Cognition 175 (2018) 101–108

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Questioning the automaticity of audiovisual correspondences

Laura M. Getz[a,b,*], Michael Kubovy[a]

[a] University of Virginia, Department of Psychology, United States
[b] Villanova University, Psychological and Brain Sciences Department, United States

ARTICLE INFO

Keywords:
Cross-modal correspondence
Audiovisual correspondence
Automaticity
Bottom-up processing
Top-down processing

ABSTRACT

An audiovisual correspondence (AVC) refers to an observer's seemingly arbitrary yet consistent matching of sensory features across the two modalities; for example, between an auditory pitch and visual size. Research on AVCs has frequently used a speeded classification procedure in which participants are asked to rapidly classify an image when it is either accompanied by a congruent or an incongruent sound (or vice versa). When, as is typically the case, classification is faster in the presence of a congruent stimulus, researchers have inferred that the AVC is automatic and bottom-up. Such an inference is incomplete because the procedure does not show that the AVC is *not* subject to top-down influences. To remedy this problem, we devised a procedure that allows us to assess the degree of "bottom-up-ness" and "top-down-ness" in the processing of an AVC. We did this in studies of AVCs between pitch and five visual features: size, height, spatial frequency, brightness, and angularity. We find that all the AVCs we studied involve *both* bottom-up *and* top-down processing, thus undermining the prevalent generalization that AVCs are automatic.

1. Introduction

Cross-modal correspondences refer to seemingly arbitrary yet consistent associations across sensory features from different sensory modalities (for reviews, see Marks, 2004; Parise, 2016; Spence, 2011). In the present paper, we focus our attention on *audiovisual* correspondences (AVCs). For example, it has been shown that people readily associate high-pitched tones with smaller objects placed higher in space. We attempt to address the issue of automaticity by creating separate measures of "bottom-up-ness" and "top-down-ness" in our assessment of AVCs between auditory *pitch* and five visual properties: *size*, *height*, *spatial frequency*, *angularity*, and *brightness*.

A majority of past research on AVCs has used a *speeded classification* paradigm. In such experiments, participants classify a multimodal stimulus according to its value on one modality while ignoring the other modality. For instance, they might be asked to report whether a stimulus was large or small while disregarding a concurrent high or low pitch. In this case, size is called the *relevant feature* and pitch is called the *irrelevant feature*. Participants encounter two main types of trials in a typical experiment: (a) on *congruent* trials, the level of the irrelevant feature matches the level of the relevant feature (a *low* pitch with a *large* stimulus); and (b) on *incongruent*trials, the level of the irrelevant feature does not match the level of the relevant feature (a *high* pitch with a *large* stimulus).[1]

Correctly classifying the relevant feature more quickly on congruent than on incongruent trials is treated as evidence that the irrelevant feature affects the processing of the relevant feature in a bottom-up fashion. For example, Evans and Treisman (2010) argue that audiovisual correspondences are "certainly automatic and independent of attention" (p. 10) and Gallace and Spence (2006) conclude that "people cannot help but process auditory information even when it is irrelevant to their visual task" (p. 1200). It is important to note that conclusions regarding automaticity are not limited to the speeded classification paradigm; for example, Parise and Spence (2012) used a speeded implicit association task to show that auditory and visual dimensions are paired together rapidly and automatically.

However, a congruency advantage alone is inadequate to imply a purely automatic, bottom-up effect for three reasons. First, contradictory evidence exists as to the replicability of the congruency advantage. A number of studies, including our own work (see S1 Motivating Experiments) and the work of other researchers (*e.g.*, Heron, Roach, Hanson, McGraw, & Whitaker, 2012; Klein, Brennan, & Gilani, 1987) have failed to show a congruency advantage on speeded detection tasks of various AVCs.

Second, the congruency advantage itself fails to show that AVCs are *immune to top-down influences*. Although the debate regarding top-down influences on perception has centered primarily on visual as opposed to cross-modal perception (*e.g.*, Firestone & Scholl, 2016; Goldstone, de Leeuw, & Landy, 2015; Vetter & Newen, 2014), there is evidence that factors such as the stimulus situation, modality characteristics, and observer

* Corresponding author at: Villanova University, Psychological and Brain Sciences Department, 800 E. Lancaster Ave., Villanova, PA 19085, United States.
E-mail address: laura.getz@villanova.edu (L.M. Getz).

[1] Some experiments, such as Gallace and Spence (2006) include a 'neutral' or unimodal condition as well, where no irrelevant stimulus value occurs (*e.g.*, no sound is played).

https://doi.org/10.1016/j.cognition.2018.02.015
Received 24 August 2017; Received in revised form 11 February 2018; Accepted 13 February 2018

Table 1
Consensus mapping for each audiovisual correspondence based on previous studies finding a significant congruency effect.

