



# Cross-modal correspondences in sine wave: Speech versus non-speech modes

Daniel Márcio Rodrigues Silva<sup>1</sup> · Samuel C. Bellini-Leite<sup>2</sup>

Published online: 14 August 2019  
© The Psychonomic Society, Inc. 2019

## Abstract

The present study aimed to investigate whether or not the so-called “bouba-kiki” effect is mediated by speech-specific representations. Sine-wave versions of naturally produced pseudowords were used as auditory stimuli in an implicit association task (IAT) and an explicit cross-modal matching (CMM) task to examine cross-modal shape-sound correspondences. A group of participants trained to hear the sine-wave stimuli as speech was compared to a group that heard them as non-speech sounds. Sound-shape correspondence effects were observed in both groups and tasks, indicating that speech-specific processing is not fundamental to the “bouba-kiki” phenomenon. Effects were similar across groups in the IAT, while in the CMM task the speech-mode group showed a stronger effect compared with the non-speech group. This indicates that, while both tasks reflect auditory-visual associations, only the CMM task is additionally sensitive to associations involving speech-specific representations.

**Keywords** Cross-modal correspondences · Sound symbolism · Bouba-kiki effect · Sine-wave speech

## Introduction

Cross-modal correspondences are tendencies to match perceptual features from different sense modalities (Deroy & Spence, 2013; Spence, 2011). A famous example is the bouba-kiki phenomenon, in which people associate pseudowords like “kiki” and “takete” with spiky shapes, and pseudowords like “bouba” and “maluma” with curved shapes (Kohler, 1947; Ramachandran & Hubbard, 2001). These nonwords combine multiple consonant and vowel features that contribute to the effect. Voiced labial stop [b], sonorants [m, l], and back round vowels [u, o] are consistently matched to curvy shapes; voiceless alveolar and velar stops [t, k], and front vowels [i, e] are consistently matched to spiky shapes (D’Onofrio, 2014; McCormick et al. 2015; Nielsen & Rendall, 2013). The bouba-kiki effect has been documented across different

cultures (Bremner et al., 2013; Styles & Gawne, 2017) and throughout development (Lockwood & Dingemans, 2015; Maurer, Pathman, & Mondloch, 2006) – the relative importance of the involved stimulus properties can vary with culture (Chen, Huang, Woods, & Spence, 2016) and developmental stages (Chow & Ciaramitaro, 2019). It is also an instance of sound symbolism, i.e., a consistent, non-arbitrary relationship between phonetic and perceptual or semantic elements establishing sound-meaning association biases in language (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016; Sidhu & Pexman, 2018).

An issue that deserves careful attention is what sorts of speech-sound representations – e.g., auditory, phonological, articulatory, lexical (Hickok & Poeppel, 2007; Monahan, 2018) – drive the bouba-kiki effect. Parise and Spence (2012) suggest that it reflects structural similarities between physical features of auditory and visual stimuli. Spiky versus curved shapes have been found to correspond with lower versus higher pitch (Marks, 1987; O’Boyle & Tarte, 1980; Walker et al., 2010), sinusoidal versus square-wave tones (the latter are richer in higher frequencies), and sharp-attack versus gradual-attack musical timbres (Adeli, Rouat, & Molotchnikoff, 2014). Accordingly, “spikiness” in speech sounds has been assigned to abrupt energy changes in voiceless compared to voiced consonants, and to the high second-formant frequency of front vowels; “curviness” is seen as

✉ Samuel C. Bellini-Leite  
samuelcblpsi@gmail.com

<sup>1</sup> Federal University of Minas Gerais (UFMG), Presidente Antônio Carlos Ave. 6627 - Campus Pampulha, Belo Horizonte, Minas Gerais, Brazil

<sup>2</sup> State University of Minas Gerais (UEMG) and Federal University of Minas Gerais (UFMG), Presidente Antônio Carlos Ave. 7545 - São Luiz, Belo Horizonte, Minas Gerais, Brazil

related to low-frequency energy due to consonant voicing (e.g., by smoothing consonant-vowel transitions) and to a lower second formant in back vowels (Fort, Martin, & Peperkamp, 2015; Knoeferle, Li, Maggioni, & Spence, 2017; Nielsen & Rendall, 2011).

