1a with the following exceptions. After the fixation dot, there was only one pitch-shape pairing presented per trial. The pitch was played for 300 ms and was accompanied by a high circle, a low circle, or no circle. Participants were instructed to press one key if the pitch was *high* and a different key if the pitch was *low*.

Data management

Participants. We did not need to remove any participants from this study based on accuracy. The average overall accuracy was 91.2%.

Reaction Time. First, we eliminated incorrect trials (6.2% of responses removed) and trials where participants failed to respond in the allotted 3,000 ms (2.6% removed). We then eliminated responses that were faster than 50ms (0.26% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested a squared inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (6.8% removed).

Results and Discussion

Reaction Time

Table 14
Height Study 1b*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive`	`Intercept`	7	0.0	1.00	0.015	0.260
`Non-additive`	`Intercept`	11	22.8	0.00	0.015	0.261

Model comparison. Our LMMS included two fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal) and `pitch difference` (large vs. octave vs. M3).We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the two fixed effects to a non-additive model including the two fixed effects and their interaction. Table 14 shows that the additive model is clearly superior to the non-additive model. The AICc difference (ΔAICc) between the two models is 22.8; "strong evidence" in favor of the better model (Jeffreys, 1961).

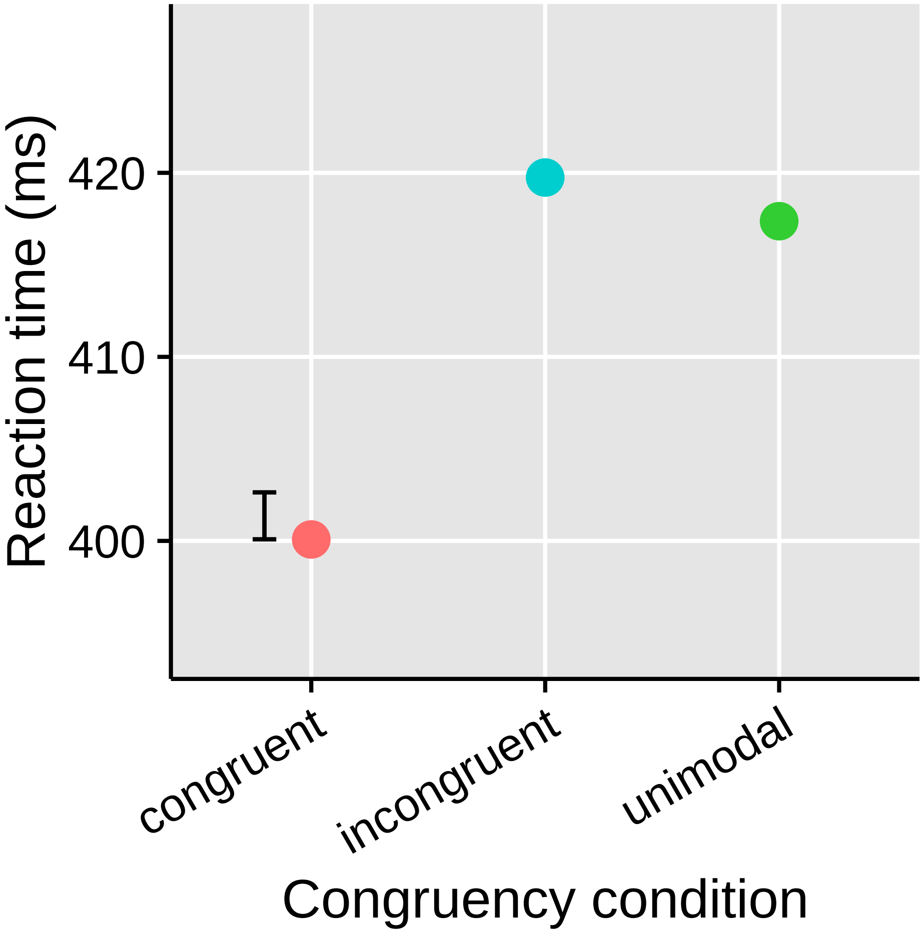


Figure 25. Height Study 1b: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars). Participants showed a congruency advantage in that they responded significant faster in the congruent condition than the incongruent and unimodal conditions.

Significant findings. We found a significant congruency effect collapsing across pitch differences (Figure 25). Participants were significantly *faster* to respond in the congruent condition than in the incongruent condition (0.05, [0.04, 0.05]) or the no sound condition (0.04, [0.03, 0.05]).

Errors

Our binomial GLMMs included two fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal) and `pitch difference` (M3 vs. octave vs. large). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model with just the two fixed effects to a non-additive model include the two fixed effects and their interaction. The non-additive model was clearly superior; the AICc difference (ΔAICc) between the two models was 9.7, which is "substantial evidence" in favor of the better model (Jeffreys, 1961).

There was a significant difference between the log-odds of correct responses for the congruent and incongruent conditions ($p \approx 0$; $z = 10.97$) and between the congruent and incongruent conditions ($p \approx 0$; $z = -4.06$). The proportion of errors in the congruent, incongruent, and unimodal conditions (with a large pitch difference) was 0.046, 0.105, and 0.03, respectively. Because the congruent condition was faster and had fewer errors than the incongruent condition (Figure 25), there is little evidence of a speed-accuracy tradeoff.

Conclusions

We again found a congruency effect between pitch and height. The difference between congruent and incongruent trials was stronger here with pitch as the relevant dimension than in Study 1a when height was the relevant dimension. This finding is in line with previous research that found auditory-relevant conditions were more strongly affected by congruency (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010).

Height Study 2: Top-down Influence [Between-participants]

We successfully replicated the pitch–height correspondence in Height Studies 1a and 1b. Next, we sought to investigate the top-down influence of task instructions on performance. For Height Study 2, we used a between-subjects design where each participants received one set of pairing instructions for the entirety of the experiment.

Method

Participants

Fifty-one U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory stimuli were the same as Height Studies 1a and 1b.

We used three pairs of visual stimuli: circles were ± 20%, ± 40% and ± 80% vertically displaced from the screen's center (see Figure 4b).

Design

We used a between-participant design in which each participant completed ten blocks of 36 trials. Participants were randomly assigned to one of four sets of experiment instructions: circle below midpoint/lower pitch, below midpoint/higher pitcher, above midpoint/higher pitch, or above midpoint/lower pitch.

Procedure

At the start of each block (of 36 trials), the instructions were presented on the screen. Participants pressed the `SPACEBAR` to proceed with the trial (on every trial). The first shape appeared for 300ms; it was accompanied by a 300ms tone. This was followed by a 500ms blank screen. Then the second shape appeared for 300ms; it was also accompanied by a 300ms tone. After both pitch/height pairings, either an 's' for shape or 'p' for pitch appeared on the screen to cue the participant to which modality to respond. Participants indicated whether the first or second stimulus met the instruction criteria (e.g., for below/lower instructions, participants would respond to the *lower* pitch if a 'p' appeared or the shape *below* the midpoint if an 's' appeared). The cue remained on the screen until the participant responded, at which point the instruction screen reappeared to begin the next trial.

Data management

Participants. We removed two participants with less than 60% accuracy on one or both stimulus dimensions. The average overall accuracy of the remaining 49 participants was 90.3%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (4% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the four fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 6.8% of responses.

Results and Discussion

Reaction Time

Table 15
Height Study 2*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	34	0.0	1.00	0.034	0.298
`Additive`	`Intercept`	8	184.8	0.00	0.014	0.272

Model comparison. Our LMMS here included five fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch or responding to shape), `height difference` (continuous), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the five fixed effects to a non-additive model including the five fixed effects and all of their two, three, four, and five-way interactions. Table 15 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 184.8, which is "decisive evidence" in favor of the better model (Jeffreys, 1961).

Significant findings. Figure 26 shows the predictors of logRT and their 95% confidence intervals. The `congruency x compatibility` interaction is the most important. Figure 27 shows the top-down influence of instructions on performance. There is a clear *congruency* advantage on compatible trials (0.19, [0.16, 0.23]), but there is also an *incongruency* advantage on incompatible trials (−0.12, [−0.16, −0.09]). However, the figure also shows that participants are significantly slower to respond to the incongruent trials in the incompatible condition than congruent trials in the compatible condition. This means that it takes longer overall to pair together the perceptually incongruent endpoints, showing some evidence of perceptual congruency impacting performance.

Another noteworthy finding comes from a three-way interaction model with `congruency` × `compatibility` and `trial number` as fixed effects.The three way-interaction is not significant (−1.66, [−5.61, 2.29]), which means the interaction strength does not change across trials.

Errors

Model comparison. Our binomial GLMMS included five categorical fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch vs. responding to shape), `height difference` (continuous) and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

Because of the low error probability overall, a non-additive model including the five fixed effects and all of their interactions failed to converge. Instead, we compared a non-additive model including the five fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the five fixed effects. The AICc difference (ΔAICc) between the two models is 27, which is "strong evidence" in favor of the non-additive model (Jeffreys, 1961).

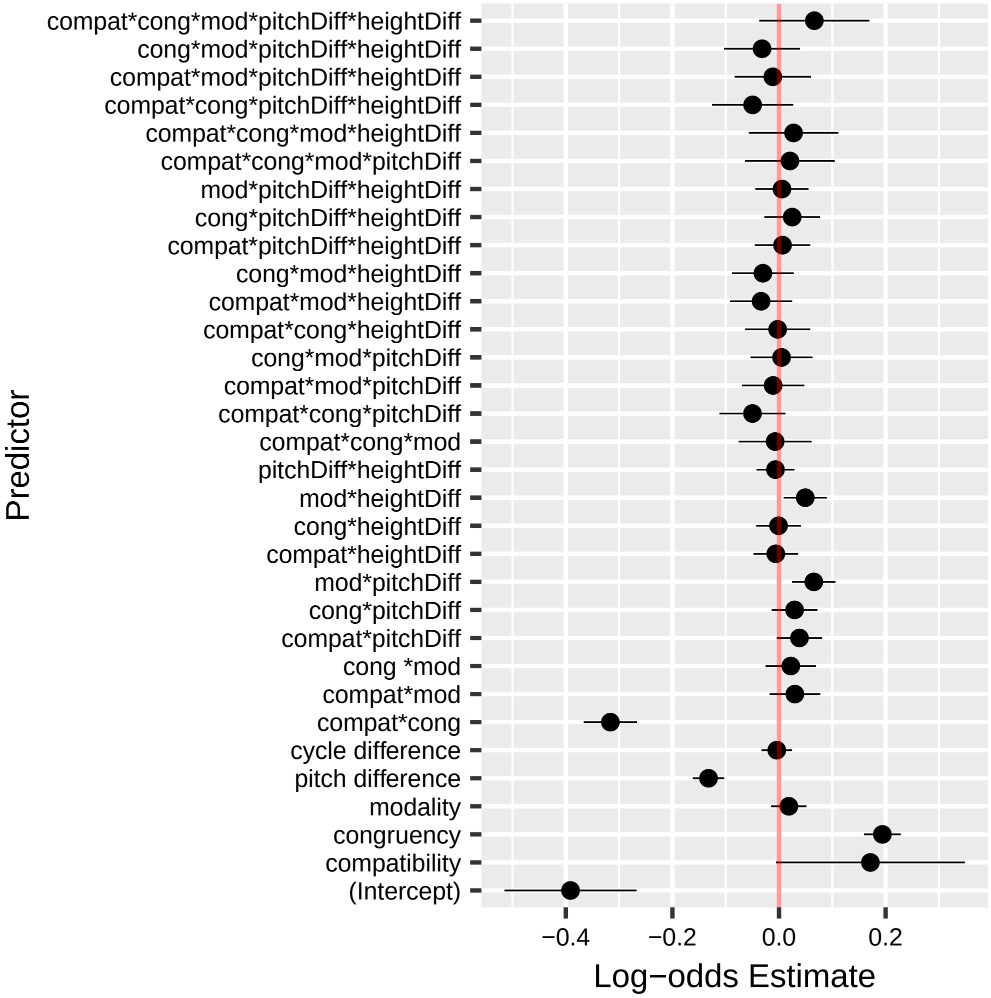


Figure 26. *Height Study 2*: Coefficient plot for the five fixed effect predictors and their interactions (with 95% confidence intervals) Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference; heightDiff = height difference.

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible trials is $p = 0.28$ ($z = 1.07$), meaning listeners make approximately the same number of errors on congruent trials regardless of compatibility condition. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 8.86$), meaning listeners make a few more errors on incongruent trials. These effects are converted to proportions and shown in Table 16. The error rates follow the same pattern as the reaction times (shown in Figure 27), meaning there is no evidence of a speed-accuracy tradeoff.

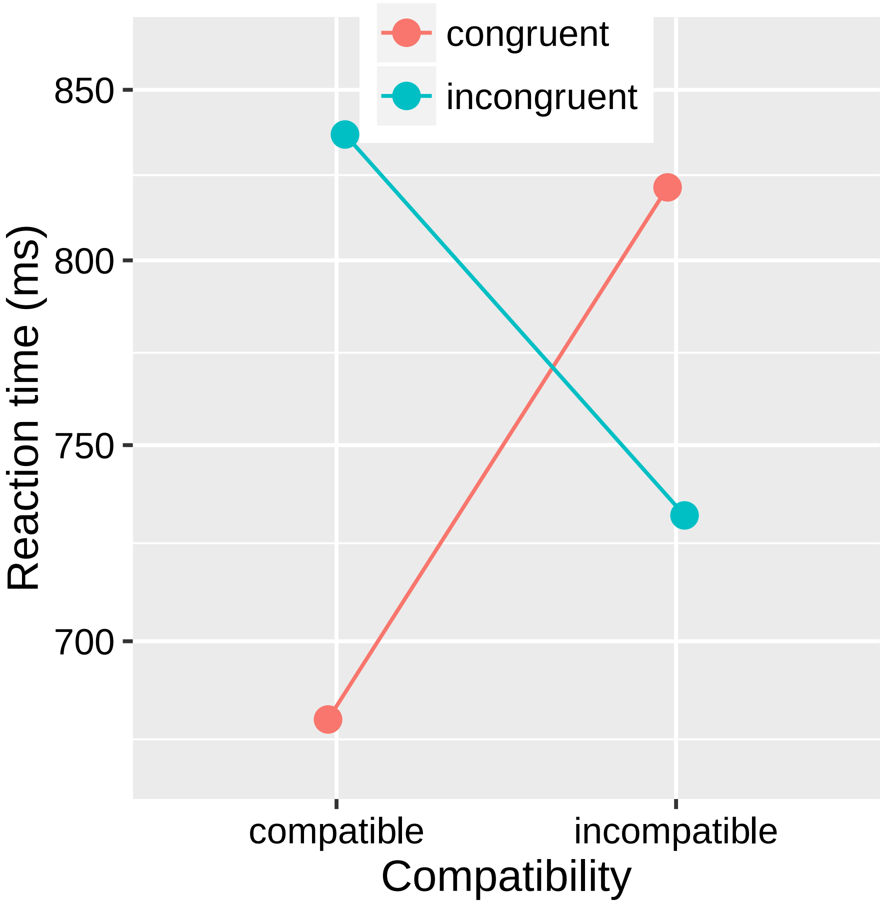


Figure 27. *Height Study 2*: Effect plot showing the interaction between perceptual congruency and instructions compatibility. (LSD bars not shown because they are not easily interpreted in a between-participant design. See text for significance.) Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

Table 16
Height Study 2*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.08	0.10
Incongruent trials	0.14	0.11

Conclusions

Despite replicating the pitch–height correspondence with one relevant dimension in Height Studies 1a and 1b, here the results clearly that most of the effect is due to the top-down influence of instructions rather than congruency using a between-participants design.

Height Study 3: Top-down Influence [Within-participants]

We found strong evidence for top-down influence in Height Study 2 using a between-participants design. In Height Study 3, we sought to replicate this effect using a within-subjects design in which each participant completed a block of each of the four types of instructions (described above).

Method

Participants

Twenty-six U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory stimuli and visual stimuli were the same as Height Study 2.

Design

We used a within-participant design in which each participant completed four blocks of 72 trials. Participants completed one block using each of four sets of experiment instructions in a random order: circle below midpoint/lower pitch, below midpoint/higher pitcher, above midpoint/higher pitch, and above midpoint/lower pitch.

Procedure

The procedure was the same as Height Study 2 except the instructions were presented on every trial rather than only on the first trial of the block.

Data management

Participants. We removed four participants with less than 60% accuracy on one or both stimulus dimensions. The average overall accuracy of the remaining 22 participants was 86.8%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (4.7% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the five fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 7.2% of responses.

Results and Discussion

Reaction Time

Table 17
Height Study 3*: Comparison of two models, ordered by ΔAICc relative to the model with the lowest (best) AICc. K is the number of parameters in the model.*

fixed effects	random effect	K	Δ AICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	34	0.0	1.00	0.043	0.315
`Additive`	`Intercept`	8	95.1	0.00	0.022	0.294

Model comparison. We compared an additive model including just the five fixed effects (see Height Study 2) to a non-additive model including the five fixed effects and all of their two, three, four-, and five-way interactions. Table 17 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 95.1, which is “very strong evidence” in favor of the better model (Jeffreys, 1961).

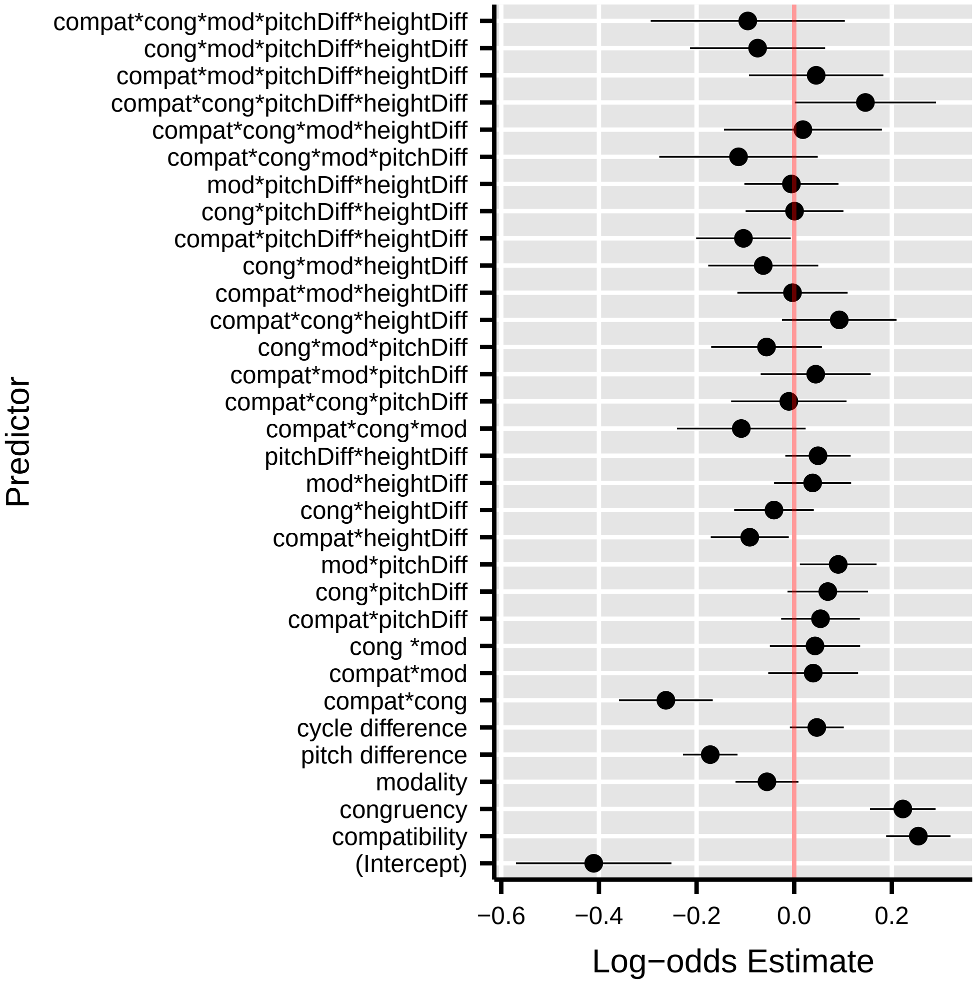


Figure 28. Height Study 3: Coefficient plot for the five fixed effect predictors and their interactions (with 95% confidence intervals). Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference; heightDiff = height difference.

Significant findings. Figure 28 shows the predictors of logRT and their 95% confidence intervals. The `congruency x compatibility` interaction is the most important. Figure 29 shows a different pattern than previous studies; here, congruency and instructions are both important.