Visual dimension	High-pitch pairing	Low-pitch pairing	Previous experiments
Size	Small	Large	Evans and Treisman (2010), Gallace and Spence (2006), Mondloch and Maurer (2004), Spector and Maurer (2009)
Height (Elevation)	High	Low	Ben-Artzi and Marks (1995), Evans and Treisman (2010), Melara and O'Brien (1987), Patching and Quinlan (2002)
Spatial frequency	High (Narrow)	Low (Wide)	Evans and Treisman (2010)
Angularity (Sharpness)	Sharp	Rounded	Marks (1987), Maurer et al. (2012), O'Boyle and Tarte (1980), Parise and Spence (2009)
Brightness (Contrast)	Bright	Dark	Marks (1974, 1987), Martino and Marks (1999), Mondloch and Maurer (2004)

processes may affect multimodal perception as well (Chen & Spence, 2017; Welch & Warren, 1980). For example, Klapetek, Ngo, and Spence (2012) conclude that the pitch–brightness AVC operates "at a more strategic (*i.e.*, rather than at an automatic or involuntary) level" (p. 1161). Similarly, others argue that AVCs are influenced by cognitive processes rather than purely the result of perceptual encoding and contend that mappings across sensory cues are highly flexible based on prior experience (Chen & Spence, 2017; Chiou & Rich, 2012; Parise, 2016).

These seemingly contradictory conclusions point to the need to quantify the degree of automaticity in AVCs rather than choosing a side in the bottom-up vs. top-down debate (cf. Spence & Deroy, 2013). This relates to the third problem with previous research, which is that there is little consensus in the literature as to what *automaticity* really means (Moors & De Houwer, 2006; Santangelo & Spence, 2008). Though determining a theoretically and pragmatically appropriate definition is beyond the scope of this paper, we agree with previous researchers who argue that automaticity should be viewed as an umbrella term (*e.g.*, Spence & Deroy, 2013). In our work, we mean automaticity in terms of a bottom-up association between the auditory and visual modalities that exists without the necessity for intentional learning and outside the influence of attention or motivation. To that end, here we report the results of a new paradigm for assessing AVCs, which we see as a first step in answering what Spence and Deroy (2013) call "a challenge of the first order" (p. 257); namely, investigating the *degree* of "bottom-up-ness" and "top-down-ness" present in a variety of AVCs.

To achieve this goal, we created a modified version of the speeded classification task, where we manipulated the stimulus-response mapping included in the instructions to participants. This allowed us to determine whether participants could pair the corresponding dimensions in either direction without a loss in reaction time (*e.g.*, pairing high pitch with small shapes vs. pairing high pitch with large shapes). This is in line with previous work showing the importance of instructions given to participants in showing that AVC processing is at least partially goal-dependent (Chiou & Rich, 2012; Klapetek et al., 2012).

In our experiments, we jointly manipulated congruence and compatibility. We defined *congruence* according to the consensus mapping of pitch onto the visual property manipulated in that study (see Table 1). For example, in the case of the pitch–size correspondence, we consider small size to be congruent with high pitch and large size congruent with low pitch. We defined *compatibility* in reference to the instructions given on each block of trials: (a) during *compatible* blocks, the instructions pair congruent endpoints of the auditory and visual dimensions (*e.g.*, participants are told to select either the large shape/low pitch or small shape/high pitch), whereas (b) during *incompatible* blocks, the instructions are reversed and now pair incongruent endpoints (participants are told to select either the large shape/high pitch or small shape/lower pitch).

This procedure allowed us to create measures of "bottom-up-ness" (BU) and "top-down-ness" (TD) based on the participants' response speed to the various conditions. Fig. 1 shows several hypothetical outcomes for experiments using our methodology. "Bottom-up-ness" refers to the ease with which participants completed the task on compatible as opposed to incompatible blocks. Slower response speeds on incompatible blocks are evidence that it is hard to pair together the incongruent dimensions and thus show a stronger bottom-up association. Fig. 1a represents the case of a strong bottom-up effect with low top-down influence: participants are slower when given instructions asking them to pair the dimensions in the non-consensus direction on incompatible blocks. "Top-down-ness" refers to how well participants followed the instructions on compatible and incompatible blocks. If participants can just as quickly and accurately pair the dimensions in the opposite, non-consensus direction (*i.e.*, on incompatible blocks), this is evidence of a stronger top-down, goal-directed influence of the instructions. Fig. 1c represents the case of high top-down influence with little evidence of a bottom-up effect: the instructions to invert the association are followed with no cost in reaction time.

Fig. 1b represents an intermediate case on both the bottom-up and top-down dimensions. In these three cases, there is a congruency advantage on compatible blocks (showing a successful replication) and an incongruency advantage on incompatible blocks (showing a successful manipulation). Though less likely, it is not inevitable that the results will show a clear trade-off between bottom-up and top-down effects. Fig. 1d represents a case where the instructions have no effect (showing a failed manipulation): participants are always faster to respond to the congruent dimensions even when the instructions ask them to pair the dimensions in the opposite direction. Fig. 1e represents a case where the auditory and visual dimensions pair together more naturally in the *opposite* direction from what has traditionally been shown, thus showing a failure to replicate previous studies.

Having separate measures for bottom-up associations and top-down influence grants us a more direct way to *quantify* the degree of automaticity present in each correspondence, thus meaningfully adding to the debate on the cognitive penetrability of audiovisual perception.