Explanations based on speech-specific rather than general auditory processing have also been offered. Particularly, properties of articulatory gestures that mimic jagged or smooth visual contours, such as sharp inflections of the tongue on the palate (Ramachandran & Hubbard, 2001) or lip rounding/stretching (Maurer et al., 2006) are thought to mediate correspondences between speech sounds and shapes. Further, Styles and Gawne (2017) suggest that failures to observe the bouba-kiki effect in speakers of Syuba and Hunjara are due to the tested pseudowords being phonologically illegal in those languages. This would imply that the effect requires sounds to be mapped onto language-specific phonological structures, which seems at odds with “general auditory” explanations and with evidence of sound symbolic effects in preverbal infants as young as 4 months (Ozturk, Krehm, & Vouloumanos, 2013; but see Fort, Weiß, Matin, & Peperkamp, 2013).

The present study aimed to investigate the roles of speech-specific and general auditory processes in the bouba-kiki phenomenon by comparing two conditions: one in which the auditory stimuli are heard as speech, and one in which the same stimuli are heard as non-speech sounds. For this, we used sine-wave speech (SWS), a spectrally reduced form of speech that can be heard as speech or non-speech depending on whether or not the listener attends to the speech-likeness of the sounds. SWS consists of sinusoidal tones (usually three) imitating time-varying properties of vocal-tract resonances (Remez, Rubin, Pisoni, & Carrell, 1981). Due to the absence of the harmonic and broadband formant structure characteristic of natural vocalizations, it sounds quite different from human speech and generally elicits no phonetic perception in naïve listeners. However, through proper instruction, listeners can direct attention to phonetic information in SWS that is sufficient to support perception of the linguistic message (Remez, Rubin, & Pisoni, 1983; Remez, Rubin, Pisoni, & Carrell, 1981). As proposed by Remez and Thomas (2013), while the vocal timbre of natural speech directs listeners’ attention to modulations caused by articulatory gestures, which engages a perceptual organization of the signal into a speech stream, SWS is not sufficient to summon such attentional setting, usually requiring further information, such as instructions. Interestingly, hearing SWS as speech versus non-speech has been found to involve functionally distinct perceptual processes and brain networks (Dehaene-Lambertz, 2005; Khoshkhou, Leonard, Mesgarani, & Chang, 2018; see also Baart, Stekelenburg, & Vroomen, 2014).

In the current experiment, participants that were trained to hear SWS stimuli as speech were compared to participants

that were not informed about the nature of the same stimuli. Comparisons were made in terms of performance on an implicit association task (IAT; Parise & Spence, 2012) and a subsequent explicit cross-modal matching (CMM) task, both used to assess sound-shape correspondences. In both cases, the tested hypothesis was that, while the correspondence effect would occur in both groups due to auditory-visual associations, it would be stronger for the speech-mode group than for the non-speech group, due to associations at speech-specific processing levels.

## Material and methods

### Participants

Fifty-four native speakers of Brazilian Portuguese (mean age: 26.4, SD: 5.0; range: 18–34 years; 29 females) participated as volunteers. They all provided written informed consent and reported no history of hearing or neurological problems. The study was approved by the local ethics committee and was conducted in accordance with the Declaration of Helsinki. Regarding the linguistic and cultural background of the participants, it is worth noting that testing languages other than English is beneficial for the field of cross-modal correspondences. We are aware of only one other study on the bouba-kiki effect in speakers of Portuguese (Godoy et al., 2018).

### Stimuli

Sine-wave versions of the legal pseudowords *maluma* and *taketa* were used as auditory stimuli in the experimental tasks. The choice of these pseudowords was based on Kohler’s (1947) seminal work and subsequent replications of the “maluma-takete” effect. Here, *taketa* was used instead of *takete* because the latter would be pronounced as [taketʃi] in most variants of Brazilian Portuguese. Three *maluma* exemplars and three *taketa* exemplars spoken by a male native speaker of Brazilian Portuguese were recorded. For the preparatory task, one exemplar of each of eight legal pseudowords was recorded. These pseudowords were formed by recombining syllables in *maluma* and *taketa* (half with two syllables from the former; half with two syllables from the latter): *ketalu*, *keluma*, *kemata*, *luketa*, *lumata*, *maketa*, *maluta*, *tamalu*. For each pseudoword exemplar, a SWS sound composed of three time-varying sinusoids corresponding to the three lower formants of the natural utterance were generated using a script for the software Praat (Boersma & Weenink, 2013) written by Darwin (2003). All SWS stimuli were normalized for equal root mean square intensity (70 dB SPL) and presented binaurally through TDH-39 headphones.