Although there is a clear *congruency* advantage in the compatible condition (0.21, [0.14, 0.28]), there is not a significant *incongruency* advantage for the incompatible condition (−0.05, [−0.12, 0.01]). Further, the figure shows that participants are significantly slower to respond to the in-

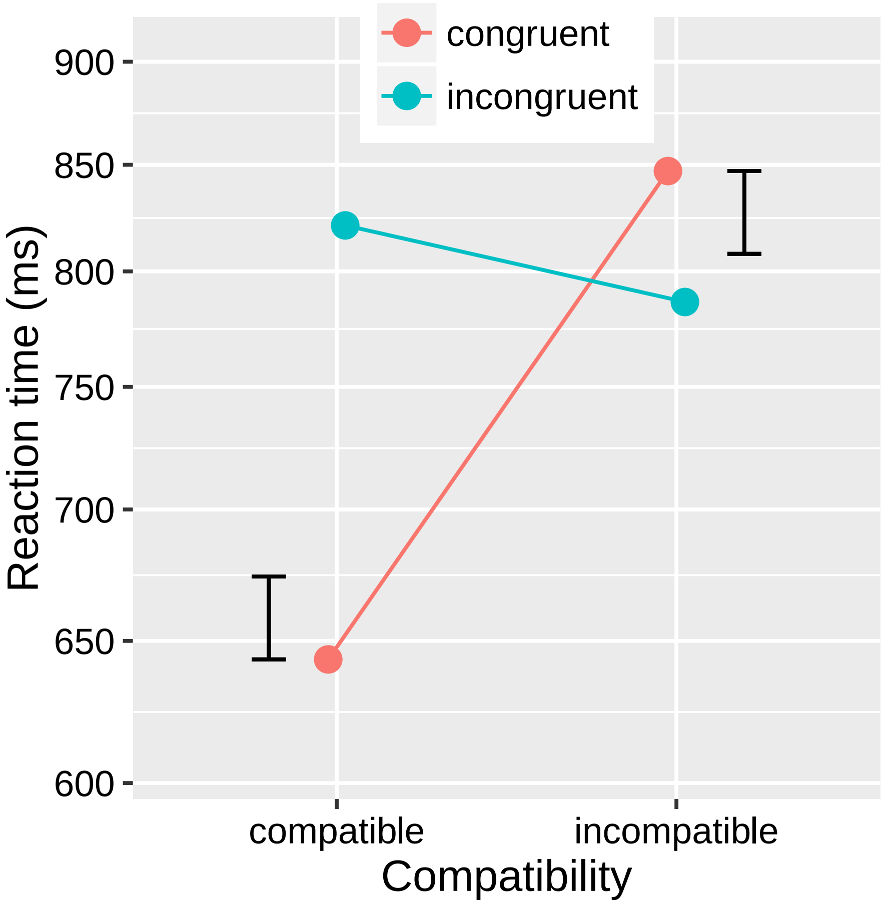


Figure 29. Height Study 3: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower `LSD` bars).

congruent trials in the incompatible condition than congruent trials in the compatible condition, meaning that overall it takes longer to pair together the perceptually incongruent endpoints.

Additionally, the three way-interaction between `congruency` × `compatibility` and `trial number` is not significant (1.61, [−3.06, 6.27]), which means the interaction strength does not change across trials.

Errors

Model comparison. Because of the low error probability overall, a non-additive model including the five fixed effects (see Height Study 2) and all of their interactions failed to converge. Instead, we compared a non-additive model including the five fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the five fixed effects. The AICc difference (ΔAICc) between the two models is 47.1, which is “strong evidence” in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible trials is $p \approx 0$ ($z = 6.36$), meaning listeners make more errors on the incompatible blocks. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 7.63$), meaning listeners make more errors on incongruent trials. These effects are converted to proportions and shown in Table 18. The error rates follow the same pattern as the reaction times (shown in Figure 29), meaning there is no evidence of a speed-accuracy tradeoff.

Table 18
Height Study 3*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.08	0.15
Incongruent trials	0.17	0.13

Conclusions

Here we found some evidence for a congruency effect in addition to a top-down effect of instructions (akin to Figure 3c) using a within-participant design. There was not a reversal of the congruency effect on incompatible trials, which shows that participants had a harder time grouping shapes below the midpoint with high pitches than with low pitches. The reaction times were also slower overall in the incompatible condition because incongruent trials were not as fast as congruent trials in the compatible condition. The results here are at odds with the results from Size Studies 1–4, which found a similar top-down influence using a between- and within-participant design. Again, this lends support to the fact that our modified speeded classification method can detect a correspondence if there is a robust effect to be discovered.

Height Study 4: Lexical Overlap

Height Study 4 investigated the effect of lexical overlap. We used the words *low* and *high* to describe both the visual elevations and auditory pitches.

Method

Participants

Twenty-four U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory stimuli and visual stimuli were the same as Height Study 2–3.

Design

Participants completed one block using each of four sets of experiment instructions in a random order: lower circle/lower pitch, lower circle/higher pitch, higher circle/higher pitch, and higher circle/lower pitch.

Procedure

The procedure was the same as Height Study 3.

Data management

Participants. We removed three participants with less than 60% accuracy on one or both stimulus dimensions. The average overall accuracy of the remaining 21 participants was 92.1%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (1.9% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the four fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 5.8% of responses.

Results and Discussion

Reaction Time

Model comparison. We compared an additive model including just the five fixed effects (see Height Study 2) to a non-additive model including the five fixed effects and all of their two, three, four, and five-way interactions. Table 19 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 18.8, which is “strong evidence” in favor of the better model (Jeffreys, 1961).

Table 19

Height Study 4*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	34	0.0	1.00	0.074	0.246
`Additive`	`Intercept`	8	18.8	0.00	0.044	0.213

Significant findings. Figure 30 shows the predictors of logRT and their 95% confidence intervals. The two significant two way interactions—`congruency` × `compatibility` and `modality` × `congruency`—are encompassed in the three-way interaction `congruency` × `compatibility` × `modality`.

Figure 31 shows that congruency and instructions are both important (as in Height Study 3). Although there is a clear *congruency* advantage on compatible trials (0.21, [0.15, 0.27]), there is not a significant *incongruency* advantage on incompatible trials (−0.06, [−0.11, 0]).

Figure 32 shows that the `congruency` × `compatibility` interaction is slightly different depending on response modality. When responding to *pitch*, instead of gaining speed when responding to incongruent trials on incompatible blocks, participants were significantly *slower* overall on incompatible blocks. This means that seeing `low` objects made it hard to choose the `high` pitch (and vice versa) regardless of what instructions were given. When responding to shape *height*, the pitch interference was not as extreme. Though not equal to the congruent-compatible condition, participants were faster to respond to incongruent trials in the incompatible condition than in the compatible condition. The slower speed overall on the incompatible case still shows that participants have difficulty pairing together the perceptually incongruent endpoints.

Additionally, the three way-interaction between `congruency` × `compatibility` and `trial number` is not significant (−0.3, [−4.64, 4.04]), which rules out a learning explanation.

Errors

Model comparison. Because of the low error probability overall, a non-additive model including the five fixed effects (see Height Study 2) and all of their interactions failed to converge. Instead, we compared a non-additive model including the five fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the five fixed effects. The AICc difference (ΔAICc) between the two models is 47.1, which is "strong evidence" in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible trials is $p \approx 0$ ($z = 6.36$), meaning listeners make more errors on the incompatible blocks. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 7.63$), meaning listeners make more errors on incongruent trials. These effects are converted to proportions and shown in Table 20. The error rates follow the same pattern as the reaction times (shown in Figure 31), meaning there is no evidence of a speed-accuracy tradeoff.

Table 20

Height Study 4*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.08	0.15
Incongruent trials	0.17	0.13

Conclusions

We again found evidence for a congruency effect in addition to a top-down effect of instructions. There was not a reversal of the congruency effect on incompatible trials, which shows that participants had a harder time grouping low shapes with higher pitches than lower pitches. Because the magnitude of the congruency effect vs. incongruency effect was similar across Height Study 3 and Height Study 4, lexical overlap did not seem to influence the results. Nonetheless, we cannot rule out that participants in Height Study 3 were thinking of the objects as ‘low’ and ‘high’ even though we gave them alternative descriptors.

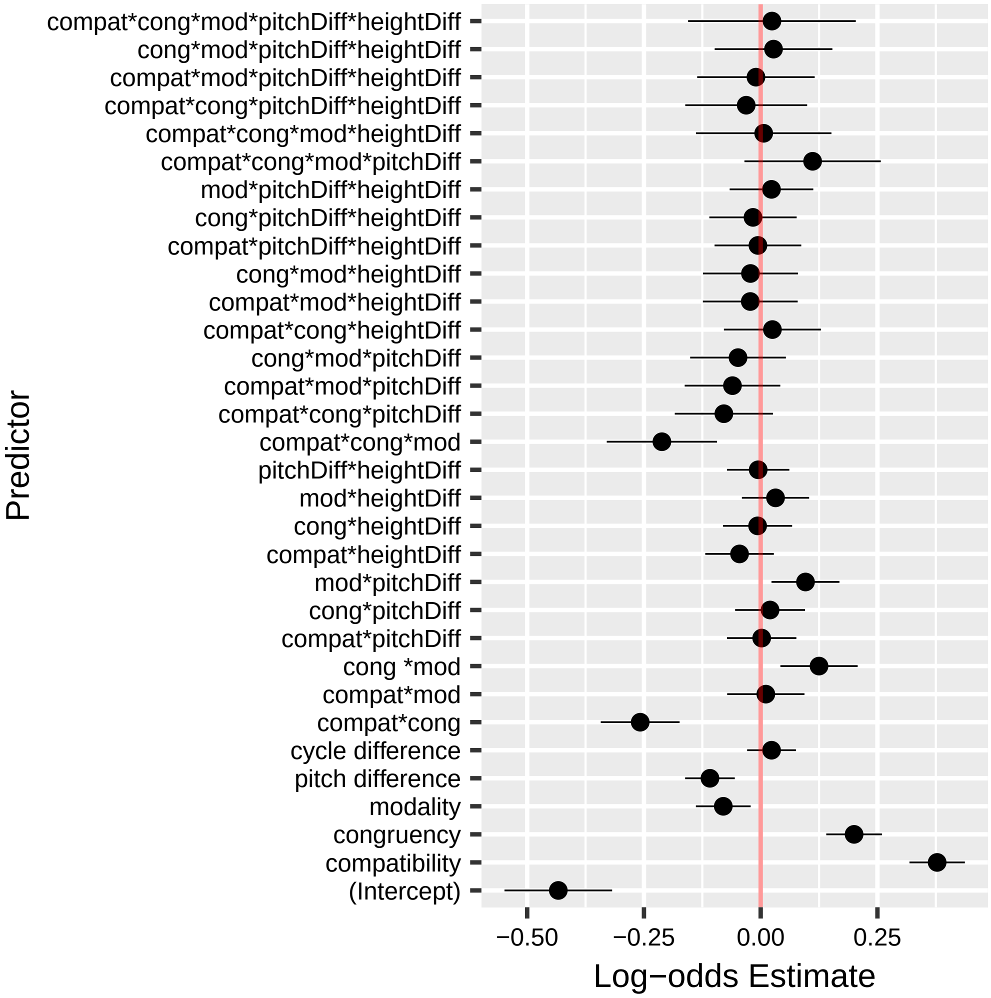


Figure 30. Height Study 4: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals) Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference; heightDiff = height difference.

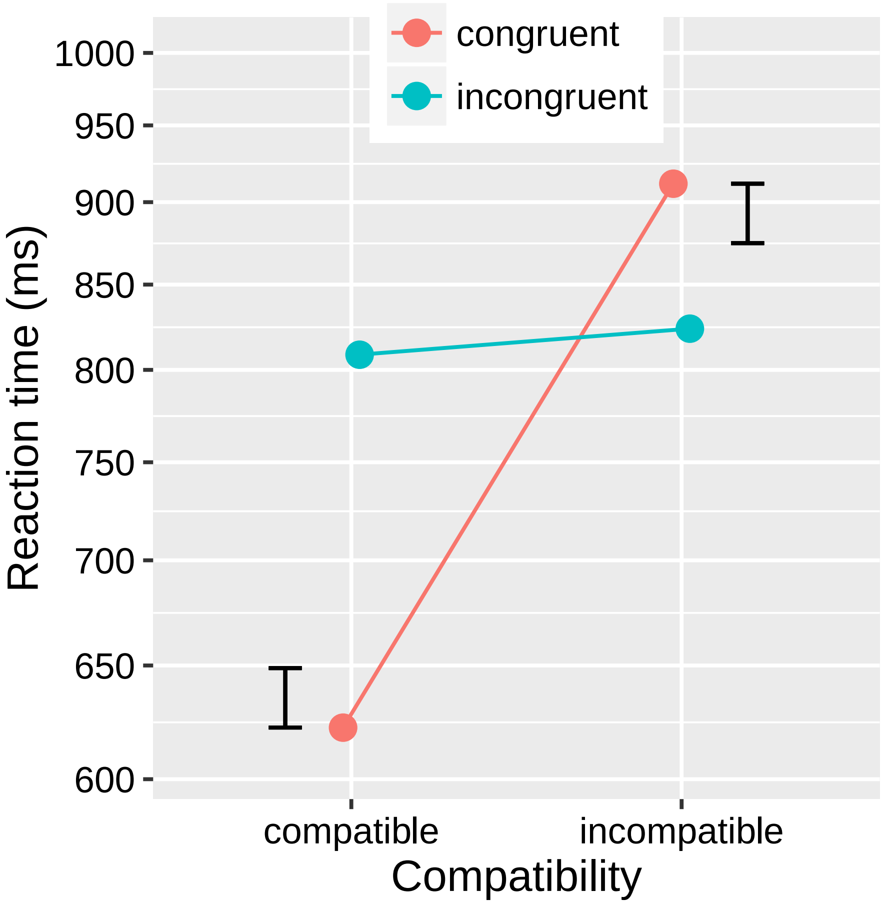


Figure 31. *Height Study 4*: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower `LSD` bars).

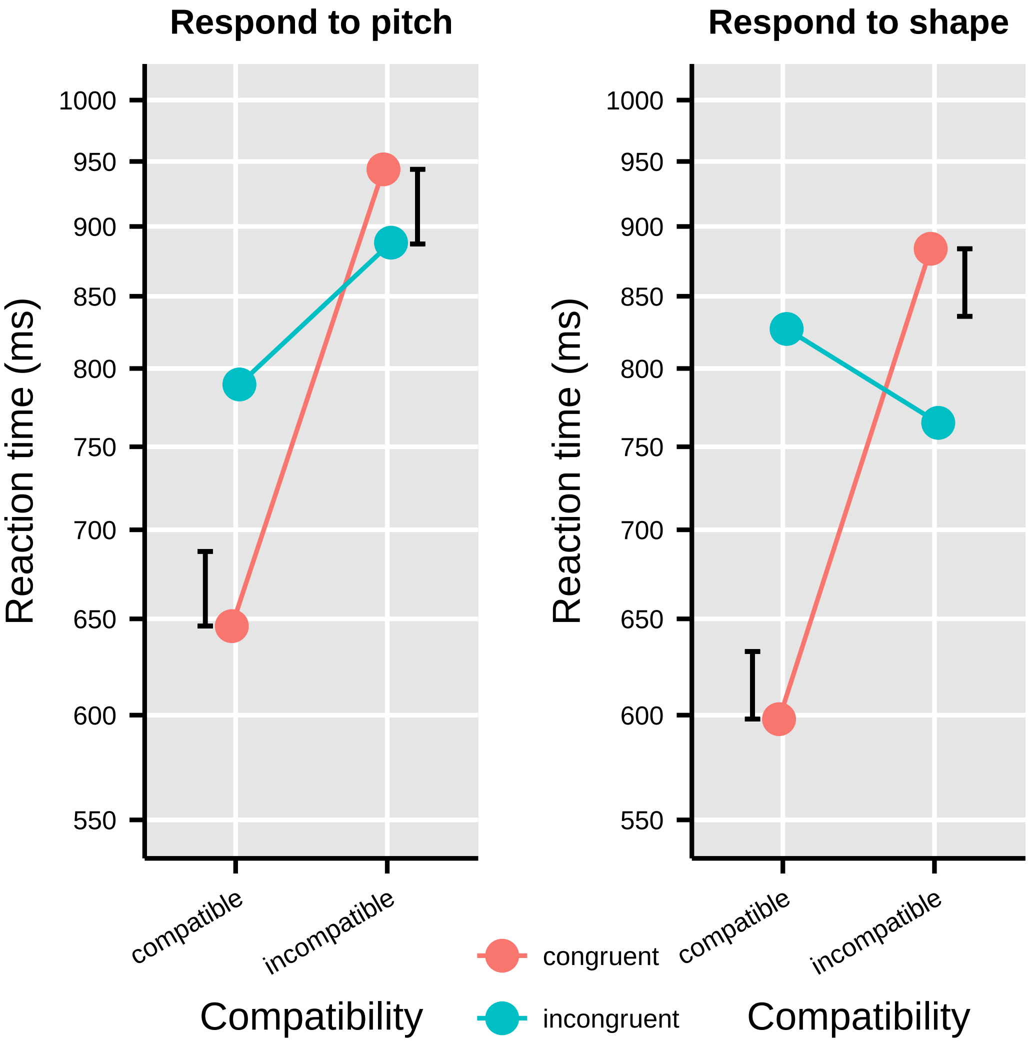


Figure 32. *Height Study 4*: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower LSD bars) by response modality (panels).

Pitch–Height Discussion

Height Studies 1a and 1b provided evidence that there is an automatic association between pitch and height. The results here diverge from Size Studies 5–7 which failed to find evidence for an automatic pitch–size correspondence. This lends support to the fact that our methods are sensitive enough to detect a correspondence if there is a robust correspondence to be discovered.

Results of Height Studies 3–4 also diverged from findings with the pitch–size correspondence. We found that although there was still a `congruency x compatibility` interaction, perceptual congruency and task instructions both influenced response speed (as in Figure 3c). In both experiments, there was not a reversal of the congruency effect on incompatible trials, which shows that participants had a harder time grouping shapes perceptually congruent endpoints, even when called to in the instructions. Using matching words (high and low) for both modalities did not seem to impact the results, though we cannot rule out the possibility that participants used those descriptors internally regardless of the words we used in our instructions.

Several height studies also provided support for previous research showing that auditory-relevant conditions are more strongly affected by congruency (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010). The reaction time difference between congruent and incongruent trials was larger in Study 1b when pitch was the relevant dimension than in Study 1a when height was the relevant dimension. In Study 4 the influence of congruency was more evident on pitch-relevant trials in that participants were significantly slower overall on the incompatible case, meaning seeing low objects made it hard to choose the high pitch (and vice versa) regardless of what instructions were given.

From these results, we conclude that although there may be evidence for an automatic congruency relationship between pitch and object height at a perceptual level, the correspondence can still be partially overridden with top-down control of attention to different task instructions.

Chapter 4: Pitch–Spatial Frequency Correspondence

Similar to pitch and height, auditory pitch and visual spatial frequency is a *semantic/lexical* correspondence (Evans & Treisman, 2010; Spence, 2011) because the words *low* and *high* are used to describe stimuli that vary in pitch and spatial frequency (i.e., the repetition rate of an object's sinusoidal components per unit of distance, or the wideness/narrowness of the object's striping pattern; see Figure 4c). The congruent endpoint mapping in this case is: low pitch/low spatial frequency and high pitch/high spatial frequency.

As in Chapter 3, here I first completed a conceptual replication of the correspondence to establish the generalizability of the pitch–spatial frequency congruency effect. In SF Study 1a, spatial frequency was the relevant dimension, and in SF Study 1b, pitch was the relevant dimension.

Following, I investigated the influence of task instructions on the pitch–spatial frequency endpoint mapping. Specifically, SF Studies 2–3 included four types of trials: two with compatible instructions and two with incompatible instructions.

- **Compatible instructions:** blocks where "perceptually congruent" endpoints are paired. In *Block 1* listeners selected either the shape with low spatial frequency or the low pitch; in *Block 2* listeners selected either the shape with high spatial frequency or the high pitch.

- **Incompatible instructions:** blocks where "perceptually incongruent" endpoints are paired. In *Block 3* listeners selected either the shape with low spatial frequency or the high pitch; in *Block 4* listeners selected either the shape with high spatial frequency or the low pitch.

Additionally, I again investigated the effect of lexical overlap. In SF Study 2, participants were asked to choose the shape with *wide stripes* or *narrow stripes* and the *low* and *high* pitch. In SF Study 3, the words *low* and *high* were used to describe both the shapes' spatial frequency and the pitches' frequency.

Having established a clear difference in the results of the pitch–size and pitch–height correspondences, the goal of Chapters 4–6 was to extend the results to other audiovisual correspondences to see whether they follow the pattern established by size or height. Doing so will allow us to uncover potential consistent properties of cross-modal correspondences that influence their stability.

SF Study 1a: Spatial Frequency-Relevant Replication

Participants were randomly assigned to complete SF Study 1a or Study 1b first. In both versions, there was one relevant dimension for the entire experiment while the other dimension remained irrelevant. In SF Study 1a, spatial frequency was the relevant dimension.