2. Method

2.1. Participants

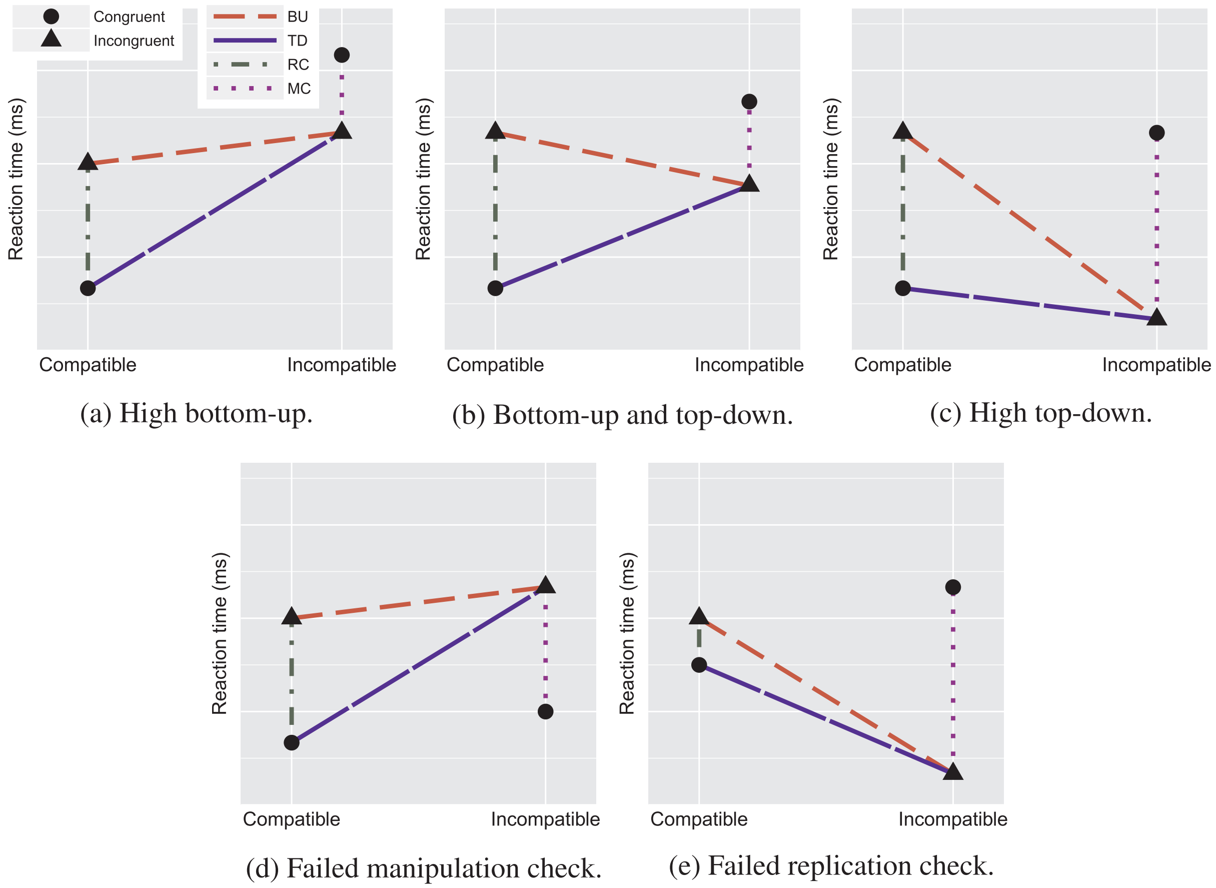
We recruited 179 University of Virginia undergraduates with normal or corrected-to-normal vision and normal hearing to participate in exchange for credit in an introductory psychology course ($n = 31$ for size; $n = 24$ for height; $n = 36$ for spatial frequency; $n = 38$ for angularity; $n = 50$ for brightness).

2.2. Stimuli

2.2.1. Auditory pitches

All sounds were sine tones with 10 ms rise and decay times. We used three frequency intervals: 'large' (300 Hz vs. 4500 Hz), 'octave' (440 Hz vs. 880 Hz), and 'M3' (a major third, 500 Hz vs. 630 Hz). The octave and M3 intervals were chosen to determine whether the effect previously found with the large interval generalized to smaller pitch differences.[2] We were not able to accurately measure the dB level of the sounds used, but they were manually adjusted to be equally loud across the various

[2] 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 only (Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002).



(a) High bottom-up.

(b) Bottom-up and top-down.

(c) High top-down.

(d) Failed manipulation check.

(e) Failed replication check.

Fig. 1. Hypothetical outcomes separating bottom-up (BU) and top-down (TD) influences, including the replication check (RC) and manipulation check (MC).

frequencies (by researcher LG prior to the experiments rather than individually by participant).

2.2.2. *Visual shapes*

Fig. 2 provides an example of the stimuli used to investigate each of the correspondences. We drew all shapes in white on a black background except for the spatial frequency stimuli (details follow).

For *size*, stimuli were generated using the ImageMagick command-line tools for Unix with Fred Weinhaus's extensions (www.fmwconcepts.com/imagemagick/randomblob/index.php). Each image included 16 randomly generated points drawn from a uniform distribution. The points were successively connected with a spline curve and a Gaussian blur was added to the lines of the image. We used two shapes each that were 200 and 325 pixels in area (Fig. 2a). For *angularity*, we generated six angular and rounded shape pairs using `Matlab`. Angular shapes included between 4 and 30 polar coordinates sorted and successively connected on a Cartesian grid. Rounded shapes were created by performing a quadratic spline on the angular shapes, thus controlling for overall size and number of edges (Fig. 2b). For *brightness*, we used three 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) colormap units (Fig. 2c). For *spatial frequency* (SF), the circles were 200 pixels in diameter and included high-contrast black and white sinusoidal gratings oriented 45° to the left presented on a gray background. We used three pairs of spatial frequency cycle differences: gratings differed by 14 (6 vs. 20), 10 (8 vs. 18), and 6 (10 vs. 16) cycles (Fig. 2d). For *height*, we used three pairs of visual stimuli: circles were ±20%, ±40% and ±80% vertically displaced from the screen's center (Fig. 2e).