Visual stimuli were six abstract shapes presented in a computer monitor positioned approximately 1.0 m from the

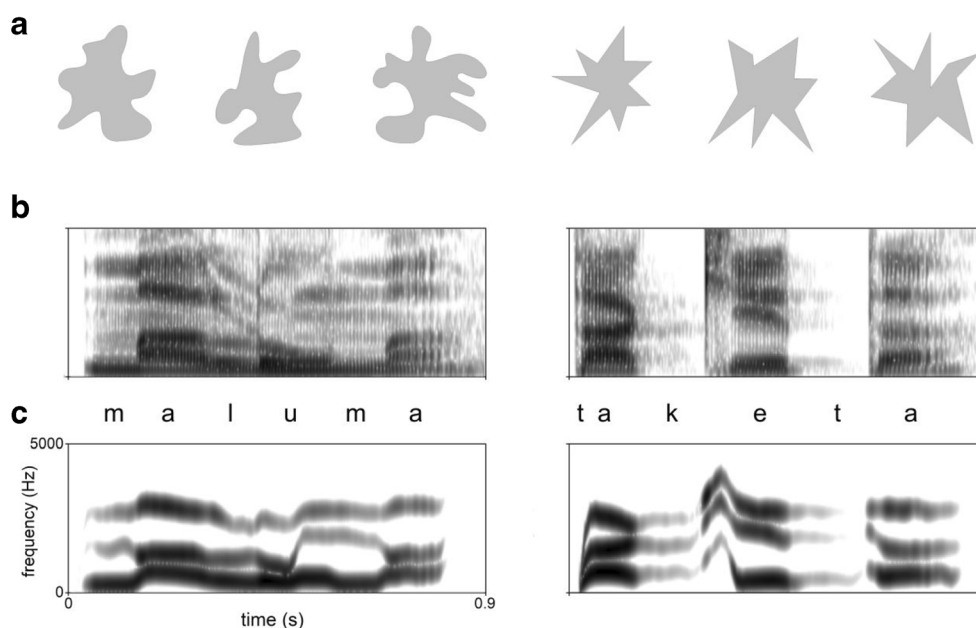
participant's eye. Three of them were spiky; the others were curved (Fig. 1). The shapes were presented in dark gray within a white rectangle subtending approximately  $7.15 \times 6.30^\circ$  of visual angle on the center of the screen with a black background.

## Design and procedures

Once participants learn to hear SWS as speech, there is no known way to induce them to revert to a non-speech mode of perception. Hence, a between-participants design was necessary to avoid confounds that would result from testing participants in speech-mode condition following a non-speech condition. In order to induce one group of participants into speech-mode and the other into non-speech mode, assigning them to different preparatory conditions before the main experimental tasks was inevitable. Simply telling a group of participants about the speech-likeness of the SWS stimuli in those tasks could be insufficient to induce most participants into the speech mode (see Remez et al., 1981); informing them specifically about the original *maluma* and *taketa* utterances could bias performance. Another possibility would be to present SWS stimuli paired with the original utterances to the speech-mode group, while presenting only SWS stimuli to the non-speech group (e.g., Vroomen & Stekelenburg, 2011), in which case the set of auditory stimuli would differ between groups. Here, we prioritized having both groups being exposed to the exact same stimuli under conditions requiring active listening.

Participants were randomly assigned to two groups ( $N = 27$ ) that differed only in the preparatory task they were to perform before the main experimental tasks. The *speech-mode* group performed a pseudoword identification task, while the *non-speech* group performed a sound localization task. The exact same SWS stimuli were used in both tasks, but only the former directed attention to the speech-like nature of the sounds. We assume that the only way the two preparatory tasks could affect subsequent tasks differently is by inducing, or failing to induce, the speech mode of perception. After completing the preparatory task, participants performed the IAT followed by the CMM task. For all tasks, the auditory or visual stimulus was presented 1.00 s after the response to the previous trial. Feedback was given in all trials of the IAT and preparatory tasks. At the end of the experiment, participants were asked to describe what kind of sound they heard during the tasks and, following the response, whether they heard the sounds as sequences of spoken syllables. Presentation software (Neurobehavioral Systems, Albany, CA, USA) was used to program and run all tasks.