Method

Participants

The same thirty-two U.Va. undergraduate students participated in SF Study 1a and SF Study 1b in exchange for credit in an introductory psychology course.

Stimuli

Visual shapes. Shapes were all presented on a gray background. The circles were 200 pixels in diameter and included high-contrast black and white sinosoidal gratings oriented 45 to the left (Figure 4c). We used three different (within-participant) spatial frequency cycle[11] pairings: a difference of 14 (6 vs. 20), 10 (8 vs. 18), and 6 (10 vs. 16).

Auditory pitches. We used the same three pitch pairings as Size Studies 1–7.

Design

Participants were randomly assigned to one of two experiment orders [spatial frequency-relevant first (Study 1a) or pitch-relevant first (Study 1b)].

For Study 1a, each of the three spatial frequency differences (14, 10, 6) was presented eight times in each of the four pitch difference conditions (large, octave, M3, and no sound), giving rise to a total of 96 trials per block. Participants completed two blocks (in a random order): (a) Participants were asked to select whether the first or second shape had *wider* stripes; (b) participants were asked to select whether the first or second shape had *narrower* stripes.

Procedure

The procedure followed the sequence of events presented in Figure 2 with one exception: instead of waiting until the end of the trial to discover whether they should respond to the pitch or the shape, participants always responded to the width of the shapes' stripes and pitch was always the irrelevant dimension. After the second stimulus appeared for 300 ms, it was replaced by a blank screen until the participant responded. Participants were told they could respond as soon as the second shape appeared. They were also told that task-irrelevant sounds would only appear on some trials.

Data management

Participants. We did not need to remove any participants. The average overall accuracy of the 32 participants was 99.3%.

Reaction Time. Using the correct responses only, we eliminated trials where responses that were faster than 50ms (0.02% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested an inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (7.5% removed).

Results and Discussion

Reaction Time

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `spatial frequency difference` (continuous), and `task` (respond to narrower stripes vs. wider stripes). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the three fixed effects to a non-additive model including the three fixed effects and all of their two and three-way interactions. Table 21 shows

[11] One cycle is equal to gray-black-gray-white-gray; i.e., one black and one white lobe.

that the non-additive model is clearly superior to the additive model. The AICc difference (ΔAICc) between the two models is 108.5; "decisive evidence" in favor of the better model (Jeffreys, 1961).

Table 21

Spatial Frequency Study 1a*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	14	0.0	1.00	0.035	0.358
`Additive`	`Intercept`	7	108.5	0.00	0.016	0.339

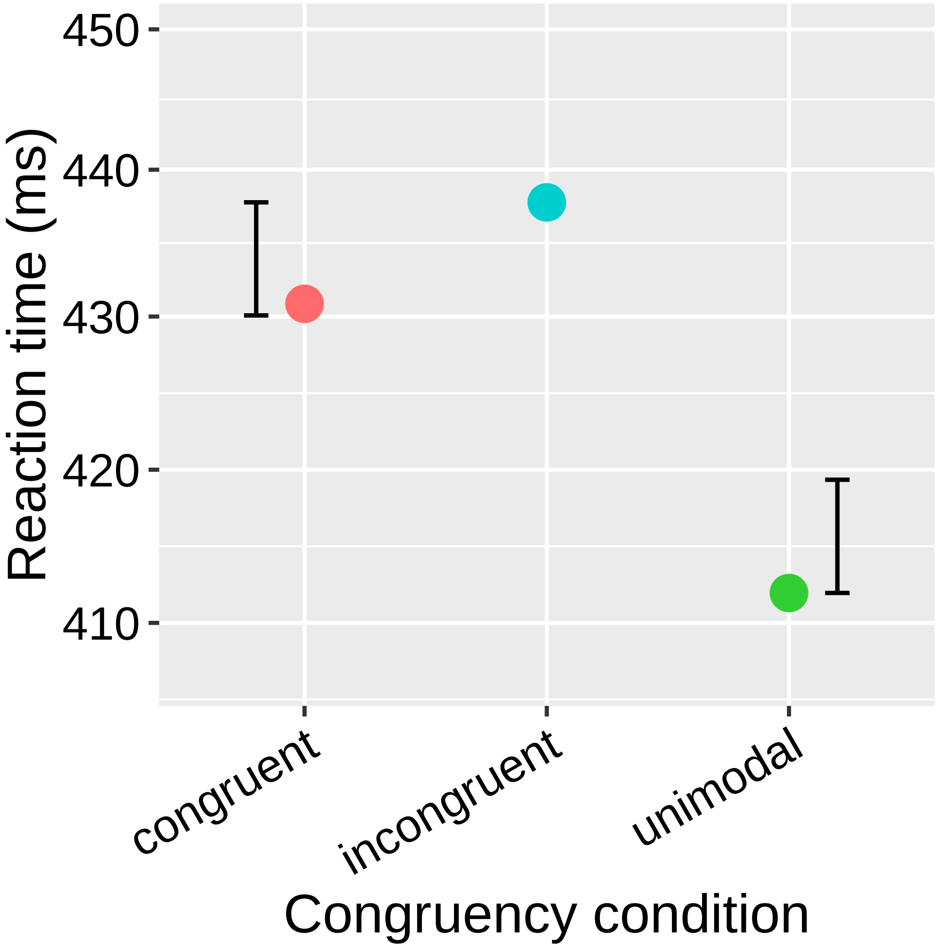


Figure 33. SF Study 1a: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars) when *spatial frequency* is the relevant dimension. Participants were significantly *slower* in any sound condition (congruent or incongruent) compared to the no sound condition.

Significant findings. We found no evidence of a congruency effect collapsing across spatial frequency differences and task (Figure 33). Instead we found that participants were significantly

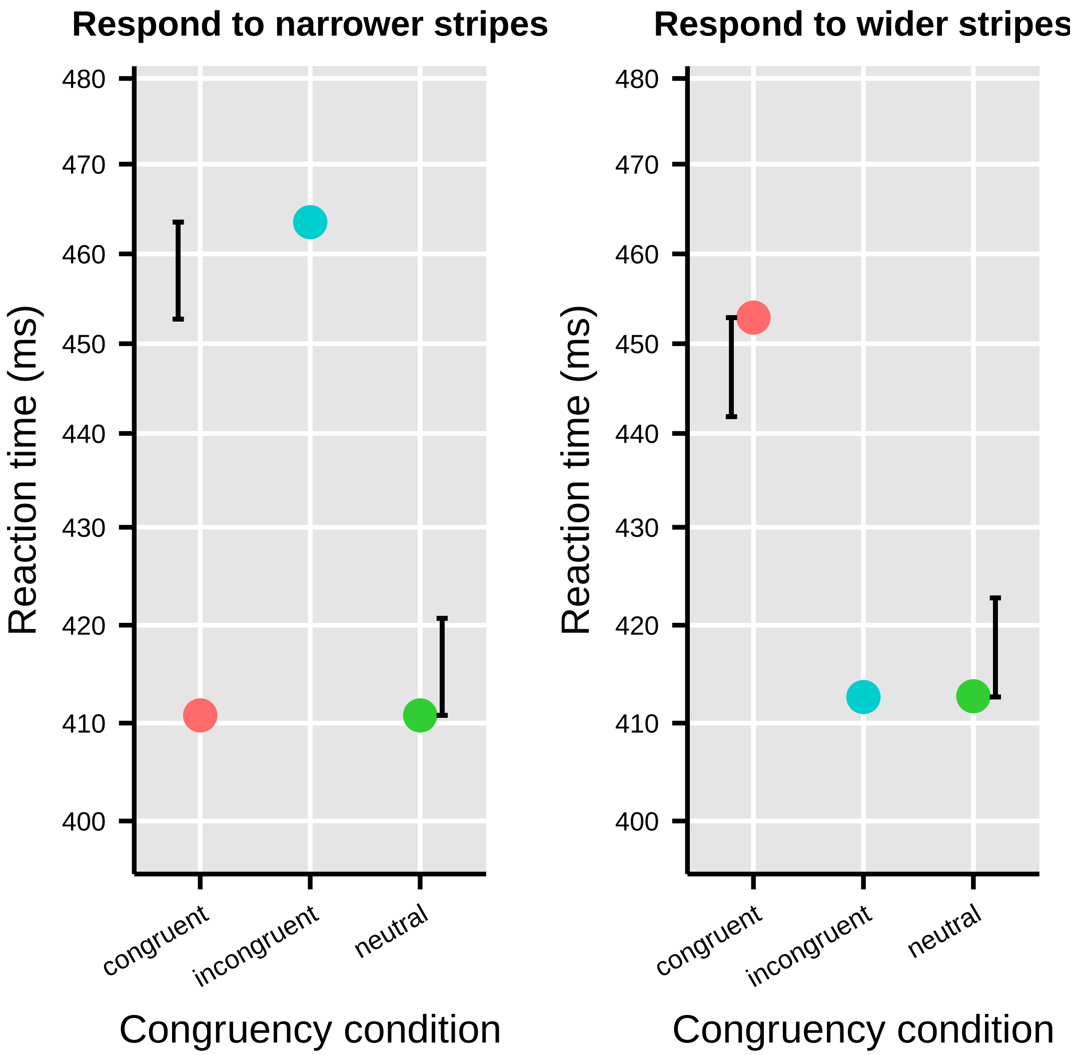


Figure 34. *SF Study 1a*: Effect plot (with upper and lower LSD bar) showing the main effect of congruency (x-axis) by task (panels) when *spatial frequency* is the relevant dimension.

faster to respond in the no sound condition (−0.04, [−0.06, −0.03]) than to either the congruent or incongruent sound conditions, which did not differ from each other (0.02, [0, 0.03]).

When we look separately at blocks requiring participants to select the *narrower* and *wider* stripes (Figure 34), we see that although the unimodal condition remains the fastest, the congruency advantage reverses between tasks. It is an open question why high pitches would be more quickly paired with whatever shape is included in the instructions (i.e., high pitch/narrower [congruent] is faster when choosing the narrower stripes whereas high pitch/wider [incongruent] is faster when choosing the wider stripes).

We also found a significant effect of spatial frequency cycle difference between the shapes (−0.04, [−0.05, −0.03]), such that participants were slowest to respond to a 6-cycle difference and fastest to respond to a 14-cycle difference, showing that participants were in fact completing the experiment as instructed.

Errors

Our poisson GLMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `spatial frequency difference` (continuous), and `task` (respond to narrower stripes vs. wider stripes). We also included the subject-by-subject variation in the intercept as a random effect. We compared an additive model including just the three fixed effects to a non-additive model including the fixed effects and their two- and three-way interactions. The AICc difference (ΔAICc) between the two models is 13.9, which is "strong evidence" in favor of the non-additive model (Jeffreys, 1961).

There was not a significant difference between the logarithm of correct responses for the congruent and incongruent conditions ($p = 0.91$; $z = -0.12$) or for the congruent and unimodal conditions ($p = 0.88$; $z = -0.15$). The proportion of errors in the congruent, incongruent, and unimodal conditions (responding to narrower stripes with the mean spatial frequency difference) was 0.004, 0.007, and 0.009, respectively. Because there is no difference in errors across conditions, we can rule out the possibility of a speed-accuracy tradeoff.

Conclusions

It is unclear why the congruency advantage reverses when the task changes from identifying the narrower stripes to identifying the wider stripes. Nonetheless, we were unable to replicate previous results finding an automatic association between pitch and spatial frequency when the visual dimension was relevant.

SF Study 1b: Pitch-Relevant Replication

SF Study 1b repeated Study 1a with pitch rather than spatial frequency as the relevant dimension.

Method

Participants

The same thirty-two U.Va. undergraduate students participated in SF Study 1a and SF Study 1b in exchange for credit in an introductory psychology course.

Stimuli

Auditory and visual stimuli were the same as SF Study 1a.

Design

For Study 1b, each of the three pitch differences (large, octave, and M3) was presented eight times in each of the four spatial frequency difference conditions (14, 10, 6, and no circles), giving rise to a total of 96 trials per block. Participants completed two blocks (in a random order): (a) Participants were asked to select whether the first or second pitch was *higher*; (b) participants were asked to select whether the first or second pitch was *lower*.

Procedure

The sequence of trials was the same as SF Study 1a except participants always responded to the frequency of the pitch. They were told that task-irrelevant circles would only appear on some trials.

Data management

Participants. We removed one participant due to low accuracy. The average overall accuracy of the remaining 31 participants was 94.1%.

Reaction Time. Using the correct responses only, we eliminated trials where responses that were faster than 50ms (0.02% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested an inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (8.8% removed).

Results and Discussion

Reaction Time

Table 22

Spatial Frequency Study 1b*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive`	`Intercept`	7	0.0	1.00	0.075	0.350
`Non-additive`	`Intercept`	14	49.0	0.00	0.076	0.351

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `pitch difference` (continuous), and `task` (respond to higher pitch vs. lower pitch). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the three fixed effects to a non-additive model including the three fixed effects and all of their two and three-way interactions. Table 22 shows that the additive model is superior to the non-additive model. The AICc difference (ΔAICc) between the two models is 49; "strong evidence" in favor of the better model (Jeffreys, 1961).

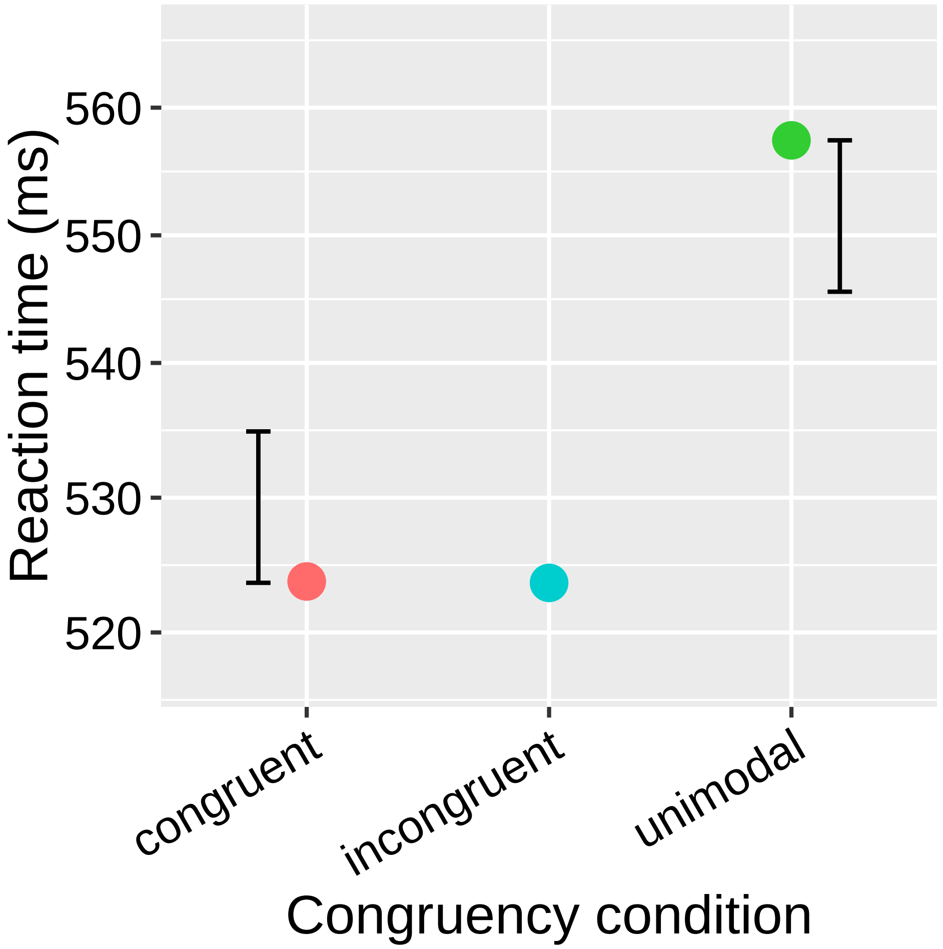


Figure 35. SF Study 1b: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars) when *pitch* is the relevant dimension. Participants were significantly *slower* in the no shape condition compared to either bimodal condition (congruent or incongruent).

Significant findings. We found no evidence of a congruency effect collapsing across pitch differences and task (Figure 35). Instead we found that participants were significantly *slower* to respond in the no shape condition (0.06, [0.04, 0.08]) than to either the congruent or incongruent conditions, which did not differ from each other (0, [−0.02, 0.02]).

We also found a significant effect of pitch difference (0.12, [0.11, 0.14]), such that participants were slowest to respond to a M3 pitch difference and fastest to respond to a large pitch difference, showing that participants were in fact completing the experiment as instructed.

Errors

Our binomial GLMM included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `task` (respond to higher pitch vs. lower pitch), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect. Because of the low error probability overall, a non-additive model failed to converge. We chose to fit an additive model with two categorical fixed effects and pitch difference as a continuous predictor since this model was clearly superior for the RT analysis as well.

There was not a significant difference between the log-odds of correct responses for the congruent and incongruent conditions ($p = 0.2$; $z = -1.28$) or for the congruent and unimodal conditions ($p = 0.61$; $z = -0.51$). The proportion of errors in the congruent, incongruent, and unimodal conditions (responding to higher pitches with the largest pitch difference) was 0.039, 0.034, and 0.037, respectively.

Conclusions

We were unable to replicate previous results finding an automatic association between pitch and spatial frequency when the auditory dimension was relevant. It remains an open question why the `unimodal` condition switched from being fastest when spatial frequency was relevant to slowest when the pitch was relevant.

SF Study 2: Top-down Influence

SF Study 1a and 1b showed a failure to replicate the pitch–spatial frequency correspondence. Next, we sought to investigate the top-down influence of task instructions on performance. For SF Study 2, we used a within-subjects design where each participants completed one block each of four pairing instructions.

Method

Participants

Thirty-two U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory and visual stimuli were the same as SF Studies 1a and 1b.

Design

We used a within-participant design in which each participant completed four blocks of 108 trials. Participants completed the four sets of experiment instructions in a random order: wider stripes/lower pitch, wider stripes/higher pitch, narrower stripes/higher pitch, or narrower stripes/lower pitch.

Procedure

At the start of each trial, the instructions were presented on the screen. Participants pressed the `SPACEBAR` to proceed with the trial. The first shape appeared for 300ms; it was accompanied by a 300ms tone. This was followed by a 500ms blank screen. Then the second shape appeared for 300ms; it was also accompanied by a 300ms tone. After both pitch/spatial frequency pairings, either an ‘s’ for shape or ‘p’ for pitch appeared on the screen to cue the participant to which modality to respond. Participants indicated whether the first or second stimulus met the instruction criteria (e.g., for wider/lower instructions, participants would respond to the *lower* pitch if a ‘p’ appeared or the *wider* stripes if an ‘s’ appeared). The cue remained on the screen until the participant responded, at which point the instruction screen reappeared to begin the next trial.

Data management

Participants. We did not remove any participants from this experiment. The average overall accuracy of the 32 participants was 90%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (5.1% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the five fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 5% of responses.

Table 23
SF Study 2*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	34	0.0	1.00	0.054	0.325
`Additive`	`Intercept`	8	388.0	0.00	0.017	0.286

Results and Discussion

Reaction Time

Model comparison. Our LMMS here included five fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch or responding to shape), `pitch difference` (continuous), and `SF cycle difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the five fixed effects to a non-additive model including the five fixed effects and all of their two, three, four, and five-way interactions. Table 23 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 388, which is "decisive evidence" in favor of the better model (Jeffreys, 1961).

Significant findings. Figure 36 shows the predictors of logRT and their 95% confidence intervals. The `congruency x compatibility` interaction is the most important.

Figure 37 shows the top-down influence of instructions on performance. Although there is a clear *congruency* advantage on compatible trials (0.29, [0.24, 0.33]), there is also a significant *incongruency* advantage on incompatible trials (−0.25, [−0.29, −0.2]). However, the figure also shows that participants are significantly slower to respond to the incongruent trials in the incompatible condition than congruent trials in the compatible condition. This means that it takes longer overall to pair together the perceptually incongruent endpoints, showing some evidence of perceptual congruency impacting performance.

Additionally, a three-way interaction model with `congruency` × `compatibility` and `trial number` as fixed effects shows that the RT difference between congruent and incongruent trials does not change as a function of trial number (−0.72, [−5.4, 3.96]), ruling out a learning explanation.