2.3. *Design & procedure*

We ran all experiments using `Matlab` (2013b–2015b) with the Psychophysics Toolbox (Version 3) extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) running on Mac Minis (OS X 10.6 or higher). We presented visual stimuli on 19-inch Dell 1901/1905 FP monitors (1280 × 1024 pixels) and sounds through Sennheiser HD 555 headphones. All procedures were in accordance with the ethical standards of the University of Virginia Institutional Review Board and all participants provided informed consent before the experiment began.

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 (*e.g.*, larger/lower and smaller/higher) and two blocks had 'incompatible' instructions (*e.g.*, larger/higher and smaller/lower).

Fig. 3 illustrates the sequence of events during each trial. At the start of each trial, participants first saw instructions detailing which stimuli they should respond to on that given block of trials; for example, on the schematic trial provided, listeners must either choose the larger of the two circles or the lower of the two pitches. They were told they would receive a cue at the end of the trial as to which modality to respond to, thus ensuring that they attended to both modalities during the trial rather than one modality being irrelevant throughout as in previous studies.

Participants pressed the `SPACEBAR` to proceed with the trial, at which point the first shape appeared for 300 ms accompanied by a 300 ms high or low tone. This was followed by a 500 ms blank screen. Then the second shape appeared for 300 ms accompanied by a 300 ms high or low tone. Stimuli for the size, angularity, brightness, and spatial frequency experiments were centered vertically and appeared left/right of the center horizontally. Stimuli for the height experiment were centered horizontally and appeared above/below the center vertically.

After both pitch/shape pairings, either a 'p' for pitch or 's' for shape 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

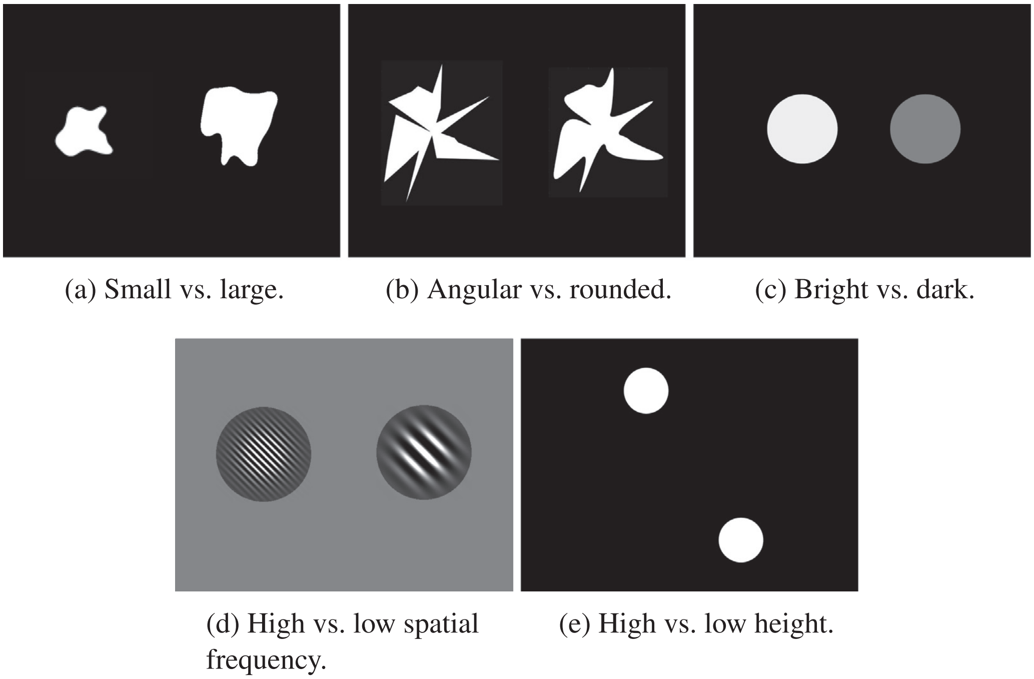


Fig. 2. Examples of visual stimuli used.

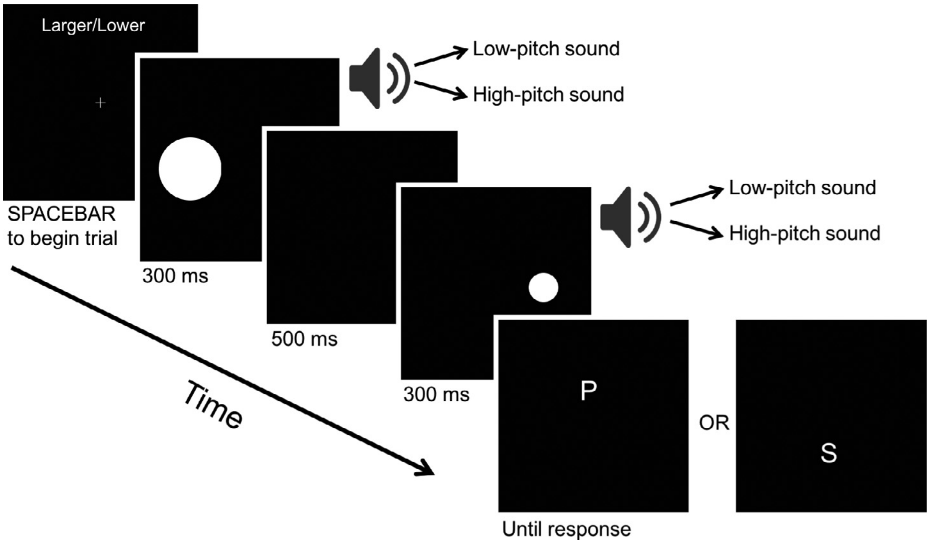


Fig. 3. Schematic illustration of the trial sequence used, shown with size stimuli.

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 a new instruction screen appeared.

2.4. Data analysis

We used R (R Development Core Team, 2016) for all of our analyses.

2.4.1. Data management

In all of the experiments, we rejected participants who did not achieve at least 60% accuracy on pitch and visual trials.

Our reaction time (RT) analysis only included correct responses.[3] For each experiment, we manipulated the RT data in three steps. First, we discarded RT ⩽ 50 ms, assuming that the participant had started their response before the modality cue appeared on the screen. Next, the RT 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 best transformation was generally logarithmic, we report logRT throughout.[4] Finally, we removed outliers that were more than three median absolute deviations (MADs) from the median RT.[5]

2.4.2. Mixed-effects modeling

We modeled our RT data with linear mixed-effects models (LMMS),

[3] Though our main focus in this paper is the RT analysis, we saw no evidence of a speed-accuracy trade-off in any experiment, and participants maintained an overall accuracy well above 90%.

[4] For experiments that suggested a different transformation, we ran the analyses both with the logRT and the alternate transformation. In no case did we find differences in significant results across transformations.

[5] MAD is a more robust measure of dispersion than standard deviation (Leys, Ley, Klein, Bernard, & Licata, 2013).

Table 2
Contrast analysis types and predicted results.

Contrast type	Measurement	Crucial comparison	Predicted direction
Bottom-Up (BU)	How much slower are participants on the incompatible blocks?	*compatible-incongruent*; incompatible-incongruent	zero or *positive*
Replication (RC)	How much of a congruency advantage exists on compatible blocks?	*compatible-congruent*; compatible-incongruent	*positive*
Top-down (TD)	How fast are trials that match the instructions?	compatible-congruent; *incompatible-incongruent*	zero or *negative*
Manipulation (MC)	How much of an incongruency advantage exists on incompatible blocks?	incompatible-congruent; *incompatible-incongruent*	*negative*

computed using the R package lme4 (Bates, Maechler, Bolker, & Walker, 2015). Because we are interested not in effects present only at an individual level but rather in generalizable effects, LMMS allow us to partition out subject-by-subject variations in model parameters 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). Full details of the fixed and random effects we used and significant interactions we found are included in S2 Methods & Results.

2.4.3. Model visualization

For our figures, we used least significant difference (LSD) bars[6] 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, Brockhoff, & Christensen, 2016) and predictmeans (Luo, Ganesh, & Koolaard, 2014). We used the average LSD value from the LMM models in the analyses. Any difference between groups larger than the height of the bar is statistically significant.

2.4.4. Contrast analysis

We performed contrast analyses using the glht function with user-defined contrasts from the R package multcomp (Hothorn, Bretz, & Westfall, 2008). This analysis used the compatibility (compatible vs. incompatible) and congruency (congruent vs. incongruent) variables to create four conditions (visualized as the black shapes in Fig. 1): compatible-congruent (CC), compatible-incongruent (CI), incompatible-congruent (IC), incompatible-incongruent (II).

Table 2 summarizes the four contrasts we included. Two contrasts make up our measurement of “bottom-up-ness”, defined as the ease with which participants completed the task on compatible as opposed to incompatible blocks. The bottom-up (BU) contrast is the logRT difference between CI and II trials (visualized by the red[7] lines in Fig. 1). If there is a strong bottom-up effect, we would expect that pairing together the incongruent dimensions would result in slower performance overall on the incompatible blocks; thus values of zero or *positive* values would be predicted. The replication (RC) contrast is the difference between the logRT on CC and CI trials (visualized by the green lines in Fig. 1). On compatible trials, a successful replication would mean participants should show a congruency advantage; thus we expect the RC comparison to yield a positive value.

An additional two contrasts make up our measurement of “top-down-ness”, defined as how well participants followed the instructions on compatible and incompatible blocks. The top-down (TD) contrast is the logRT difference between CC and II trials (visualized by the blue lines in Fig. 1). If there is a strong top-down effect, we would expect participants to be just as fast when asked to pair together the incongruent dimensions; thus values of zero or *negative* values would be predicted. The manipulation (MC) contrast is the difference between the logRT on IC and II trials (visualized by the purple lines in Fig. 1). On incompatible trials, a successful manipulation would mean participants show an incongruency advantage (as those dimensions are now paired in the instructions); thus we expect the MC comparison to yield a negative value.

In Sections 3.1–3.5, we report 95% confidence intervals for the contrast parameters instead of reporting null-hypothesis tests. These CIs may be interpreted as tests of significance: if the confidence interval for an estimated contrast does not straddle zero, this estimate may be considered significant at $\alpha < 0.05$.