Errors

Model comparison. Our binomial GLMMS included five fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch vs. responding to shape), `SF cycle difference` (continuous), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

Because of the low error probability overall, a non-additive model including the five fixed effects and all of their two-, three-, four-, and five-way interactions failed to converge. Instead, we compared a non-additive model including the five fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the five fixed effects. The AICc difference

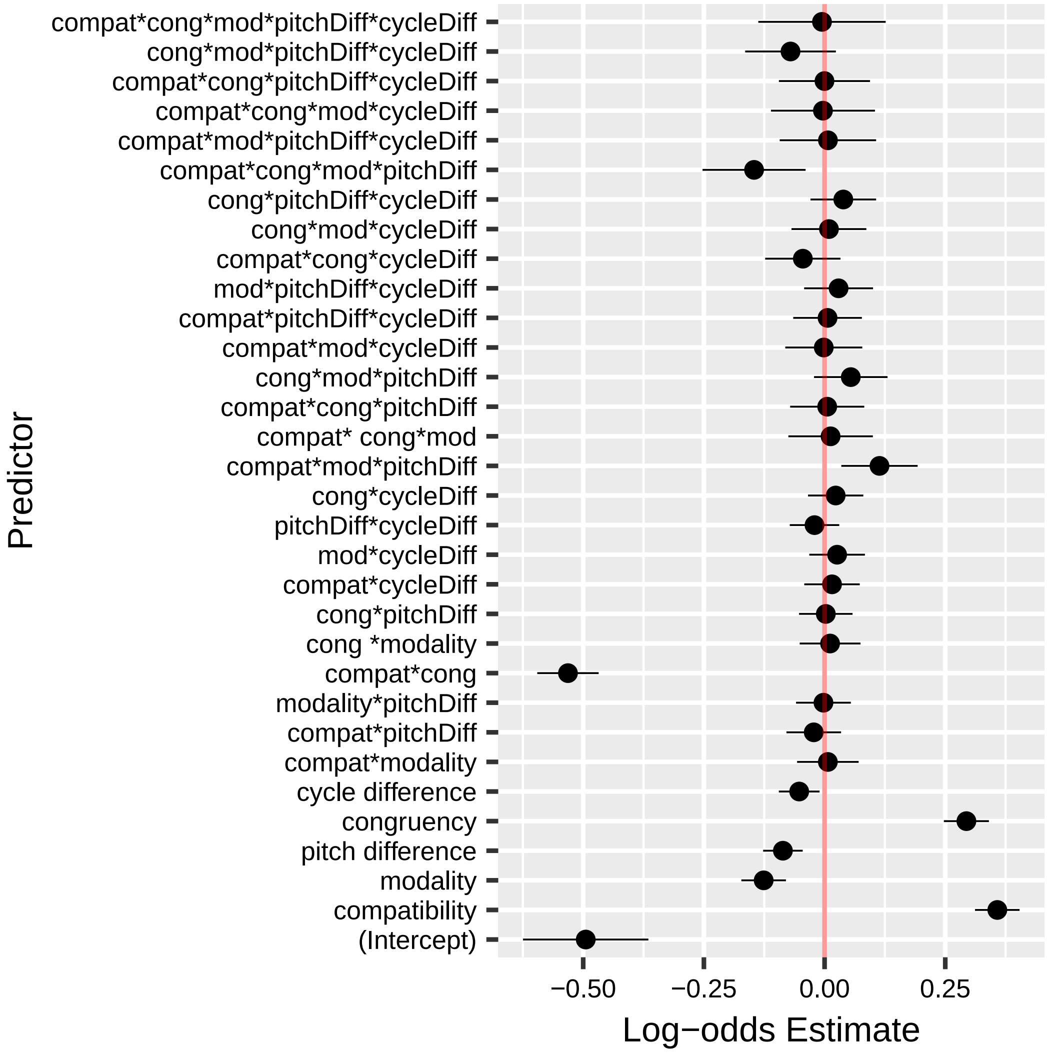


Figure 36. *SF Study 2*: Coefficient plot for the five fixed effect predictors and their interactions (with 95% confidence intervals) Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference; cycleDiff = cycle difference.

(ΔAICc) between the two models is 115.7, which is “decisive evidence” in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible blocks is $p \approx 0$ ($z = 3.4$), meaning listeners make more errors on incompatible blocks. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 8.29$), meaning listeners make a more errors on incongruent trials. These effects are converted to proportions and shown in Table 34. We can rule out a speed-accuracy tradeoff because errors and reaction times (shown in Figure 37) follow the same pattern.

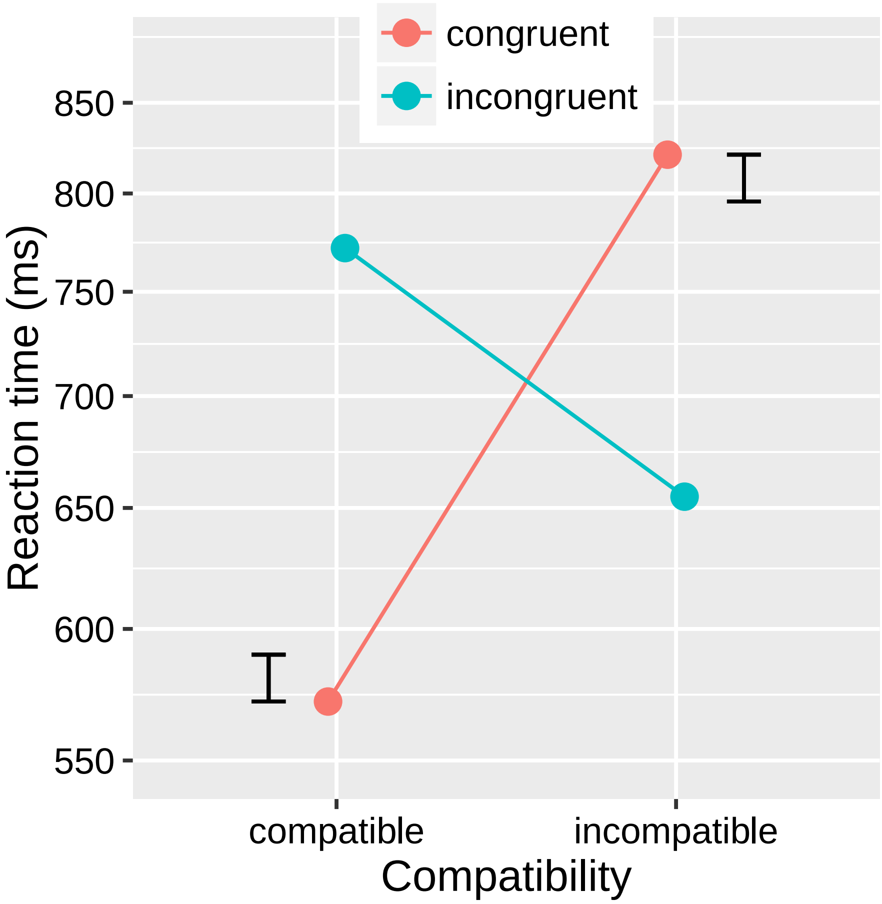


Figure 37. *SF Study 2*: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower LSD bars). Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

Table 24
SF Study 2*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.10	0.13
Incongruent trials	0.18	0.07

Conclusions

The results here are a hybrid of the pitch–size and pitch–height results. We did find top-down influence of instructions on performance with the pitch–spatial frequency correspondence. However, the congruency to incongruency advantage reversal was not as extreme as in the pitch–size correspondence, showing that there is some evidence for an automatic association between pitch and spatial frequency. Despite this, differing instructions seem to be mainly driving the results.

SF Study 3: Lexical Overlap

SF Study 3 investigated the effect of lexical overlap. We used the words *high* and *low* to describe both the spatial frequencies and auditory pitches. It is worth noting that although ‘high’ and ‘low’ do technically describe the stimuli in Figure 4c, few participants knew the term *spatial frequency* or used ‘high’ and ‘low’ as natural descriptors for the stimuli. They were much more likely to use the ‘narrow’ and ‘wide’ stripe descriptors we used in SF Study 2.

Method

Participants

Thirty-six U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory and visual stimuli were the same as SF Study 2.

Design

Participants completed one block using each of four sets of experiment instructions in a random order: lower spatial frequency/lower pitch, lower spatial frequency/higher pitch, higher spatial frequency/higher pitch, and higher spatial frequency/lower pitch.

Procedure

The procedure was the same as SF Study 2.

Data management

Participants. We removed one participant with less than 60% accuracy. The average overall accuracy of the remaining 35 participants was 93.2%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (3.2% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the five fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 5.7% of responses.

Results and Discussion

Reaction Time

Table 25
SF Study 3*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	34	0.0	1.00	0.048	0.328
`Additive`	`Intercept`	8	275.1	0.00	0.022	0.297

Model comparison. We compared an additive model including just the five fixed effects (see SF Study 2) to a non-additive model including the five fixed effects and all of their two, three, four, and five-way interactions. Table 25 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 275.1, which is "decisive evidence" in favor of the better model (Jeffreys, 1961).

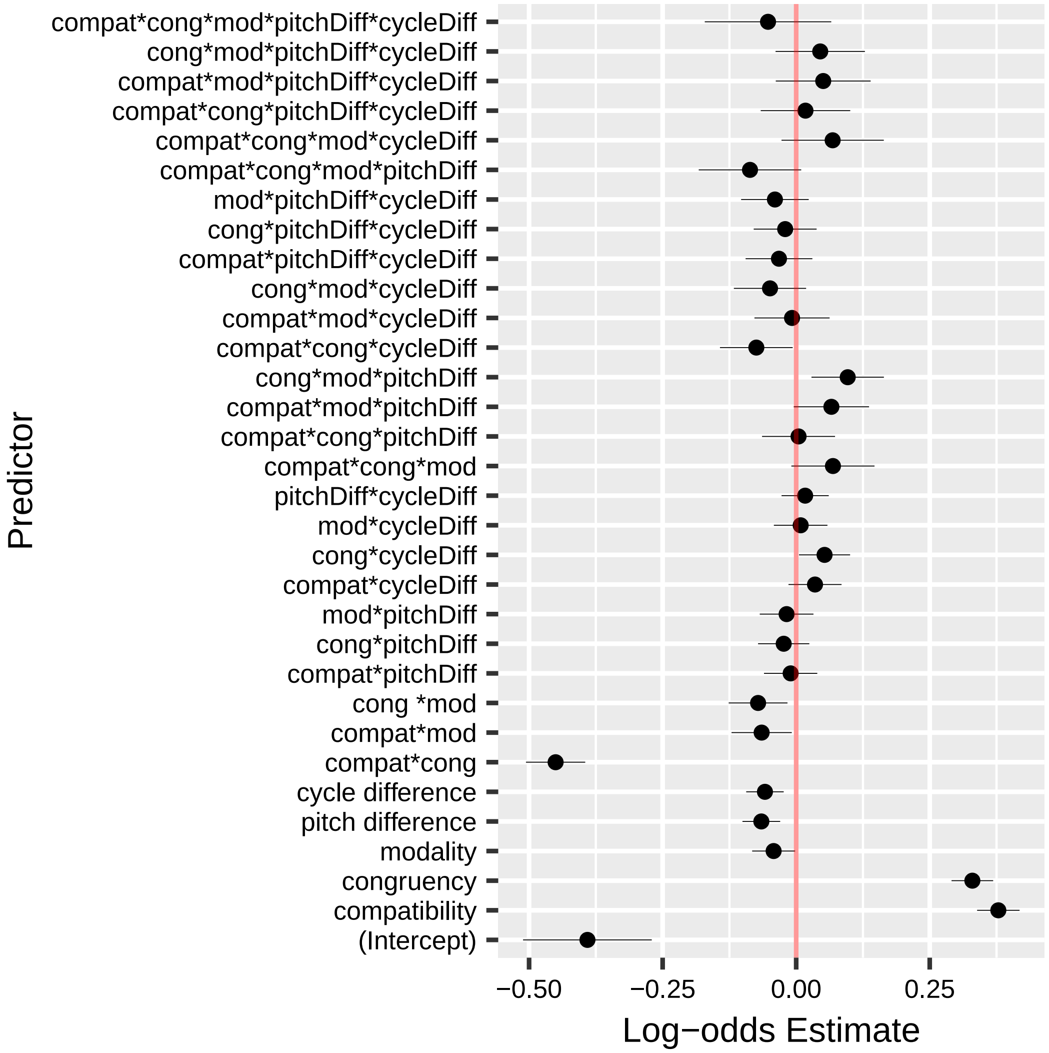


Figure 38. SF Study 3: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals) Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference.

Significant findings. Figure 38 shows the predictors of logRT and their 95% confidence intervals. The `congruency x compatibility` interaction is the most important.

Figure 39 shows that instructions and perceptual congruency both impacted the results. Although there is a clear *congruency* advantage on compatible trials (0.33, [0.29, 0.37]), there is also a moderate *incongruency* advantage on incompatible trials (−0.13, [−0.17, −0.09]). Further, the

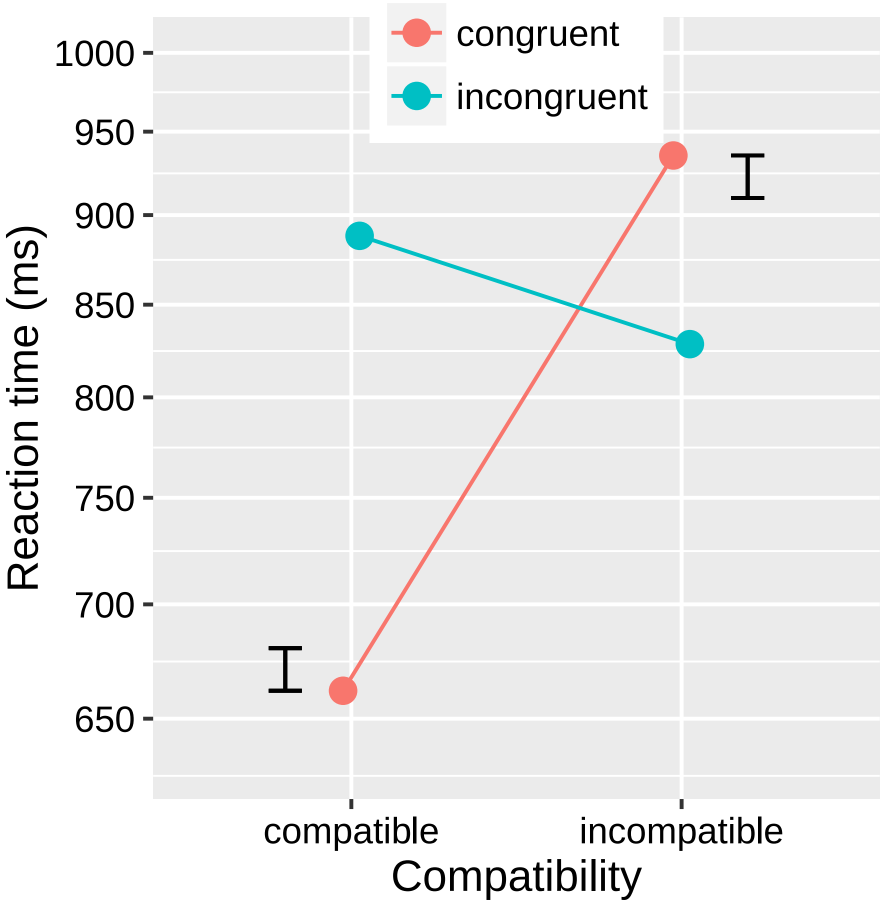


Figure 39. SF Study 3: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower `LSD` bars). Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

figure shows that participants were significantly slower to respond to the incongruent incompatible trials than congruent compatible trials, meaning that overall it takes longer to pair together the perceptually incongruent endpoints.

Additionally, a three-way interaction model with `congruency` × `compatibility` and `trial number` as fixed effects shows that the RT difference between congruent and incongruent trials does not change as a function of trial number (−0.14, [−4.62, 4.33]).

Errors

Model comparison. Because of the low error probability overall, a non-additive model including the five fixed effects (see SF Study 2) and all of their interactions failed to converge. Instead, we compared a non-additive model including the five fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the five fixed effects. The AICc difference (ΔAICc) between the two models is 56.3, which is "very strong evidence" in favor of the non-additive model (Jeffreys, 1961).

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible blocks is $p \approx 0$ ($z = 3.4$), meaning listeners make more errors on incompatible blocks. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 8.29$), meaning listeners make a more errors on incongruent trials. These effects are converted to proportions and shown in Table 34. We can rule out a speed-accuracy tradeoff because errors and reaction times (shown in Figure 39) follow the same pattern.

Table 26

SF Study 3*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.04	0.08
Incongruent trials	0.09	0.07

Conclusions

The results here are again a hybrid of the pitch–size and pitch–height results. We found some evidence for a congruency effect in addition to a top-down effect of instructions when the words to describe the two modalities overlapped. There was only a moderate reversal of the congruency effect on incompatible trials, which shows that participants had a harder time grouping low spatial frequency shapes with high pitches than with low pitches. The reaction times were also slower overall in the incompatible condition because incongruent trials were not as fast as congruent trials in the compatible condition.

With the height correspondence, lexical overlap did not influence the results. Here however, using the words ‘high’ and ‘low’ to describe both pitches and spatial frequencies made a difference. When given different words, there was less evidence of a congruency effect, but when given the same words for the two modalities, participants had a harder time grouping the incongruent endpoints.

Pitch–Spatial Frequency Discussion

SF Studies 1a and 1b failed to provide evidence for an automatic association between pitch and spatial frequency. Although the unimodal condition reversed from being slower or faster than the bimodal conditions, the congruent and incongruent conditions never differed in overall reaction time. Surprisingly, the results also differed depending on the task we gave participants; when asked to select the narrower shape, participants were faster to pair narrow shapes and high pitches, but when asked to select the wider shape, participants were faster to pair wider shapes with high pitches. The reason for this congruency reversal effect remains an open question, but nonetheless shows that there is no fixed pairing between pitch and spatial frequency at a perceptual level.

In SF Studies 2 and 3, the results were a hybrid between pitch–size (as in Figure 3b) and pitch–height (as in Figure 3c). In both studies, there was a `congruency x compatibility` interaction, showing that task instructions impacted performance. However, there was also some evidence that participants had a harder time grouping perceptually incongruent endpoints even when called to in the instructions.

Further, unlike with the pitch–height correspondence, here there was a small effect of lexical overlap; when given the same words for the two modalities, participants had a harder time grouping the incongruent endpoints. It is possible that because spatial frequency was a relatively unknown concept, the memory load was simply too high for participants. Alternatively, the effect of lexical overlap could provide further evidence that the pitch–spatial frequency correspondence is a decision-level effect because congruency affected the results more when the dimensions are given overlapping labels.

From these results, we conclude that there is some evidence for an automatic congruency relationship between pitch and spatial frequency at a perceptual level. However, most of the evidence points to top-down, decision level influences on pitch–spatial frequency pairings.

Chapter 5: Pitch–Brightness Correspondence

Auditory pitch and visual brightness may be a *structural* correspondence because it is unlikely to arise from environmental associations (Maurer et al., 2012), yet there is evidence for a non-arbitrary mapping between the dimensions. Synesthetes and control participants match pitch to brightness in a consistent way, with higher pitches eliciting brighter/lighter colors (Hubbard, 1996; Ward, Huckstep, & Tsakanikos, 2006). Most most three-year-olds also match a darker ball with a low-pitched sound and lighter ball with a high-pitched sound in a two-alternative forced-choice task (Mondloch & Maurer, 2004). Additionally, differences in brightness affect both the speed and accuracy of pitch judgements (Marks, 1974, 1987a). The congruent endpoint mapping in this case is: dark/low and bright/high.

It is worth noting here that although most studies looking at these effects use the term *brightness*, the relationship seems to hold with both *brighter* stimuli (e.g., luminous spots in darkness) and *lighter* stimuli (e.g., surface differences) being paired with higher pitches (Marks, 1987a). When discussing surface differences, it is also necessary to keep in mind the background surface. For example, studies on the number–luminance correspondence find that it is actually luminance *contrast* that influences the pairing such that higher numbers are paired with the value that contrasts more with the background (i.e., with a black background, the congruent pairing is brighter/higher, but with a white background the pairing reverses to darker/higher; Cohen Kadosh, Cohen Kadosh, & Henik, 2008; Cohen Kadosh & Henik, 2006; Gebuis & van der Smagt, 2011).

As in Chapters 3 & 4, here I first completed a conceptual replication of the correspondence to establish the generalizability of the pitch–brightness congruency effect. In Bright Study 1a, brightness was the relevant dimension, and in Bright Study 1b, pitch was the relevant dimension.

Following, I investigated the influence of task instructions on the pitch–brightness endpoint mapping. Specifically, Bright Study 2 included four types of trials: two with compatible instructions and two with incompatible instructions.

- **Compatible instructions:** blocks where "perceptually congruent" endpoints are paired. In *Block 1* listeners selected either the dark shape or low pitch; in *Block 2* listeners selected either the bright shape or high pitch.
- **Incompatible instructions:** blocks where "perceptually incongruent" endpoints are paired. In *Block 3* listeners selected either the dark shape or high pitch; in *Block 4* listeners selected either the bright shape or low pitch.

Bright Study 1a: Brightness-Relevant Replication

Participants were randomly assigned to complete Study 1a or Study 1b first. In both versions, there was one relevant dimension for the entire experiment while the other dimension remained irrelevant. In Bright Study 1a, brightness was the relevant dimension.

Method

Participants

The same forty-six U.Va. undergraduate students participated in Bright Study 1a and Bright Study 1b in exchange for credit in an introductory psychology course.

Stimuli

Visual shapes. Shapes were all presented on a black screen (see Figure 4d); we used three different (within-participant) brightness pairings based on the 256-entry `Matlab` gray colormap (0 = black, 255 = white): a difference of 200 (50 vs. 250), 150 (75 vs. 225), and 100 (100 vs. 200).

Auditory pitches. We used the same three pitch pairings as Size Studies 1–2.

Design

Participants were randomly assigned to one of two experiment orders [bright-relevant first (Study 1a) or pitch-relevant first (Study 1b)].

For Study 1a, each of the three brightness differences (200, 150, 100) was presented 12 times in each of the four pitch difference conditions (large, octave, M3, and no sound), giving rise to a total of 144 trials per block. Participants completed two blocks (in a random order): (a) Participants were asked to select whether the first or second shape was *brighter*; (b) participants were asked to select whether the first or second shape was *darker*.

Procedure

The procedure followed the sequence of events presented in Figure 2 with one exception: instead of waiting until the end of the trial to discover whether they should respond to the pitch or the shape, participants always responded to the brightness of the shape and pitch was always the irrelevant dimension. After the second stimulus appeared for 300 ms, it was replaced by a blank screen until the participant responded. Participants were told they could respond as soon as the second shape appeared. They were also told that task-irrelevant sounds would only appear on some trials.

Data management

Participants. In Bright Study 1, we removed one participant with less than 60% accuracy and four participants who completed fewer than 60% of the total trials. The average overall accuracy of the remaining 41 participants was 98.7%.

Reaction Time. Using the correct responses only, we eliminated trials where responses that were faster than 50ms (0.03% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested an inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (11.6% removed).

Results and Discussion

Reaction Time

Table 27

Bright Study 1a*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	14	0.0	1.00	0.023	0.351
`Additive`	`Intercept`	7	80.1	0.00	0.013	0.340

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `brightness difference` (continuous), and `task` (respond to brighter shape vs. darker shape). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the three fixed effects to a non-additive model including the three fixed effects and all of their two and three-way interactions. Table 27 shows that the non-additive model is clearly superior to the additive model. The AICc difference (ΔAICc) between the two models is 80.1; "very strong evidence" in favor of the better model (Jeffreys, 1961).

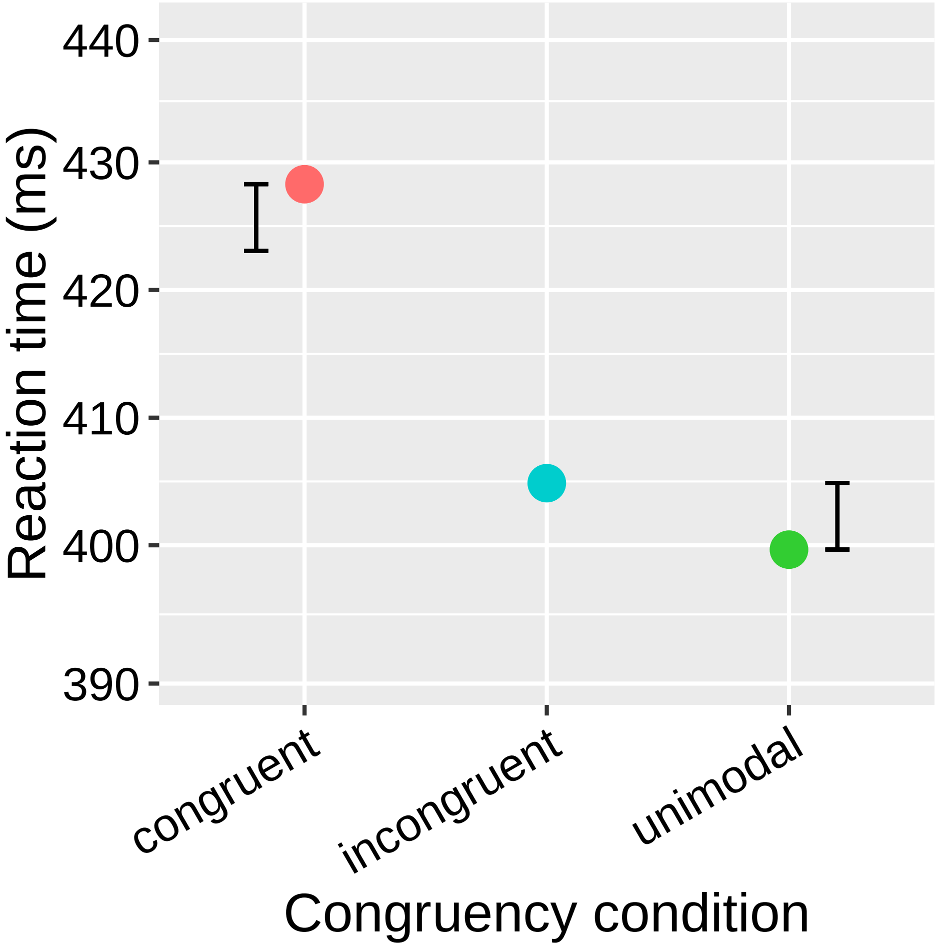


Figure 40. Bright Study 1a: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars) when *brightness* is the relevant dimension. Participants did not show a congruency effect; in fact, RTs were significantly *slower* in the congruent condition than both the incongruent and unimodal conditions.

Significant findings. We found no evidence of a congruency effect collapsing across brightness differences and task (Figure 40). In fact, participants were significantly *faster* to respond in the incongruent condition (−0.13, [−0.14, −0.11]) and no sound condition −0.1, [−0.12, −0.08]) than in the congruent condition.

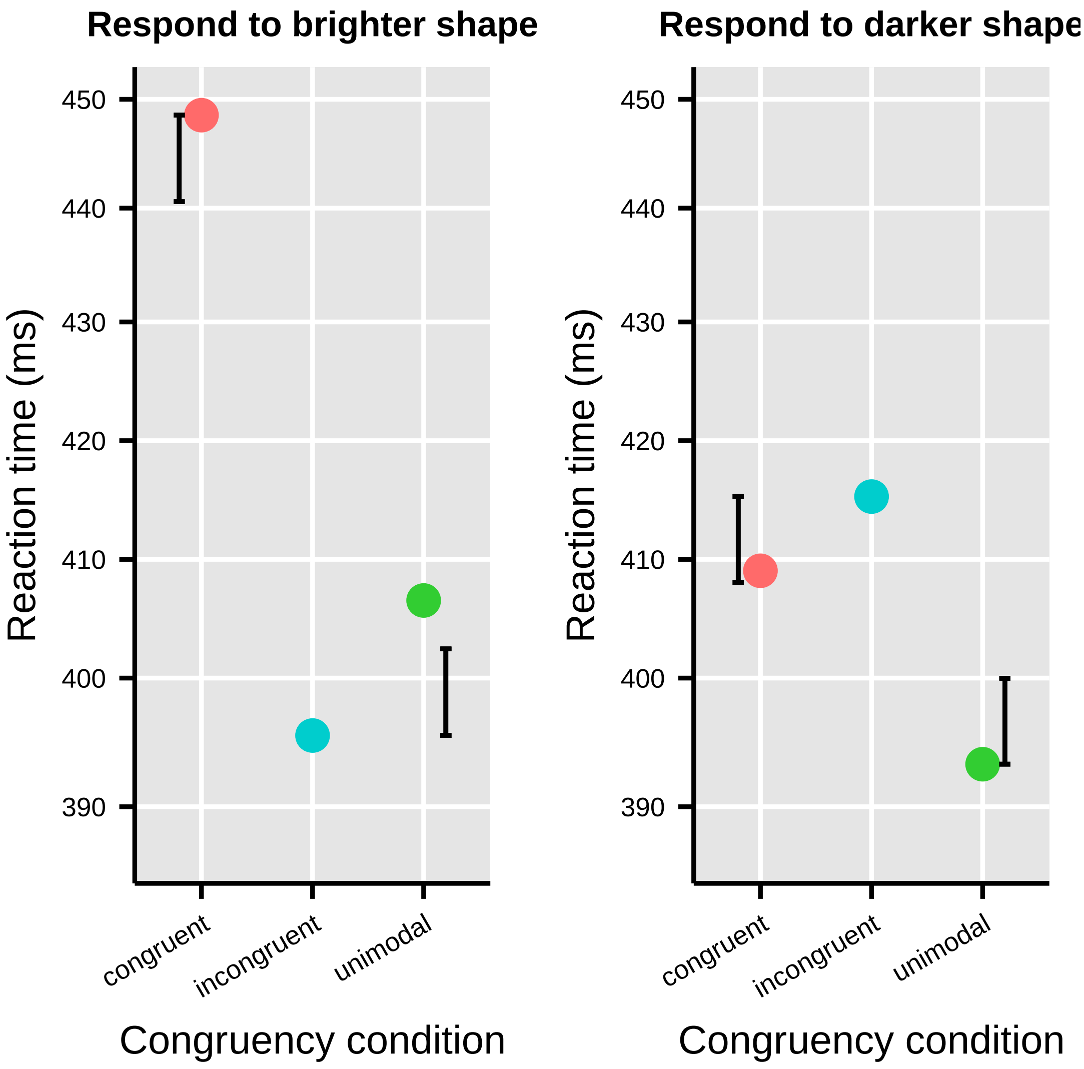


Figure 41. Bright Study 1a: Effect plot (with upper and lower LSD bar) showing the main effect of congruency (x-axis) by task (panels) when *brightness* is the relevant dimension.

When we look separately at blocks requiring participants to select the *brighter* shape and *darker* shape, we see that the congruency disadvantage is only true on the *brighter* task (Figure 41). It is an open question why the congruency effect reverses when participants are asked which shape is brighter compared to which shape is darker.

There was also a significant effect of brightness difference, such that participants are slower to respond to a smaller difference between the visual stimuli (−0.02, [−0.03, −0.01]), which shows that participants were completing the experiment as instructed.

Errors

Our poisson GLMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `brightness difference` (continuous), and `task` (respond to brighter vs. darker shape). We also included the subject-by-subject variation in the intercept as a random effect. We compared an additive model including just the three fixed effects to a non-additive model including the fixed effects and their two- and three-way interactions. The AICc difference (ΔAICc) between the two models is 13.9, which is "strong evidence" in favor of the non-additive model (Jeffreys, 1961).

There was not a significant difference between the logarithm of correct responses for the congruent and incongruent conditions ($p = 0.8$; $z = 0.26$) or for the congruent and unimodal conditions ($p = 0.94$; $z = -0.08$). The proportion of errors in the congruent, incongruent, and unimodal conditions (responding to brighter shapes with the mean brightness difference) was 0.013, 0.008, and 0.015, respectively. Because there is no difference in errors across conditions, we can rule out the possibility of a speed-accuracy tradeoff.

Conclusions

It is unclear why the congruency advantage reverses when the task changes from identifying the brighter shape to identifying the darker shape. Nonetheless, we were unable to replicate previous results finding an automatic association between pitch and brightness when the visual dimension was relevant.

Bright Study 1b: Pitch-Relevant Replication

Bright Study 1b repeated Study 1a with pitch rather than brightness as the relevant dimension.

Method

Participants

The same forty-six U.Va. undergraduate students participated in Bright Study 1a and Bright Study 1b in exchange for credit in an introductory psychology course.

Stimuli

Auditory and visual stimuli were the same as Bright Study 1a.

Design

For Study 1b, each of the three pitch differences (large, octave, and M3) was presented 12 times in each of the four brightness difference conditions (200, 150, 100, and no circles), giving rise to a total of 144 trials per block. Participants completed two blocks (in a random order): (a) Participants were asked to select whether the first or second pitch was *higher*; (b) participants were asked to select whether the first or second pitch was *lower*.

Procedure

The sequence of trials was the same as Bright Study 1a except participants always responded to the height of the pitch. They were told that task-irrelevant circles would only appear on some trials.

Data management

Participants. We did not need to remove any participants based on accuracy. We removed three participants who completed fewer than 60% of the trials. The average overall accuracy of the remaining 43 participants was 93.5%.

Reaction Time. Using the correct responses only, we eliminated trials where responses that were faster than 50ms (0.06% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested an inverse square root transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (6.7% removed).

Results and Discussion

Reaction Time

Table 28

Bright Study 1b*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive`	`Intercept`	7	0.0	1.00	0.038	0.409
`Non-additive`	`Intercept`	14	57.2	0.00	0.039	0.409

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `pitch difference` (continuous), and `task` (respond to higher pitch vs. lower pitch). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the three fixed effects to a non-additive model including the three fixed effects and all of their two and three-way interactions. Table 28 shows that the additive model is superior to the non-additive model. The AICc difference (ΔAICc) between the two models is 57.2; "very strong evidence" in favor of the better model (Jeffreys, 1961).

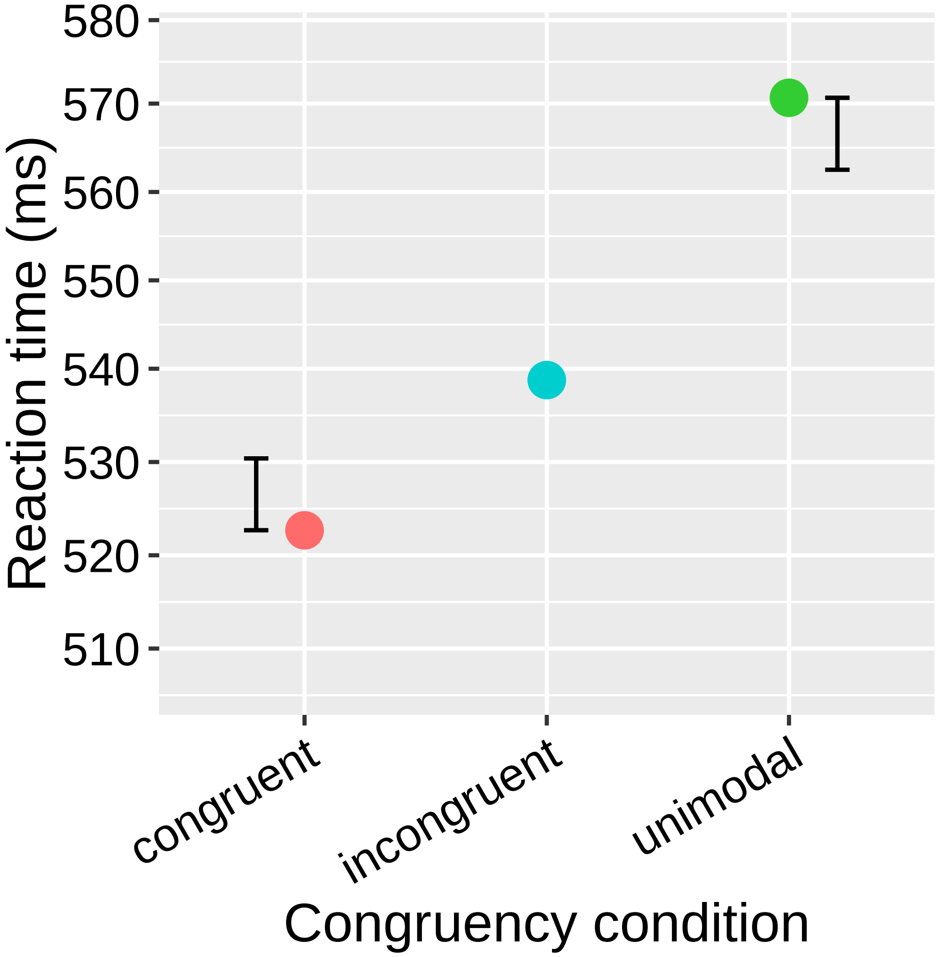


Figure 42. Bright Study 1b: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars) when *pitch* is the relevant dimension. Participants showed a congruency effect in that RTs were faster in the congruent condition than the incongruent and unimodal conditions.

Significant findings. We found a significant congruency effect collapsing across pitch differences and task (Figure 42). Participants were significantly *faster* to respond in the congruent condition than the no sound condition (0.09, [0.07, 0.1]) and the incongruent condition (0.03, [0.02, 0.04]).

Additionally, there was a significant effect of pitch difference, such that participants were slower to respond to a smaller pitch difference (0.09, [0.085, 0.1]), which shows that participants were completing the experiment as instructed.

Errors

Our binomial GLMM included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `task` (respond to higher pitch vs. lower pitch), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect. Because of the low error probability overall, a non-additive model failed to converge. We chose to fit an additive model with two categorical fixed effects and pitch difference as a continuous predictor since this model was clearly superior for the RT analysis as well.

There was a significant difference between the log-odds of correct responses for the congruent and incongruent conditions ($p \approx 0$; $z = 10.79$) and for the congruent and unimodal conditions ($p = 0.02$; $z = 2.36$). The proportion of errors in the congruent, incongruent, and unimodal conditions (responding to larger shapes with a 50-pixel size difference) was 0.02, 0.051, and 0.026, respectively. Because there is a congruency effect in errors as well as reaction time (as shown in Figure 42), we can rule out the possibility of a speed-accuracy tradeoff.

Conclusions

We found a marginal congruency effect between low pitches and dark shapes (and high pitches/bright shapes) when pitch was the relevant dimension. Findings here are similar to Height Study 1b, and are in line with previous research that found the pitch-relevant condition was more strongly affected by congruency (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010).

Bright Study 2: Top-down Influence

Together, Bright Studies 1a and 1b provide mixed evidence for an automatic association between pitch and brightness. Next, we sought to investigate the top-down influence of task instructions on performance. For Bright Study 2, we used a within-subjects design where each participants completed one block each of four pairing instructions.

Method

Participants

Fifty U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory and visual stimuli were the same as Bright Studies 1a and 1b.

Design

We used a within-participant design in which each participant completed four blocks of 72 trials. Participants completed the four sets of experiment instructions in a random order: darker circle/lower pitch, darker circle/higher pitcher, brighter circle/higher pitch, or brighter circle/lower pitch.

Procedure

At the start of each trial, the instructions were presented on the screen. Participants pressed the `SPACEBAR` to proceed with the trial. The first shape appeared for 300ms; it was accompanied by a 300ms tone. This was followed by a 500ms blank screen. Then the second shape appeared for 300ms; it was also accompanied by a 300ms tone. After both pitch/brightness pairings, either an ‘s’ for shape or ‘p’ for pitch appeared on the screen to cue the participant to which modality to respond. Participants indicated whether the first or second stimulus met the instruction criteria (e.g., for darker/lower instructions, participants would respond to the *lower* pitch if a ‘p’ appeared or the *darker* shape if an ‘s’ appeared). The cue remained on the screen until the participant responded, at which point the instruction screen reappeared to begin the next trial.

Data management

Participants. We removed one participant with less than 60% accuracy. The average overall accuracy of the remaining 49 participants was 91.7%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (5.8% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the five fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 5.4% of responses.

Table 29
Bright Study 2*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	34	0.0	1.00	0.078	0.352
`Additive`	`Intercept`	8	864.0	0.00	0.014	0.279

Results and Discussion

Reaction Time

Model comparison. Our LMMS included five fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch or responding to shape), `bright difference` (continuous), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the five fixed effects to a non-additive model including the five fixed effects and all of their two-, three-, four-, and five-way interactions. Table 29 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 864, which is “decisive evidence” in favor of the better model (Jeffreys, 1961) .

Significant findings. Figure 43 shows the predictors of logRT and their 95% confidence intervals. The `congruency x compatibility` interaction is the most important.

Figure 44 shows that instructions and perceptual congruency both impacted the results. Although there is a clear *congruency* advantage on compatible trials (0.43, [0.39, 0.47]), there is also a significant *incongruency* advantage on incompatible trials (−0.31, [−0.36, −0.27]). Further, the figure shows that participants are significantly slower to respond to the incongruent incompatible trials than congruent compatible trials, which means that it takes longer overall to pair together the perceptually incongruent endpoints.

Additionally, a three-way interaction model with `congruency` × `compatibility` and `trial number` as fixed effects shows that the RT difference between congruent and incongruent trials does not change as a function of trial number (0.58, [−3.96, 5.13]), ruling out a learning explanation.

Errors

Model comparison. Our binomial GLMMS included five fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch vs. responding to shape), `brightness difference` (continuous), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

Because of the low error probability overall, a non-additive model including the five fixed effects and all of their two-, three-, four-and five-way interactions failed to converge. Instead, we compared a non-additive model including the five fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the five fixed effects. The AICc difference (ΔAICc) between the two models is 25.3, which is “strong evidence” in favor of the non-additive model (Jeffreys, 1961).