3. Results

Below we report the bottom-up and top-down contrast estimates for each correspondence, ordered by decreasing top-down influence. Fig. 4 provides a depiction of the results of the four contrast types (see Table 2). In each case, the replication and manipulation checks were successful (*i.e.*, positive values for replication and negative values for manipulation).

3.1. Size

As shown in Fig. 4a, the RT in the incompatible-incongruent condition was equal to the compatible-congruent condition (TD = 0.004 $[-0.027, 0.035]$). The RT in the incompatible-incongruent condition was significantly faster than the compatible-incongruent condition (BU = -0.052 $[-0.083, -0.020]$). This means that participants followed the instructions without a loss in reaction time when the response pairing was reversed, resulting in high top-down influence and little evidence of a bottom-up association.

3.2. Angularity

Unlike with pitch–size, Fig. 4b shows that the RT in the incompatible-incongruent condition was not as fast as in the compatible-congruent condition, resulting in less top-down influence (TD = 0.125 $[0.091, 0.159]$). The RT in the incompatible-incongruent condition was still significantly faster than the compatible-incongruent condition, resulting in a similarly small bottom-up effect (BU = -0.052 $[-0.087, -0.017]$).

3.3. Brightness

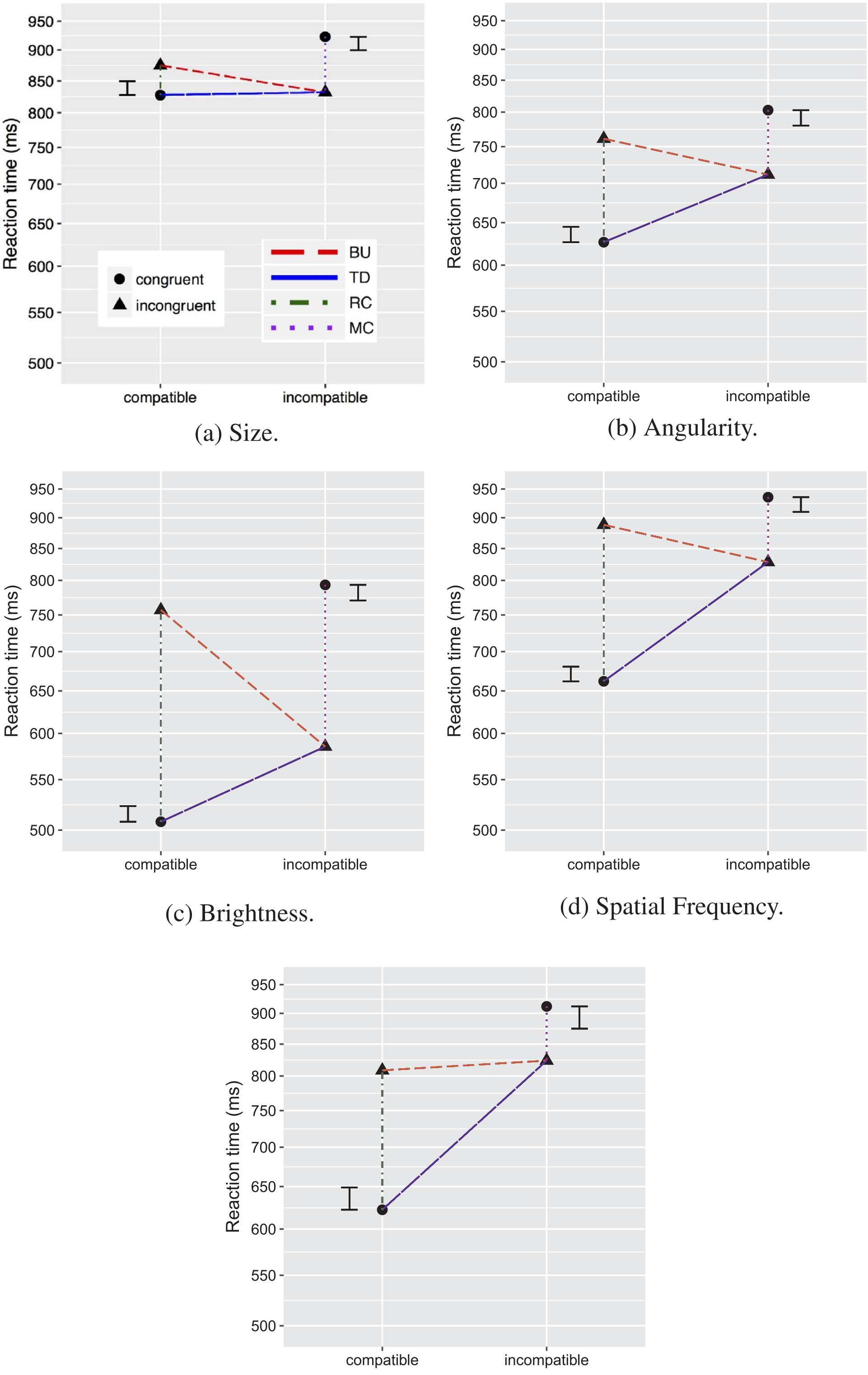
Fig. 4c shows similar top-down influence as pitch–angularity (TD = 0.138 $[0.101, 0.174]$). However, the bottom-up effect is even smaller here, as the RT in the incompatible-incongruent condition was much faster than in the compatible-incongruent condition (BU = -0.253 $[-0.289, -0.217]$).

3.4. Spatial frequency

Fig. 4d shows a similar bottom-up effect as pitch–size and

[6] 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. We thus use LSD bars to visualize the model predictions in order to avoid these large error bars that make inferences regarding significance difficult.

[7] For interpretation of color in Figs. 1, 4, and 5, the reader is referred to the web version of this article.



(a) Size.

(b) Angularity.

(c) Brightness.

(d) Spatial Frequency.

(e) Height.

Fig. 4. Effect plots showing RTs (back-transformed to RT from `logRT` for ease of interpretation) for congruency and compatibility conditions for each audiovisual correspondence. Each figure includes upper and lower `LSD` bars: any difference between groups larger than the height of the `LSD` bar is statistically significant. The four congruency × compatibility conditions were used to create the bottom-up (BU), top-down (TD), replication check (RC), and manipulation check (MC) contrasts: each contrast was calculated as a comparison between the two linked shapes.

pitch–angularity (BU = -0.071 $[-0.104, -0.039]$). Additionally, the top-down influence is weaker here than the previously-mentioned correspondences (TD = 0.206 $[0.173, 0.240]$).

3.5. Height

As shown in Fig. 4e, the RT in the incompatible-incongruent condition is slower than the compatible-congruent *and* compatible-

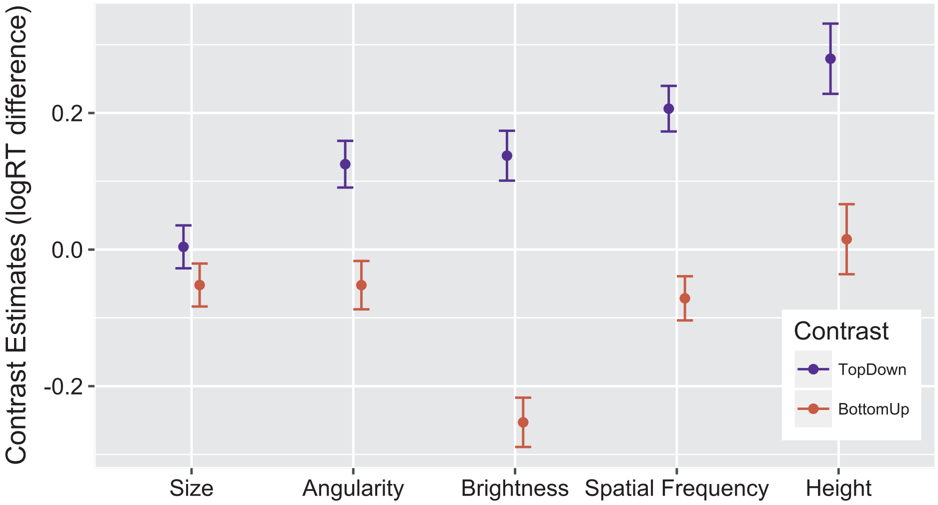


Fig. 5. Bottom-up (BU) and Top-down (TD) estimates with 95% confidence intervals by correspondence. A strong top-down effect results in a zero or *negative* estimate, meaning size has the strongest top-down influence. A strong bottom-up effect results in a zero or *positive* estimate, meaning height has the strongest bottom-up association.

incongruent conditions, providing the strongest evidence for a bottom-up effect (BU = 0.015 [−0.036, 0.067]) and only weak support for top-down influence (TD = 0.28 [0.228, 0.331]).

3.6. Summary

Fig. 5 provides a summary of the bottom-up and top-down contrast estimates for the five correspondences. We found that top-down influence was present to varying degrees in all five cases, with the strongest influence on size and weakest influence on height. Additionally, most correspondences had at least a small bottom-up effect as well, with the strongest effect on height and weakest effect on brightness.

An additional pitch–size experiment using percussion tones (rather than sine tones) and experiments of the pitch–height and pitch–spatial frequency correspondence using different words to describe the auditory and visual modalities (rather than overlapping words) are included in S3 Additional Experiments. The results do not add anything critical to the interpretation, but provide interesting insights into the influence of timbre and lexical overlap.

4. Discussion

By creating a novel paradigm to separately measure bottom-up and top-down effects, we provide a first step in being able to *quantify* the degree of automaticity present in a number of audiovisual correspondences. This fills an important gap in the literature that has largely focused on a dichotomous view of automaticity rather than focusing on the extent to which various automaticity criteria are met (cf. Spence & Deroy, 2013). Together, the results of our five experiments point to the fact that AVCs *jointly* rely on bottom-up and top-down processing rather than being *solely* explained by an automatic association (Evans & Treisman, 2010; Parise & Spence, 2012) or *solely* operating at a strategic, top-down level (Chiou & Rich, 2012; Klapetek et al., 2012).

It is important to note that given the nature of our cuing task, it would have been surprising (though not impossible; see Fig. 1d) to *not* see a top-down effect of instructions. However, what was more important to us than just whether we saw top-down influence was to assess the degree of influence across the various AVCs tested. Our results point to the fact that a greater degree of top-down influence may be indicative of the visual dimension providing a less natural metaphor to describe pitch (in English and many other languages; Eitan & Timmers, 2010).