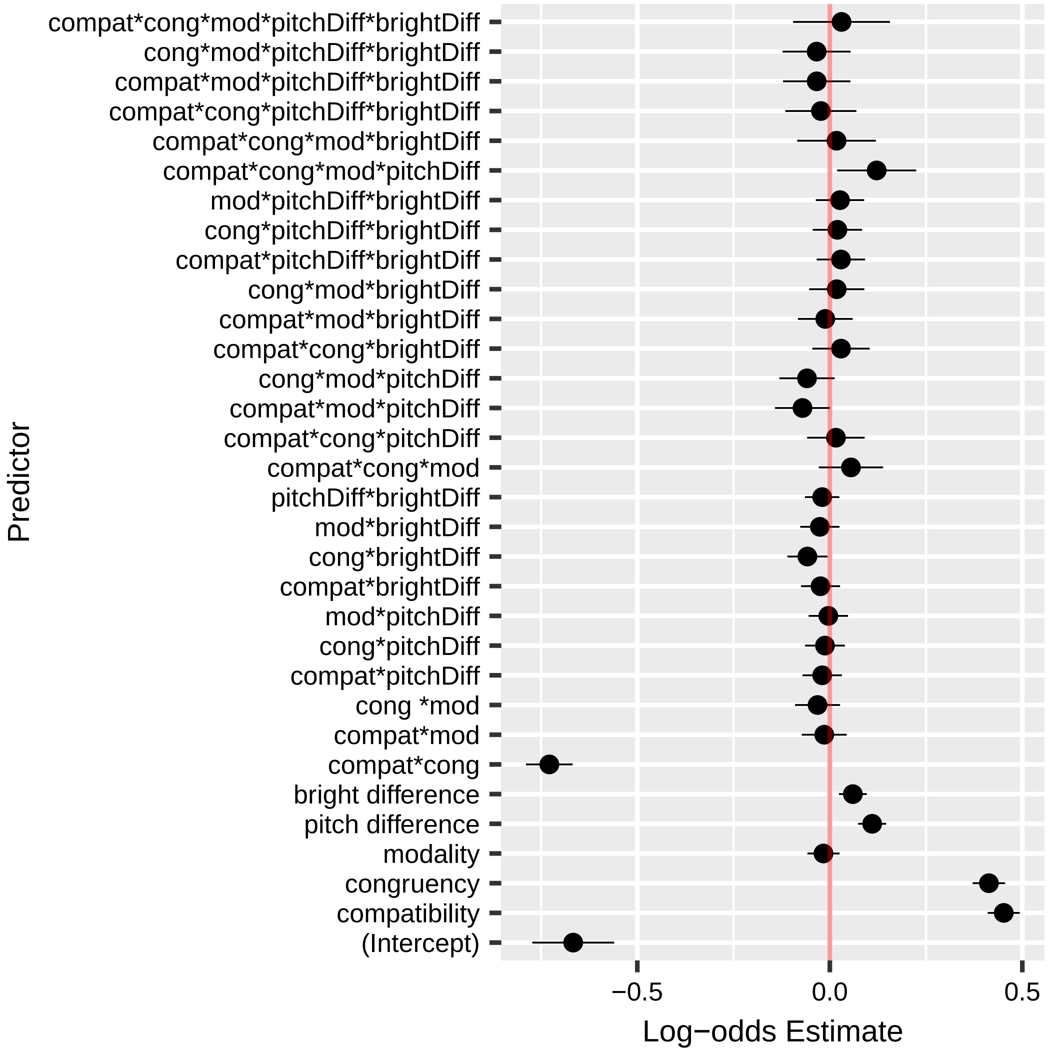


Figure 43. Bright Study 2: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals) Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference; brightDiff = brightness difference.

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible blocks is $p \approx 0$ ($z = 8.64$), meaning listeners make more errors on incompatible blocks. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 10.92$), meaning listeners make a more errors on incongruent trials. These effects are converted to proportions and shown in Table 30. The error rates follow a similar pattern as the reaction times (shown in Figure 44), meaning there is not a speed-accuracy tradeoff.

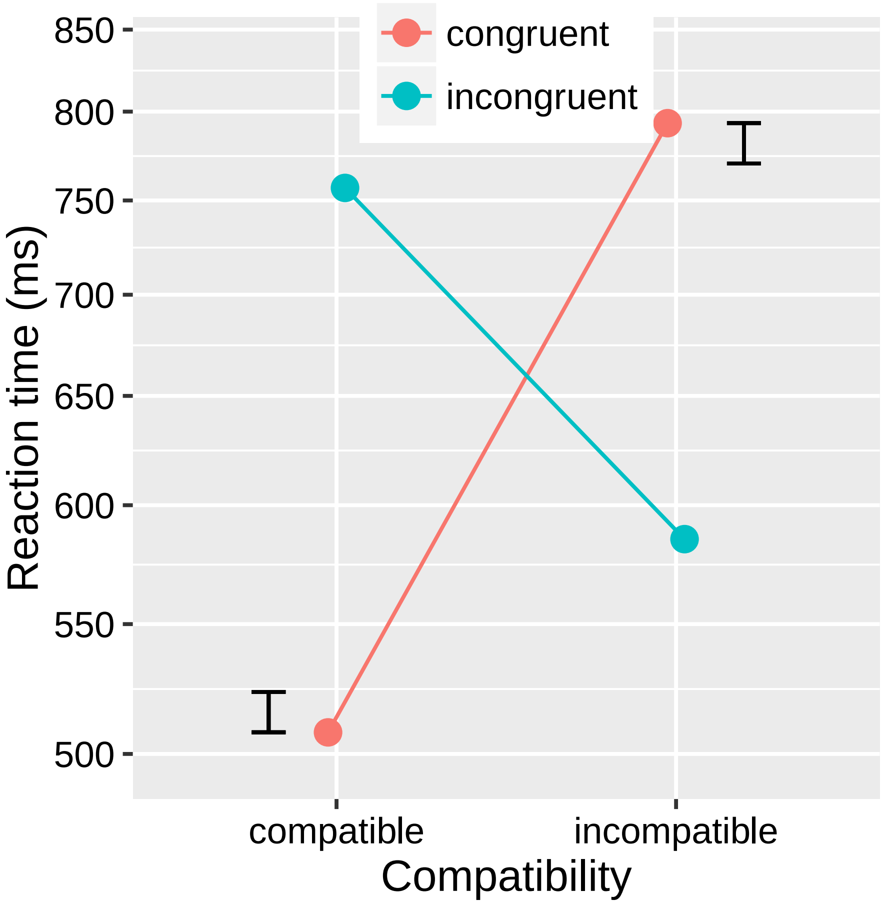


Figure 44. Bright Study 2: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower LSD bars). Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

Table 30
Bright Study 2*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.04	0.09
Incongruent trials	0.11	0.11

Conclusions

The pitch–brightness results are again a hybrid of pitch–size and pitch–height results. We found top-down influence of instructions on performance, but the congruency to incongruency advantage reversal was not as complete as in the pitch–size correspondence. This shows that although there is some evidence for an automatic association between pitch and spatial frequency, differing instructions seem to be mainly driving the results.

Pitch–Brightness Discussion

Bright Studies 1a and 1b provided mixed support for an automatic association between pitch and brightness. In Study 1a, the results differed depending on the task we gave participants; when asked to select the brighter shape, participants were faster to pair dark shapes and high pitches (i.e., the perceptually incongruent dimensions), but when asked to select the darker shape, participants were faster in the unimodal condition than either of the bimodal conditions. The reason for this difference remains an open question, but nonetheless shows that there does not appear to be a fixed pairing when the visual dimension is relevant.

In contrast, Bright Study 1b found a significant congruency effect when the auditory dimension is relevant. This supports the findings of Height Study 1b and previous research that found auditory-relevant conditions to be more strongly affected by congruency than visual-relevant conditions (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010).

Bright Study 2 also showed mixed support for an automatic association. We found a `congruency` × `compatibility` interaction, showing that task instructions impacted performance. However, there was also some evidence that participants had a harder time grouping perceptually incongruent endpoints even when called to in the instructions. The results here are a hybrid between pitch–size (as in Figure 3b) and pitch–height (as in Figure 3c).

From these results, we conclude that there is some evidence for an automatic congruency relationship between pitch and brightness at a perceptual level. However, most of the evidence points to top-down, decision level influences on pitch–brightness pairings.

Chapter 6: Pitch–Sharpness Correspondence

Like pitch–brightness, auditory pitch and visual sharpness may be an example of a *structural* correspondence. There is no obvious connection between the roundness/sharpness of an object and the frequency of the sound it makes, yet infants and adults alike consistently match endpoints suggesting an intrinsic mapping between brain areas responsible for auditory and visual perception (Marks, 1987a; Maurer et al., 2012; O'Boyle & Tarte, 1980; Parise & Spence, 2009). The congruent endpoint mapping in this case is: round/low and sharp/high.

As in Chapters 3, 4, & 5, here I first completed a conceptual replication of the correspondence to establish the generalizability of the pitch–sharpness congruency effect. In Sharp Study 1a, sharpness was the relevant dimension, and in Sharp Study 1b, pitch was the relevant dimension.

Following, I investigated the influence of task instructions on the pitch–sharpness endpoint mapping. Specifically, Sharp Study 2 included four types of trials: two with compatible instructions and two with incompatible instructions.

- **Compatible instructions:** blocks where "perceptually congruent" endpoints are paired. In *Block 1* listeners selected either the round shape or low pitch; in *Block 2* listeners selected either the sharp shape or high pitch.

- **Incompatible instructions:** blocks where "perceptually incongruent" endpoints are paired. In *Block 3* listeners selected either the round shape or high pitch; in *Block 4* listeners selected either the sharp shape or low pitch.

Sharp Study 1a: Sharpness-Relevant Replication

Participants were randomly assigned to complete Sharp Study 1a or Study 1b first. In both versions, there was one relevant dimension for the entire experiment while the other dimension remained irrelevant. In Sharp Study 1a, sharpness was the relevant dimension.

Method

Participants

The same thirty-four U.Va. undergraduate students participated in Sharp Study 1a and Sharp Study 1b in exchange for credit in an introductory psychology course.

Stimuli

Visual shapes. Six sharp and rounded shape pairings were generated using Matlab (MATLAB, 2013). All shapes were set to be a white stimulus on a black background. Sharp shapes included between 4-30 polar coordinates sorted and successively connected on a Cartesian grid. Rounded shapes were created by performing a quadratic spline on the sharp shapes, thus controlling for overall size and number of edges (see Figure 4e).

Auditory pitches. We used the same three pitch pairings as Size Studies 1–2.

Design

Participants were randomly assigned to one of two experiment orders [sharp-relevant first (Study 1a) or pitch-relevant first (Study 1b)].

For Study 1a, each of the six sharpness pairings was presented 4 times in each of the four pitch difference conditions (large, octave, M3, and no sound), giving rise to a total of 108 trials per block. Participants completed two blocks (in a random order): (a) Participants were asked to select whether the first or second shape was *sharper*; (b) participants were asked to select whether the first or second shape was *rounder*.

Procedure

The procedure followed the sequence of events presented in Figure 2 with one exception: instead of waiting until the end of the trial to discover whether they should respond to the pitch or the shape, participants always responded to the sharpness of the shape and pitch was always the irrelevant dimension. After the second stimulus appeared for 300 ms, it was replaced by a blank screen until the participant responded. Participants were told they could respond as soon as the second shape appeared. They were also told that task-irrelevant sounds would only appear on some trials.

Data management

Participants. We did not need to remove any participants from Sharp Study 1a. The average overall accuracy of the 34 participants was 99.1%.

Reaction Time. Using the correct responses only, we eliminated trials where responses that were faster than 50ms (0.03% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested an inverse transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (8.9% removed).

Results and Discussion

Reaction Time

Table 31

Sharp Study 1a*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive`	`Intercept`	11	0.0	1.00	0.005	0.362
`Non-additive`	`Intercept`	38	171.6	0.00	0.007	0.363

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `sharpness pairing` (1–6), and `task` (respond to sharper shape vs. rounder shape). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the three fixed effects to a non-additive model including the three fixed effects and all of their two and three-way interactions. Table 31 shows that the additive model is clearly superior to the non-additive model. The AICc difference (ΔAICc) between the two models is 171.6; "decisive evidence" in favor of the better model (Jeffreys, 1961).

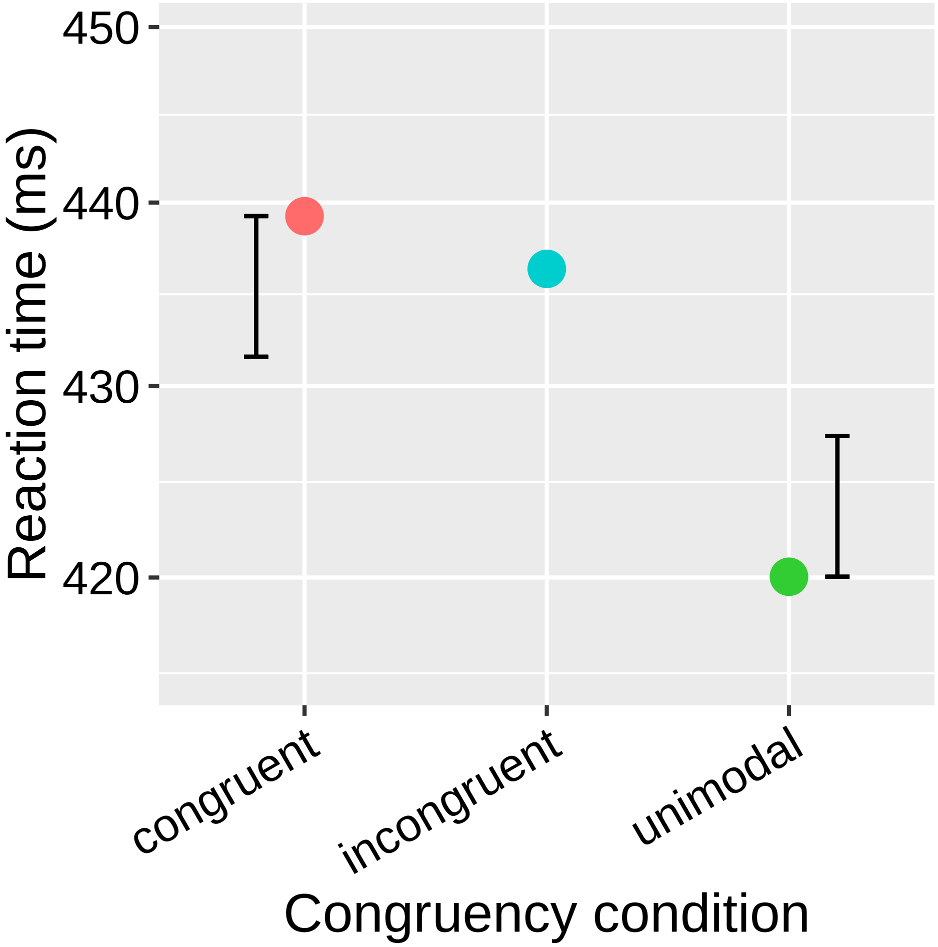


Figure 45. *Sharp Study 1a*: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars) when *sharpness* is the relevant dimension. Participants were significantly *slower* to respond to any sound (congruent or incongruent) compared to no sound.

Significant findings. We found no evidence of a congruency effect collapsing across sharpness differences and task (Figure 45). Participants were significantly *faster* to respond in the no sound condition (−0.05, [−0.06, −0.03]) than in the congruent or incongruent conditions, which did not differ from each other (−0.01, [−0.02, 0.01]). There was also not a significant effect of sharpness pairing or task instructions.

Errors

Our poisson GLMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `sharpness pairing` (1-6), and `task` (respond to sharper vs. rounder shape). We also included the subject-by-subject variation in the intercept as a random effect. We compared an additive model including just the three fixed effects to a non-additive model including the fixed effects and their two- and three-way interactions. The AICc difference (ΔAICc) between the two models is 53.6, which is “very strong evidence” in favor of the non-additive model (Jeffreys, 1961).

There was not a significant difference between the logarithm of correct responses for the congruent and incongruent conditions ($p = 1$; $z = 0$) or for the congruent and unimodal conditions ($p = 0.85$; $z = -0.19$). The proportion of errors in the congruent, incongruent, and unimodal conditions (responding to rounder shapes with sharpness pair 1) was 0.007, 0.007, and 0.012, respectively. Because there is no difference in errors across conditions, we can rule out a speed-accuracy tradeoff.

Conclusions

We were unable to replicate previous results finding an automatic association between pitch and sharpness when the visual dimension was relevant.

Sharp Study 1b: Pitch-Relevant Replication

Sharp Study 1b repeated Study 1a with pitch rather than sharpness as the relevant dimension.

Method

Participants

The same thirty-four U.Va. undergraduate students participated in Sharp Study 1a and Sharp Study 1b in exchange for credit in an introductory psychology course.

Stimuli

Auditory and visual stimuli were the same as Sharp Study 1a.

Design

For Study 1b, each of the three pitch differences (large, octave, and M3) was presented four times in each of the seven sharpness difference conditions (1–6, and no shapes), giving rise to a total of 108 trials per block. Participants completed two blocks (in a random order): (a) Participants were asked to select whether the first or second pitch was *higher*; (b) participants were asked to select whether the first or second pitch was *lower*.

Procedure

The sequence of trials was the same as Sharp Study 1a except participants always responded to the height of the pitch. They were told that task-irrelevant shapes would only appear on some trials.

Data management

Participants. We removed two participants based on accuracy. The average overall accuracy of the remaining 32 participants was 93.7%.

Reaction Time. Using the correct responses only, we eliminated trials where responses that were faster than 50ms (0.04% removed). We next performed a Box-Cox analysis on the reaction time data; the result suggested an inverse square root transformation, but we used a log transformation for consistency. Finally, we removed outliers that were more than three MADs from the median RT (6.9% removed).

Results and Discussion

Reaction Time

Table 32
Sharp Study 1b*: Comparison of two models predicting reaction time, ordered by* Δ*AICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Additive`	`Intercept`	7	0.0	1.00	0.050	0.401
`Non-additive`	`Intercept`	14	51.7	0.00	0.050	0.401

Model comparison. Our LMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `pitch difference` (continuous), and `task` (respond to higher pitch vs. lower pitch). We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the three fixed effects to a non-additive model including the three fixed effects and all of their two and three-way interactions. Table 32 shows that the additive model is superior to the non-additive model. The AICc difference (ΔAICc) between the two models is 51.7; "very strong evidence" in favor of the better model (Jeffreys, 1961).

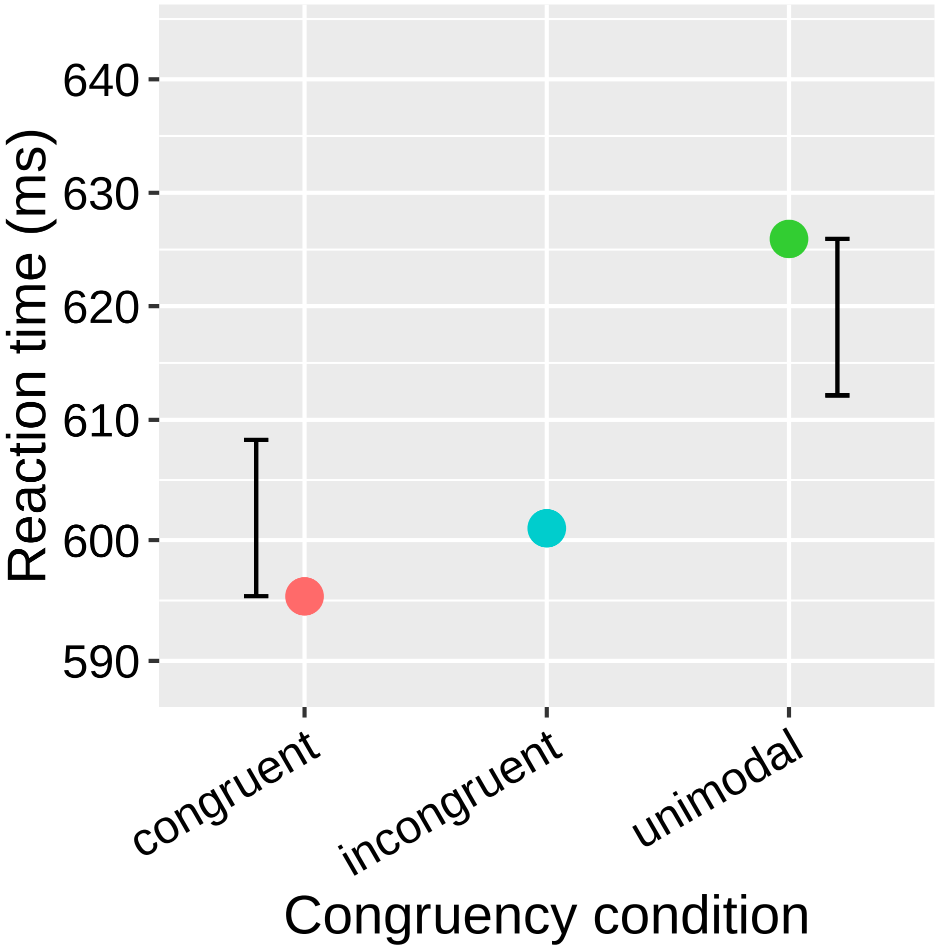


Figure 46. Sharp Study 1b: Effect plot showing the main effect of perceptual congruency (with upper and lower LSD bars) when *pitch* is the relevant dimension. Participants were *faster* to respond to either bimodal condition (congruent or incongruent) compared to the no shape condition.