The correspondence with the least top-down influence (and largest bottom-up association) was *height*, which is the dominant metaphor English speakers use when talking about pitch. This is not to imply that the overlapping verbal labels are the cause of the implicit association (*e.g.*, see S3.2), but rather that this language-specific metaphor shapes people's nonlinguistic representations of musical pitch to such an extent that they may not realize they are using such a spatial metaphor (Dolscheid, Shayan, Majid, & Casasanto, 2013). Indeed, cross-cultural (*e.g.*, Parkinson, Kohler, Sievers, & Wheatley, 2012) and developmental (*e.g.*, Fernández-Prieto, Navarra, & Pons, 2015; Walker et al., 2010) work has provided evidence that the pitch–height association is not purely semantic in nature as infants and individuals from cultures who use other pitch metaphors still show effects of pitch directionality on elevation judgments. Some studies have even found evidence of cross-modal associations in non-human animals (*e.g.*, chimpanzees, Ludwig, Adachi, & Matsuzawa, 2011) showing that the association cannot purely be a linguistic phenomenon.

At the other extreme, *size* was influenced the most by top-down processing. English speakers never use the terms 'small' and 'large' to describe 'high' and 'low' pitches, and thus the association seems easier to override with different task instructions. Pitch-spatial frequency, brightness, and angularity fall between the two extremes set by size and height. Though *spatial frequency* can also be described using the words 'high' and 'low', this usage was generally unfamiliar to our participants, who thought that it made more sense to describe the width of the stripes as narrow (*i.e.*, high SF) or wide (*i.e.*, low SF). Also, spatial frequency may relate more to the physical property of auditory frequency (*i.e.*, higher repetition rate corresponding to a higher frequency) rather than to our metaphorical understanding of pitch. *Angularity* and *brightness* are also adjectives occasionally used to describe pitch (*e.g.*, high sounds are described as 'jagged', 'shrill', and 'bright'; low sounds are described as 'dark', 'full' and 'round'), but these relate more to timbral than pure tone differences.

In summary, the fact that all of the AVCs we investigated showed at least *some* top-down influence contrasts previous studies that concluded AVCs are automatic. However, the fact that most of the AVCs also showed at least *some* degree of a bottom-up association contrasts studies arguing for complete top-down control. We therefore conclude that it is likely that there is a more natural dimensional pairing for these correspondences that exists automatically due to structural brain connectivity, natural scene statistics, or semantic overlap (Parise, 2016; Spence, 2011). Nonetheless, when the instructions ask participants to pair the dimensions in the opposite direction, little experience or learning is needed to recouple the associations, leading to top-down effects of the task. The only time significant relearning is required to override the natural inclination is with a strong metaphor such as pitch and height. Note that although we discuss this top-down effect in terms of changing the "coupling" between the auditory and visual dimensions, we are not attempting to make claims about whether the dimensions fuse into a single complex dimension or whether observers necessarily decide the auditory and visual features belong to the same object, as with the unity assumption (cf., Chen & Spence, 2017). It is also an open question whether similar conclusions would extend to correspondences between other audiovisual dimensions or other pairs of modalities.

Acknowledgements

The work presented here was part of LMG's dissertation work; thanks to her committee (Michael Kubovy, Dennis Proffitt, C. Daniel Meliza, and Filip Loncke) for their helpful insights. Others who provided helpful discussions include Steven Scheid, Daniel Shanahan, and Joseph Toscano. LMG is now supported by a Mendel Science Experience Post-doctoral Fellowship from the Villanova University College of Liberal Arts and Sciences.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2018.02.015.

References

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. http://dx.doi.org/10.1016/j.jml.2007.12.005.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

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

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

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Chen, Y. C., & Spence, C. (2017). Assessing the role of the 'unity assumption' on multisensory integration: A review. *Frontiers in Psychology, 8*(445), 1–22.

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception, 41*, 339–353.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological 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.

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

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

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences, 39*, 1–77.

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* (2nd ed.). Thousand Oaks, CA: Sage. < http://socserv.socsci.mcmaster.ca/jfox/Books/Companion > .

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

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

Heron, J., Roach, N. W., Hanson, J. V., McGraw, P. V., & Whitaker, D. (2012). Audiovisual time perception is spatially specific. *Experimental Brain Research, 218*, 477–485.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*(3), 346–363.

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Do crossmodal correspondences enhance the facilitatory effect of auditory cues on visual search? *Attention, Perception, and Psychophysics, 74*, 1154–1167.

Klein, R. M., Brennan, M., & Gilani, A. (1987). November. Covert cross-modality orienting of attention in space. *Annual meeting of the psychonomics society*. Seattle.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception, 36*(ECVP Abstract Supplement).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (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).

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.

Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (pan troglodytes) and humans. *Proceedings of the National Academy of Sciences, 108*(51), 20661–20665.

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

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.). *Handbook of multisensory processes* (pp. 85–105). Cambridge, MA: MIT Press.

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

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.

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

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*, 297–326.

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.

Parise, C. V. (2016). Crossmodal correspondences: Standing issues and experimental guidelines. *Multisensory Research, 29*, 7–28.

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.

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research, 220*(3–4), 319–333.

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.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

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.

Santangelo, V., & Spence, C. (2008). Is the exogenous orienting of spatial attention truly automatic? Evidence from unimodal and multisensory studies. *Consciousness and Cognition, 17*, 989–1015.

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

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

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition, 22*, 245–260.

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., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science, 21*, 21–25.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 3*, 638–667.