Significant findings. We found no evidence of a congruency effect collapsing across pitch differences and task (Figure 46). Participants were significantly *slower* to respond in the no sound condition (0.05, [0.03, 0.07]) than either of the sound conditions (congruent and incongruent), which did not differ from each other (0.01, [−0.01, 0.03]).

Additionally, there was a significant effect of pitch difference, such that participants were slower to respond to a smaller pitch difference (0.12, [0.11, 0.14]), which shows that participants were completing the experiment as instructed.

Errors

Our binomial GLMMS included three fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent vs. unimodal), `task` (respond to higher pitch vs. lower pitch), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect. Because of the low error probability overall, a non-additive model failed to converge. We chose to fit an additive model with two categorical fixed effects and pitch difference as a continuous predictor since this model was clearly superior for the RT analysis as well.

There was a significant difference between the log-odds of correct responses for the congruent and incongruent conditions ($p \approx 0$; $z = 4.52$) and a marginal difference between the congruent and unimodal conditions ($p = 0.08$; $z = 1.74$). The proportion of errors in the congruent, incongruent, and unimodal conditions (responding to higher pitches with the average pitch difference) was 0.026, 0.045, and 0.032, respectively. Because the fastest condition (as shown in Figure 46) also has the least number of errors, we can rule out a speed-accuracy tradeoff.

Conclusions

We were unable to replicate previous results finding an automatic association between pitch and sharpness when the auditory dimension was relevant.

Sharp Study 2: Top-down Influence

Sharp Study 1a and 1b showed a failure to replicate the pitch–sharpness correspondence. Next, we sought to investigate the top-down influence of task instructions on performance. For Sharp Study 2, we used a within-subjects design where each participants completed one block each of four pairing instructions.

Method

Participants

Thirty-eight U.Va. undergraduate students participated in exchange for credit in an introductory psychology course.

Stimuli

The auditory and visual stimuli were the same as Sharp Studies 1a and 1b.

Design

We used a within-participant design in which each participant completed four blocks of 72 trials. Participants completed the four sets of experiment instructions in a random order: rounder shape/lower pitch, rounder shape/higher pitcher, sharper shape/higher pitch, or sharper shape/lower pitch.

Procedure

At the start of each trial, the instructions were presented on the screen. Participants pressed the `SPACEBAR` to proceed with the trial. The first shape appeared for 300ms; it was accompanied by a 300ms tone. This was followed by a 500ms blank screen. Then the second shape appeared for 300ms; it was also accompanied by a 300ms tone. After both pitch/sharpness pairings, either an ‘s’ for shape or ‘p’ for pitch appeared on the screen to cue the participant to which modality to respond. Participants indicated whether the first or second stimulus met the instruction criteria (e.g., for rounder/lower instructions, participants would respond to the *lower* pitch if a ‘p’ appeared or the *rounder* shape if an ‘s’ appeared). The cue remained on the screen until the participant responded, at which point the instruction screen reappeared to begin the next trial.

Data management

Participants. We removed one participant who completed fewer than 60% accuracy of the total trials. The average overall accuracy of the remaining 37 participants was 87.8%.

Reaction Time. After eliminating trials with incorrect responses, we next eliminated responses that were faster than 50ms (2.5% of responses). We then performed a Box-Cox analysis on the reaction time (RT) data based on an additive model of the five fixed effect predictors; the result suggested a logarithmic transformation. Finally, we removed outliers that were more than three MADs from the median RT, which removed an additional 5.1% of responses.

Table 33
Sharp Study 2*: Comparison of two models predicting reaction time, ordered by ΔAICc.*

fixed effect	random effect	K	ΔAICc	weight	$R^2_{\text{marg.}}$	$R^2_{\text{cond.}}$
`Non-additive`	`Intercept`	18	0.0	1.00	0.037	0.429
`Additive`	`Intercept`	7	213.4	0.00	0.017	0.413

Results and Discussion

Reaction Time

Model comparison. Our LMMS included four fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch or responding to shape), and `pitch difference` (continuous).[12] We also included the subject-by-subject variation in the intercept as a random effect.

We compared an additive model including just the four fixed effects to a non-additive model including the four fixed effects and all of their two-, three-, and four-way interactions. Table 33 shows that the non-additive model is clearly superior to the additive-only model. The AICc difference (ΔAICc) between the two models is 213.4, which is "decisive evidence" in favor of the better model (Jeffreys, 1961).

Significant findings. Figure 47 shows the predictors of logRT and their 95% confidence intervals. The `congruency x compatibility` interaction is the most important.

Figure 48 shows that instructions and perceptual congruency both impacted the results. Although there is a clear *congruency* advantage on compatible trials (0.27, [0.22, 0.31]), there is also a significant *incongruency* advantage on incompatible trials (−0.16, [−0.2, −0.12]). Further, the figure shows that participants are significantly slower to respond to the incongruent incompatible trials than congruent compatible trials, which means that it takes longer overall to pair together the perceptually incongruent endpoints.

Additionally, a three-way interaction model with `congruency` × `compatibility` and `trial number` as fixed effects shows that the RT difference between congruent and incongruent trials does not change as a function of trial number (1.69, [−2.19, 5.57]).

Errors

Model comparison. Our binomial GLMMS included four fixed effects: `congruency` (perceptually congruent vs. perceptually incongruent), `instruction compatibility` (compatible vs. incompatible), `modality` (responding to pitch vs. responding to shape), and `pitch difference` (continuous). We also included the subject-by-subject variation in the intercept as a random effect.

Because of the low error probability overall, a non-additive model including the five fixed effects and all of their two-, three-, and four-way interactions failed to converge. Instead, we compared a non-additive model including the four fixed effects and the `congruency` × `compatibility` two-way interaction to an additive model including just the four fixed effects. The AICc difference (ΔAICc) between the two models is 351.7, which is "decisive evidence" in favor of the non-additive model (Jeffreys, 1961).

[12] I did not include sharpness pairing as a factor because there was no particular difference ordering between the various shapes used (similar to the blobs in the size studies) and Sharp Study 1a found no RT difference across pairings.

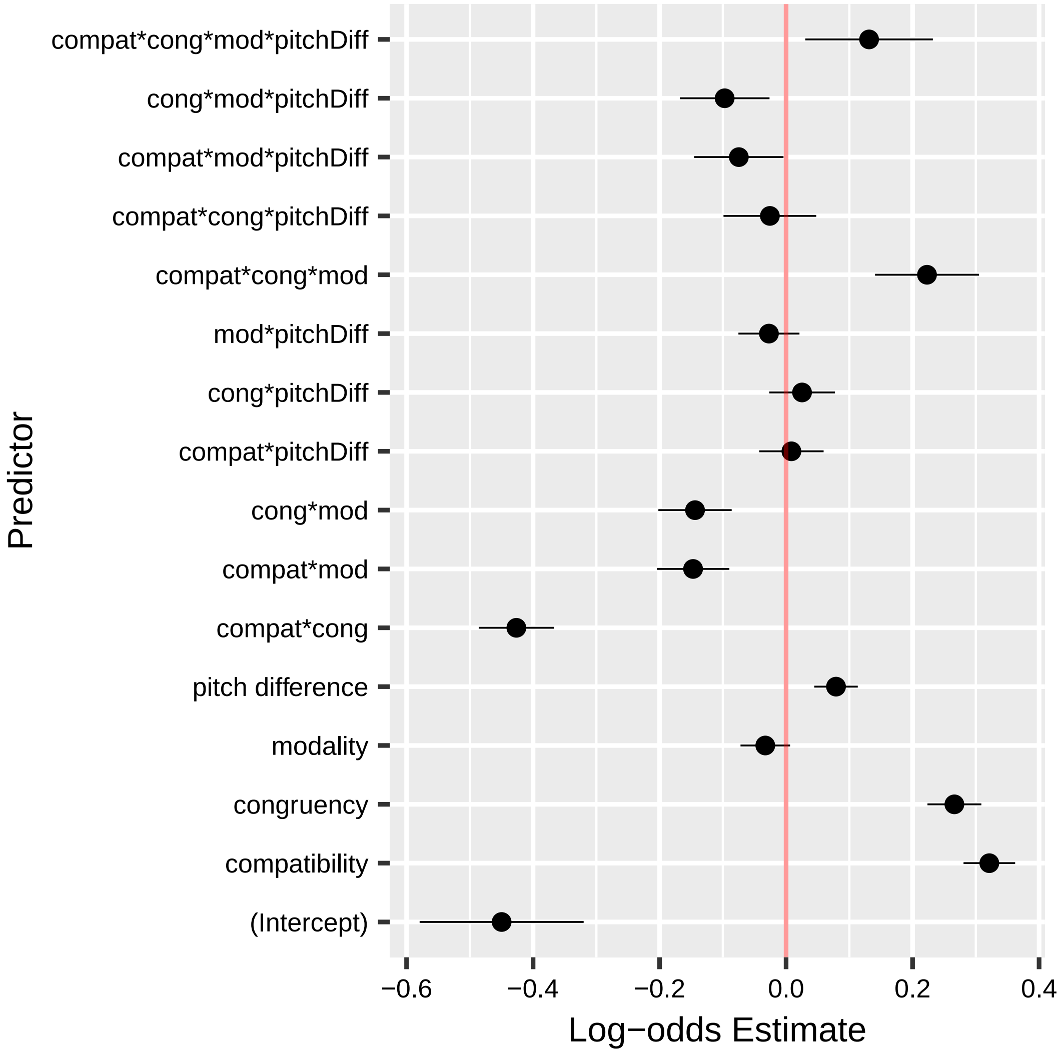


Figure 47. *Sharp Study 2*: Coefficient plot for the four fixed effect predictors and their interactions (with 95% confidence intervals) Compat = compatibility; Cong = congruency; Mod = modality; pitchDiff = pitch difference.

Significant findings. In the congruent condition, the probability that there is no difference between the log-odds of correct responses for compatible and incompatible blocks is $p \approx 0$ ($z = 14.18$), meaning listeners make more errors on incompatible blocks. In the compatible condition, the probability that there is no difference between the log-odds of correct responses for congruent and incongruent trials is $p \approx 0$ ($z = 15.58$), meaning listeners make a more errors on incongruent trials. These effects are converted to proportions and shown in Table 34. The error rates follow the same pattern as the reaction times (shown in Figure 48), meaning there is no evidence of a speed-accuracy tradeoff.

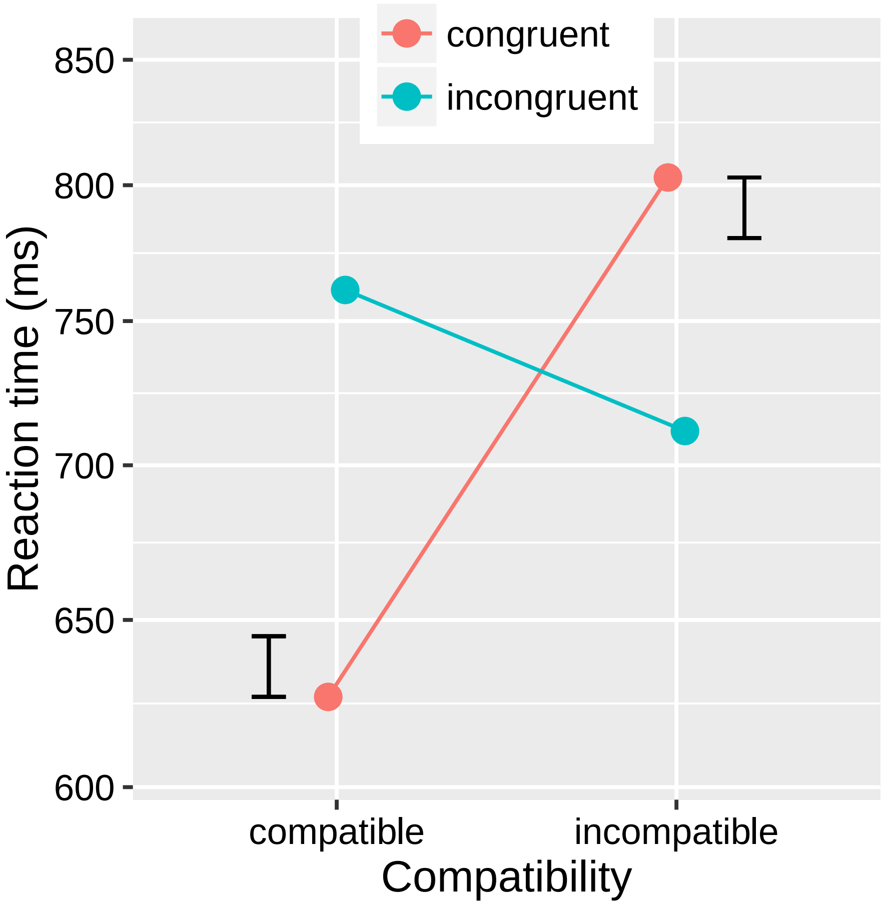


Figure 48. *Sharp Study 2*: Effect plot showing the interaction between perceptual congruency and instructions compatibility (with upper and lower LSD bars). Whatever instructions are given on that block, participants are faster to respond when those dimensions are paired together.

Table 34

Sharp Study 2*: Proportion of errors across compatibility and congruency conditions.*

	Compatible blocks	Incompatible blocks
Congruent trials	0.05	0.19
Incongruent trials	0.19	0.08

Conclusions

As with pitch–spatial frequency and pitch–brightness, the results here are a hybrid of the pitch–size and pitch–height results. We found top-down influence of instructions on performance with the pitch–sharpness correspondence. However, the congruency to incongruency advantage reversal was not as extreme as in the pitch–size correspondence, showing that there is some evidence for an automatic association between pitch and sharpness. Despite this, differing instructions seem to be mainly driving the results.

Pitch–Sharpness Discussion

Sharp Studies 1a and 1b failed to provide evidence for an automatic association between pitch and sharpness. Although the unimodal condition reversed from being slower or faster than the bimodal conditions, the congruent and incongruent conditions never differed in overall reaction time. In Sharp Study 2, the results were a hybrid between pitch–size (as in Figure 3b) and pitch–height (as in Figure 3c). There was a `congruency` × `compatibility` interaction, showing that task instructions impacted performance. There was also some evidence that participants had a harder time grouping perceptually congruent endpoints even when called to in the instructions. From these results, we conclude that there is some evidence for an automatic congruency relationship between pitch and sharpness at a perceptual level. However, most of the evidence points to top-down, decision level influences on pitch–sharpness pairings.

Chapter 7: General Discussion

Recently it has been proposed that our knowledge of cross-modal correspondences may be a useful solution to the problem of deciding when and how to bind information from our different senses to form a single muiltimodal percept (Spence, 2011). If individuals have a consistent mapping of sensory features across modalities, it can guide them to decide when inputs across modalities should be combined.

In this dissertation, I conducted several experiments to determine the **replicability** of correspondences between auditory pitch and visual dimensions of size (Size Studies 5-7), height (Height Studies 1a & 1b), spatial frequency (SF Studies 1a & 1b), brightness (Bright Studies 1a & 1b), and sharpness (Sharp Studies 1a & 1b). I failed to replicate seven out of ten correspondences. Because the matching of features in most cases was *not* automatic and *not* stable across slight changes in methodology, audiovisual correspondences may *not* be a reliable solution to the binding problem.

Another current debate in visual perception is the extent to which perception is "cognitively impenetrable" to higher-order cognition. Some argue that basic perception cannot operate outside of cognitive influence (Clark, 2013; Collins & Olson, 2014; Goldstone et al., 2015; Vetter & Newen, 2014) while others contend that there is not enough evidence to show perception changes under the influence of motivation, emotion, attention, or memory processes (Firestone & Scholl, in press).

To determine whether audiovisual correspondences are subject to **top-down cognitive influences**, we asked participants to pair dimensions in a way incongruent with "natural" mappings based on environmental correlations or language knowledge (Size Studies 1-4, Height Studies 2-4, SF Studies 2-3, Bright Study 2, Sharp Study 2). I found that differing instructions changed response speed, which supports previous research finding that higher-order cognitive processes can influence basic perception.

Replicability

I found that audiovisual correspondences do not rely *solely* on bottom-up processing, and in many cases, may rely *entirely* on top-down processing. Despite a number of previous studies showing a congruency advantage using a speeded classification paradigm (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010; Gallace & Spence, 2006; Marks, 1987a; Patching & Quinlan, 2002), I failed to replicate seven out of ten correspondences. Table 35 provides a summary of the conceptual and direct replication experiments.

There are three reasons to believe this is a real failure to replicate rather than a problem with our methodology. *First*, when there is a strong correspondence to be detected, our method is capable of detecting a significant effect. There was a significant congruency effect for three of the correspondences; the height of auditory pitch stimuli affected the classification of visual heights, the height of visual stimuli affected the classification of auditory pitch, and the brightness of visual stimuli affected the classification of auditory pitch. All of the other experiments had a similar number of participants and thus similar power to detect a significant effect. *Second*, manipulation checks found that participants were completing the task as requested and not randomly selecting a response. In each experiment the main effect of pitch difference or visual difference was significant. Larger differences between pitch and visual stimuli resulted in faster responses compared to smaller differences. *Third*, there was little evidence of a speed-accuracy tradeoff. Participants made few errors overall, and when they did, they normally made more errors on slower conditions.

Table 35
Chapters 2–6*: Replicability summary. Only three correspondences showed a significant congruency effect: the height of auditory pitch stimuli affected the classification of visual heights, the height of visual stimuli affected the classification of auditory pitch, and the brightness of visual stimuli affected the classification of auditory pitch.*

Correspondence	**Relevant Dimension**	
	Visual	**Auditory**
Pitch–size	failure to replicate [Size Study 5, 6]	failure to replicate [Size Study 7]
Pitch–height	**congruency effect** [Height Study 1a]	**congruency effect** [Height Study 1b]
Pitch–spatial frequency	failure to replicate [SF Study 1a]	failure to replicate [SF Study 1b]
Pitch–brightness	failure to replicate [Bright Study 1a]	**congruency effect** [Bright Study 1b]
Pitch–sharpness	failure to replicate [Sharp Study 1a]	failure to replicate [Sharp Study 1b]

The distinction between unimodal and bimodal conditions remains unclear. In most of the visual-relevant studies (i.e., size [conceptual replication], spatial frequency, brightness, sharpness), the unimodal condition was *faster* than the two bimodal conditions, which did not differ from each other. In most of the pitch-relevant conditions (i.e., spatial frequency, sharpness), the unimodal condition was *slower* than the two bimodal conditions, which did not differ from each other. There was also an advantage to the bimodal conditions in the size-relevant direct replication, which replicates Gallace and Spence (2006). It is unclear why in some cases the redundant modality seems to be distracting whereas in other cases it facilitates reaction time.

In the spatial frequency and bright studies, the congruency advantage reversed depending on the instructions used in the visual-relevant condition. When asked to select the narrower shape, participants were faster to pair narrow shapes and high pitches, but when asked to select the wider shape, participants were faster to pair wider shapes with high pitches. When asked to select the brighter shape, participants were faster to pair dark shapes and high pitches, but when asked to select the darker shape, participants were faster in the unimodal condition than either of the bimodal conditions. The reason for this congruency reversal effect remains an open question, but nonetheless shows that there is no fixed pairing between pitch and spatial frequency or brightness at a perceptual level.

Finally, the replication results suggest a bias towards visual processing in general, with the visual dimension seemingly harder to ignore than pitch when it is irrelevant to the task. In the height and bright studies, there was a stronger congruency effect in the pitch-relevant condition compared to the visual-relevant condition. This is in line with previous research that found auditory-relevant conditions to be more strongly affected by congruency than visual-relevant conditions (Ben-Artzi & Marks, 1995; Evans & Treisman, 2010).

Top-down Influence

As with the replicability results, the results here point to the fact that audiovisual correspondences either *jointly* rely on bottom-up and top-down processing or are *solely* the result of top-down effects. I found that differences in task instructions and lexical overlap changed participants' response times to the audiovisual mappings. These results show the varying degrees to which audiovisual correspondences are subject to the top-down influences.

Figure 49 provides a summary of the top-down influence experiments. Instead of the three alternatives predicted in Figure 3, our results produced three slightly different outcomes.

- Instructions (Figure 3b)
  - There is a congruency effect on compatible blocks (i.e., faster RT on `congruent` than `incongruent` trials)
  - There is an *equal* or *larger* incongruency effect on incompatible blocks (i.e., faster RT on `incongruent` than `congruent` trials)
  - There is *no* RT advantage for the compatible blocks overall (i.e., RT for the `congruent compatible` and the `incongruent incompatible` trials are the same)
- Hybrid
  - There is a congruency effect on compatible blocks
  - There is a *smaller* (but still significant) incongruency effect on incompatible blocks
  - There *is* an RT advantage for the compatible blocks overall (i.e., faster RT for the `congruent compatible` than the `incongruent incompatible` trials)
- Instructions + Congruency (Figure 3c)
  - There is a congruency effect on compatible blocks
  - There is *no* incongruency effect on incompatible blocks
  - There is an RT advantage for the compatible blocks overall

The strength of the metaphor used to describe auditory pitch gets stronger moving along the **Instructions** to **Hybrid** to **Instructions + Congruency** continuum. *Size* was only influenced by top-down processing and showed no evidence of an automatic, bottom-up association. In our culture, we never use the terms 'small' and 'large' to describe 'high' and 'low' pitches. At the other end of the continuum, *height* was the only metaphor that was automatic and stable across replications and was the most influenced by congruency in addition to top-down processes. 'High' and 'low' is the dominant metaphor we use when talking about pitch in our culture.

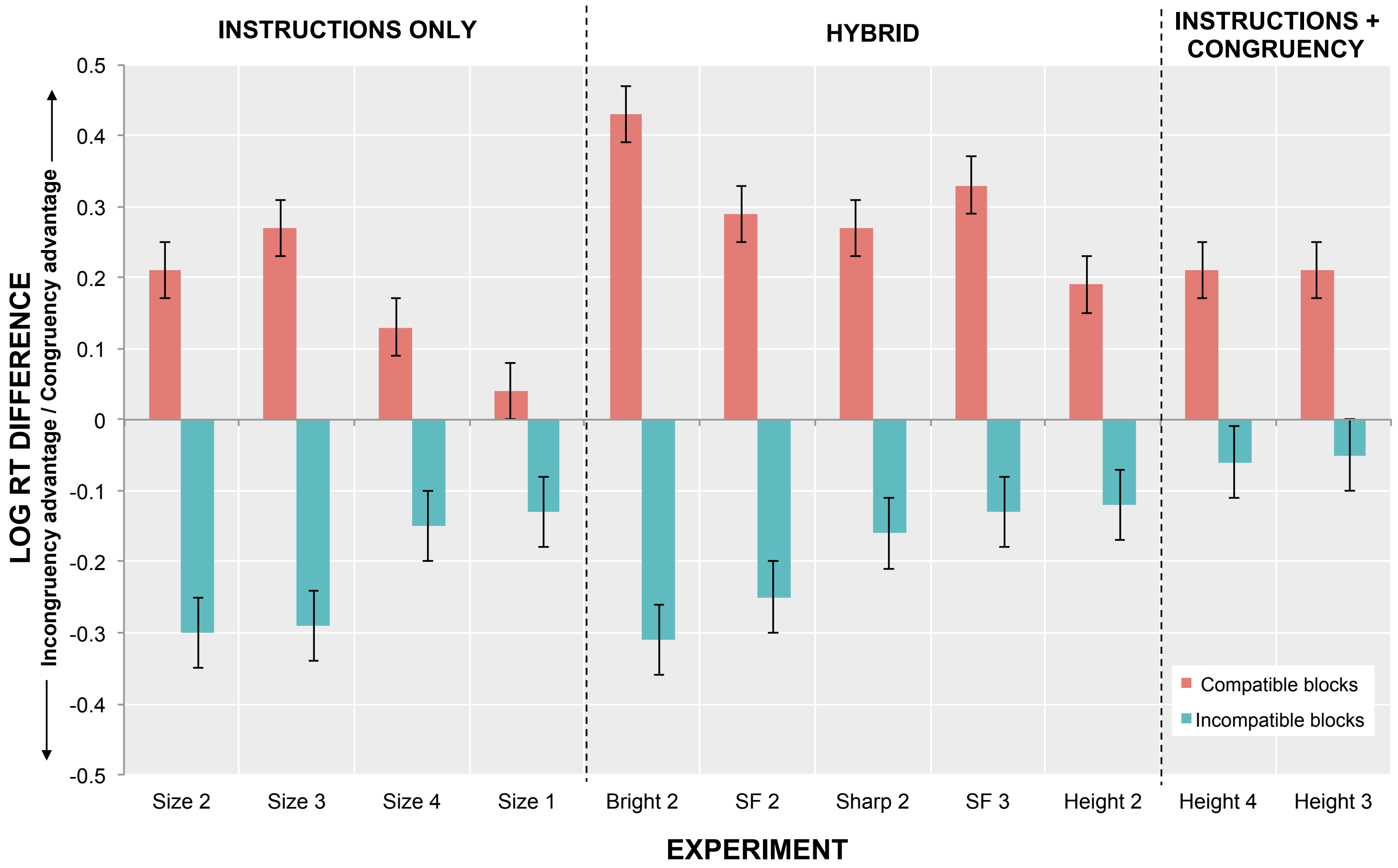


Figure 49. *Chapters 2–6*: Log RT difference (with 95% confidence intervals) between congruent and incongruent trials on compatible blocks (red) and incompatible blocks (blue). Positive values indicate a congruency advantage (congruent faster than incongruent) and negative values indicate an incongruency advantage (incongruent faster than congruent). Experiments are ordered by the strength of the incongruency advantage relative to the congruency advantage. The strength of the metaphor used to describe auditory pitch is stronger moving along the **Instructions** to **Hybrid** to **Instructions + Congruency** continuum.

Though *spatial frequency* can also be described using the words 'high' and 'low', this tended to be an unfamiliar definition for our participants, who anecdotally reported that narrow and wide stripes made more intuitive sense than high and low spatial frequency. Also, spatial frequency may relate more to the physical property of auditory frequency (i.e., higher repetition rate = higher frequency) rather than our metaphorical understanding of pitch. Thus it makes sense that the results trend towards pitch–height but show more influence of instructions than height.

Sharpness and *brightness* are also metaphors we occasionally use to describe pitch (e.g., high sounds are described as 'sharp', 'shrill', and 'bright'; low sounds are described as 'dark', 'full' and 'round'), but these relate more to timbral than pure tone differences. Firm conclusions regarding the ordering of the middle correspondences on the continuum await future studies, but the speculative ordering I provide is in keeping with top-down influence of language processing.

Open Questions

The most basic open question that remains is how far this methodology can extend. All of the correspondences included in my dissertation involved audiovisual correspondences with pitch as the auditory dimension. Given the possible metaphorical relationships between other auditory and visual dimensions, future research should extend the visual correspondences used here to auditory loudness and timbre. Beyond audiovisual correspondences, it will also be of interest to look at correspondences between other modalities, such as tastes/flavors and sounds (Crisinel & Spence, 2010; Gallace, Boschin, & Spence, 2011; Mesz, Trevisan, & Sigman, 2011; Simner, Cuskley, & Kirby, 2010), odors and shapes/colors (Gilbert, Martin, & Kemp, 1996; Kemp & Gilbert, 1997; Seo et al., 2010), and vision and touch (Martino & Marks, 2000; Morgan, Goodson, & Jones, 1975; P. Walker, Francis, & Walker, 2010), just to name a few.

A second open question relates to ecological validity. Because the majority of my studies used sine tones and circles or randomly-generated shapes presented on a computer screen, it is possible that the modalities were not tightly bound because they are not representative of sounds or shapes in the environment (De Gelder & Bertelson, 2003). Thus future research using pairings that are more environmentally-valid, such as animal or instrument sounds/pictures, may find that the pairings are harder to reverse.

Third, as noted in the introduction, by investigating replicability and top-down influences on a number of audiovisual correspondences, I hoped to offer a principled way to categorize cross-modal correspondences as an alternative to Spence's (2011) structural, statistical, and semantic categories. This is necessary because of the lack of distinction between innate vs. learned correspondences and between perceptual vs. decisional consequences and problems regarding potential overlapping and missing categories. Despite my proposed addition of a *metaphorical* correspondence (similar to P. Walker's (2012) suggestion of a semantic correspondence based on connotative meaning rather than lexical overalp), an alternative classification is unfortunately not obvious given the results from my five audiovisual correspondences. Though an alternative classification remains for future research, I echo Spence's argument from his review:

> To make further progress ... I would argue that researchers will need to make a much clearer distinction between the various kinds of crossmodal correspondences that have been reported to date. This may be especially important given that they may reflect the influence of different underlying neural substrates, and may even have qualitatively different effects on human perception and performance (Spence, 2011, p.988).

Fourth, although I used a paradigm updated from previous studies, it is still an extension of the speeded classification paradigm. Therefore, it is an open question whether top-down influences would be evident in other tasks as well. For example, one could investigate memory for object features involving either congruent or incongruent pairings. Another task may be to investigate either implicit preference (using eye tracking looking time) or explicit preference (using rating scales) for congruent vs. incongruent pairings, similar to methods used in previous studies of preference for unimodal features (Palumbo & Bertamini, 2016; Palumbo, Ruta, & Bertamini, 2015). Converging operations are important to validate the results found here using speeded classification.

Another open question that relates to bottom-up vs. top-down processing relates to neural activity. I have argued that audiovisual correspondences occur at a later, decision level of processing rather than an early, purely perceptual level. One way to validate this claim would be to investigate the timing of the effect using electroencephalogram (EEG). Though studies have looked at congruency effect timing in other correspondences (e.g., Rampone, Makin, & Bertamini, 2014), no studies I am aware of have used EEG to assess effect timing of audiovisual correspondences. Doing so will allow us to see whether differences exist between congruent and incongruent trials and at what stage of processing the potential overlap between modalities occurs.

Finally, to address the question of different correspondence types, more developmental and cross-cultural work needs to be done to address distinctions between innate vs. learned correspondences and perceptual vs. decisional consequences. A plethora of cultural work has been done with the pitch–height correspondence (Ashley, 2004; Dolscheid et al., 2013; Eitan & Timmers, 2010; Parkinson et al., 2012), but similar studies are necessary for other correspondences. Methods to investigate correspondences with young infants are also needed (e.g., Lewkowicz & Turkewitz, 1980). The combination of neural, cultural, and developmental studies will help create a better categorization scheme for cross-modal correspondences.

Why the Differences from Previous Results?

My dissertation strongly questions the assumption of automaticity prevalent in the cross-modal correspondence literature. I would argue that audiovisual correspondences are a consequence of later decision-level influences rather than being truly automatic perceptual effects.

The difference between my findings and previous research could be *methodological*: audiovisual correspondences may exist under very specific conditions only. However, this would not explain my failure to directly replicate the pitch–size correspondence. Also, if true, this would mean cross-modal correspondences are irrelevant and uninteresting if they do not stand up to changes in methodology.

Second, the difference could be *analytical*: my stricter analysis approach may be the explanation for the lack of replication. I took a more robust approach than previous studies which I believe to be more accurate, including a reaction time transformation, removing outliers, and using mixed-effects models. However, although I found a main effect of congruency using repeated-measures ANOVA on the direct replication (Size Study 6) of Gallace and Spence (2006), post-hoc analyses revealed no significant difference between the congruent and incongruent conditions. It is therefore unlikely that the failure to replicate is due solely to my analysis approach.

Third, the difference could be *theoretical*: I empirically tested for top-down influences rather than assuming a bottom-up association. It is possible that there is a slight natural advantage to the pairings that exist in the environment which would result in a tenuous effect that sometimes fails to replicate. Further, when specifically asked to pair the modalities in the opposite direction, the natural advantage requires very little rewiring or relearning, which would explain the congruency reversal in most of the correspondences. The only time relearning would be required to override the natural inclination is with a strong metaphor such as pitch and height.

Though my results are at odds with the majority of previous research in this field, several recent findings support the role of top-down processing in sensory integration more generally. Research on perceptual decision-making has shown that evidence from sensory domains is accumulated separately in redundant conditions and flexibly coupled together at a later stage of processing (Otto, Dassy, & Mamassian, 2013; Otto & Mamassian, 2012). Research on individual differences in spatial and temporal integration has shown that binding is stable across time but seems to be task specific, which argues against a single, global parameter that is responsible for sensory integration (Odegaard & Shams, 2016). Even Spence's (2007) review of audiovisual multisensory integration notes that "it is not always that easy to distinguish between the contribution of structural (or bottom-up) and cognitive (or more top-down) factors to multisensory integration" (p. 68). Clearly then, an important avenue for continued research is to understand the interaction between bottom-up and top-down processes in audiovisual correspondences.

References

Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, *66*(3), 912–918.

Ashley, R. (2004). Musical pitch space across modalities: Spatial and other mappings through language and culture. In S. D. Lipscomb, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the 8th International Conference on Music Perception and Cognition* (p. 64-72).

Atkinson, J. E. (1978). Correlation analysis of the physiological factors controlling fundamental voice frequency. *Journal of the Acoustic Society of America*, *63*, 211-222.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi: 10.1016/j.jml.2007.12.005

Bartoń, K. (2014). `MuMIn`: Multi-model inference [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=MuMIn (R package version 1.10.5)

Bates, D., Maechler, M., & Bolker, B. (2014). `lme4`: Linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 1.1-7)

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*, 1151-1162.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289–300.

Bond, B., & Stevens, S. (1969). Cross-modality matching of brightness to loudness by 5-year-olds. *Perception & Psychophysics*, *6*, 337-339.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*, 211-252.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological Science*, *9*, 211–214.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, UK: Cambridge University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204.

Cohen Kadosh, R., Cohen Kadosh, K., & Henik, A. (2008). When brightness counts: The neuronal correlate of numerical–luminance interference. *Cerebral Cortex*, *18*(2), 337-343.

Cohen Kadosh, R., & Henik, A. (2006). A common representation for semantic and physical properties: A cognitive-anatomical approach. *Experimental Psychology*, *53*(6), 87-94.

Collins, J. A., & Olson, I. R. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic Bulletin & Review*, *21*, 843–860.

Coward, S. W., & Stevens, C. J. (2004). Extracting meaning from sound: Nomic mappings, everyday listening, and perceiving object size from frequency. *Psychological Record*, *54*, 349–364.

Crisinel, A.-S., & Spence, C. (2010). As bitter as a trombone: Synesthetic correspondences in nonsynesthetes between tastes/ flavors and musical notes. *Attention, Perception, & Psychophysics*, *72*, 1994–2002.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences*, *7*(10), 460-467.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Pychological Science*, *24*(5), 613-621.

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*, 405-422.

Esterman, M., Verstynen, T., Ivry, R. B., & Robertson, L. C. (2006). Coming unbound: Disrupting automatic integration of synesthetic color and graphemes by tms of the right parietal lobe. *Journal of Cognitive Neuroscience*, *18*, 1570–1576.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6: 1-12.

Firestone, C., & Scholl, B. J. (in press). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences*.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*(15), 1-27.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second ed.). Thousand Oaks CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gallace, A., Boschin, E., & Spence, C. (2011). On the taste of "bouba" and "kiki": An exploration of word–food associations in neurologically normal participants. *Cognitive Neuroscience*, *5*(1), 34-46.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*, 1191-1203.

Gebuis, T., & van der Smagt, M. J. (2011). Incongruence in number–luminance congruency effects. *Attention, Perception, & Psychophysics*, *73*, 259-265.

Gilbert, A. N., Martin, R., & Kemp, S. E. (1996). Cross-modal correspondence between vision and olfaction: The color of smells. *American Journal of Psychology*, *109*, 335–351.

Goldstone, R. L., de Leeuw, J. R., & Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, *135*, 24-29.

Grassi, M. (2005). Do we hear size or sound? Balls dropped on plates. *Perception & Psychophysics*, *67*, 274-284.

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *American Journal of Psychology*, *109*, 219-238.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford Universaity Press.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R^2 GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946.

Kemp, S. E., & Gilbert, A. N. (1997). Odor intensity and color lightness are correlated sensory dimensions. *American Journal of Psychology*, *110*, 35–46.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016). lmertest: Tests in linear mixed effects models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=lmerTest (R package version 2.0-30)

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*, 278-292.

Lakoff, G., & Johnson, M. (1980a). The metaphorical structure of the human conceptual system. *Cognitive Science*, *4*(2), 195-208.

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

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology*, *16*, 597-607.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764-766.

Luo, D., Ganesh, S., & Koolaard, J. (2014). predictmeans: Calculate predicted means for linear models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=predictmeans (R package version 0.99)

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *American Journal of Psychology*, *87*, 173-188.

Marks, L. E. (1978). *The unity of the senses: Interrelations among the modalities*. New York: Academic Press.

Marks, L. E. (1987a). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 384-394.

Marks, L. E., Ben-Artzi, E., & Lakatos, S. (2003). Cross-modal interactions in auditory and visual discrimination. *International Journal of Psychophysiology*, *50*, 125-145.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*, 903-923.

Martino, G., & Marks, L. E. (2000). Cross-modal interaction between vision and touch: The role of synesthetic correspondence. *Perception*, *29*, 745–754.

MATLAB. (2013). *version 8.2.0 (r2013b)*. Natick, Massachusetts: The MathWorks Inc.

Maurer, D., Gibson, L. C., & Spector, F. (2012). Infant synaesthesia: New insights into the development of multisensory perception. In A. J. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory development* (pp. 229–250). Oxford, UK: Oxford University Press.

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, *116*, 323-336.

Mesz, B., Trevisan, M. A., & Sigman, M. (2011). The taste of music. *Perception*, *40*(6), 209-219.

Mondloch, C., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 133–136.

Morgan, G. A., Goodson, F. E., & Jones, T. (1975). Age differences in the associations between felt temperatures and color choices. *American Journal of Psychology*, *88*, 125–130.

Muggleton, N., Tsakanikos, E., Walsh, V., & Ward, J. (2007). Disruption of synaesthesia following TMS of the right posterior parietal cortex. *Neuropsychologia*, *45*, 1582-1585.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

O'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.

Odegaard, B., & Shams, L. (2016). The brain's tendency to bind audiovisual signals is stable but not general. *Psychological Science*.

Ohala, J. J. (1997). Sound symbolism. In *Proceedings of the 4th Seoul International Conference on Linguistics [SICOL]* (p. 98-103).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716: 1-8.

Otto, T. U., Dassy, B., & Mamassian, P. (2013). Principles of multisensory behavior. *Journal of Neuroscience*, *33*(17), 7463-7474.

Otto, T. U., & Mamassian, P. (2012). Noise and correlations in parallel perceptual decision making. *Current Biology*, *22*(15), 1391-1396.

Palumbo, L., & Bertamini, M. (2016). The curvature effect: A comparison between preference tasks. *Empirical Studies of the Arts*, *34*, 35-52.

Palumbo, L., Ruta, N., & Bertamini, M. (2015). Comparing angular and curved shapes in terms of implicit associations and approach/avoidance responses. *PLoS ONE*, *10*(10), e0140043.

Parise, C. V., & Spence, C. (2009). When birds of a feather flock together: Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE*, *4*, e5664.

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*, 854-861.

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*, 755-775.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2016). nlme: Linear and nonlinear mixed effects models [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=nlme` (R package version 3.1-125)

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Ramachandran, V. S., & Hubbard, E. M. (2001). Synesthesia—a window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3-34.

Rampone, G., Makin, A. D. J., & Bertamini, M. (2014). Electrophysiological analysis of the affective congruence between pattern regularity and word valence. *Neuropsychologia*, *58*, 107-117.

Roffler, S. K., & Butler, R. A. (1968). Localization of tonal stimuli in the vertical plane. *The Journal of the Acoustical Society of America*, *43*, 1260-1266.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology,*, *12*, 225–239.

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1791-1810.

Seo, H. S., Arshamian, A., Schemmer, K., Scheer, I., Sander, T., Ritter, G., & Hummel, T. (2010). Cross-modal integration between odors and abstract symbols. *Neuroscience Letters*, *478*(3), 175-178.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Simner, J., Cuskley, C., & Kirby, S. (2010). What sound does that taste? Cross-modal mapping across gustation and audition. *Perception*, *39*, 553–569.

Spector, F., & Maurer, D. (2009). Synesthesia: A new approach to understanding the development of perception. *Developmental Psychologysychology*, *45*, 175-189.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science & Technology*, *28*, 61–70.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*, 971-995.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99-114.

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, *27*, 62–75.

Walker, P. (2012). Cross-sensory correspondences and cross talk between dimensions of connotative meaning: Visual angularity is hard, high-pitched, and bright. *Attention, Perception, & Psychophysics*, *74*(8), 1792-1809.

Walker, P., Francis, B. J., & Walker, L. (2010). The brightness-weight illusion: Darker objects look heavier but feel lighter. *Experimental Psychology*, *57*(6), 462-469.

Walker, R. (1985). Mental imagery and musical concepts: Some evidence from the congenitally blind. *Bulletin of the Council for Research in Music Education*, *85*, 229-237.

Walker, R. (1987). The effects of culture, environment, age, and music training on choices of visual metaphors for sound. *Perception & Psychophysics*, *42*, 491-502.

Walsh, V. (2003). A theory of magnitude: Common cortical metrices of time, space and quality. *Trends in Cognitive Sciences*, *7*, 483–488.

Ward, J., Huckstep, B., & Tsakanikos, E. (2006). Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all? *Cortex*, *42*, 264-280.