**Preparatory tasks** In both preparatory tasks, SWS pseudowords *ketalu*, *keluma*, *kemata*, *luketa*, *lumata*, *maketa*, *maluta*, and *tamalu* were used as stimuli, each presented twice in a randomized sequence. The sound localization task given to the *non-speech* group consisted of 16 trials in each of which a SWS pseudoword was presented and the participant was required to press a button indicating whether the sound was to the left or right. In each trial, either the left or right channel was attenuated by either 15 or 10 dB. Attenuation direction



**Fig. 1** Visual and auditory stimuli. (a) The spiky and curved shapes used as visual stimuli in the explicit task. The first and fourth shapes (from left to right) were used in the implicit association task. (b) Spectrograms of an

original *maluma* utterance (left) and an original *taketa* utterance (right). (c) Sine-wave speech versions of the utterances represented in B

and degree were counterbalanced and randomly assigned across trials.

The *speech-mode* group performed a pseudoword identification task. Along with the auditory stimulus, a pair of written pseudowords appeared in the center of the screen (one above the other) in each trial. One of the written pseudowords corresponded to the accompanying SWS stimulus; the other was drawn randomly from the remaining seven written pseudowords. The participant was required to press a button indicating which written pseudoword matched the auditory stimulus.

The auditory stimuli were the same in the two preparatory tasks (including left-right amplitude differences). While both tasks required active listening to the stimulus set, they critically differed in that only pseudoword identification required mapping sounds into linguistic units and involved orthographically presented pseudowords – two features that can reasonably be assumed to prime participants to attend to the speech-likeness of the stimuli.

The average proportion of correct responses was 95.5% for the *non-speech* group (in the sound localization task) and 91.5% for the *speech-mode* group (in the pseudoword identification task). This difference between groups was significant (Wilcoxon rank sum test:  $W = 149$ ;  $p = .013$ ). More importantly, the high proportions of correct responses suggest that participants of both groups had little difficulty in performing the preparatory task and were successfully primed to attend to either sound location or speech-related features.

**Implicit association task (IAT)** Two visual shapes (curved and spiky) and two SWS pseudoword exemplars (*maluma* and *taketa*) were used as stimuli (Fig. 1). The participant was asked to keep the index fingers resting on two response keys. Each of the 12 blocks of this task was composed of three phases: *teaching*, *training*, and *test*. In each of four *teaching* trials, an auditory or visual stimulus was presented along with an arrow indicating the corresponding response key. An arrow pointing to the left (right), in the left (right) inferior corner of the screen, indicated that the left (right) key should be pressed in response to the current stimulus. Presentation order was randomized with the constraint that auditory and visual stimuli should alternate.

The *training* and *test* phases within a block were identical except for the number of trials (eight and 16, respectively) and the instruction for the participant to respond as quickly as possible (without sacrificing accuracy) during the test. In each trial, an auditory or visual stimulus was presented and the participant had to respond according to the stimulus-response mapping specified in the *teaching* phase. Presentation order was randomized (with no stimulus being presented in two immediately successive trials). Accuracy and response time were recorded only during the test phase. There were three blocks for each of the four possible mappings in

which one auditory and one visual stimulus was mapped onto each response key. The 12 blocks were presented in randomized order. In six *congruent* blocks, the sound *taketa* (*maluma*) was mapped onto the same response key as the spiky (curved) shape. In other six *incongruent* blocks, *taketa* (*maluma*) was mapped onto the same key as the curved (*spiky*) shape. A brief pause was allowed between any two blocks.

**Cross-modal matching (CMM) task** Six trials were presented in which the participant heard a SWS pseudoword and used a visual analog scale to indicate whether it matched better to one of two shapes presented side-by-side on the screen. Unlike traditional binary forced-choice paradigms, this task is sensitive to gradient, subcategorical detail in stimuli (Schellinger, Munson, & Edwards, 2017) and allows participants to respond “none of the shapes.” Two keyboard keys were used to scroll a gray square along a horizontal white bar at the bottom of the screen, indicating to what degree the participant thought the sound matched the left or the right shape (middle meaning “indifferent”). Three *maluma* and three *taketa* exemplars were used as auditory stimuli. Six curvy-spiky pairs formed by the shapes shown in Fig. 1 were used as visual stimuli. Spiky and curved shapes appeared an equal number of times (three) to the right and to the left. Presentation order was randomized – three consecutive exemplars of the same pseudoword were not allowed.

## Results

In the *speech-mode* group, most participants reported hearing pseudowords or Portuguese words, while most participants in the *non-speech* group reported hearing whistles and/or “electronic sounds” – *taketa* being often described as “treble” relative to *maluma*. Data from four participants of the *speech-mode* group were excluded either for not giving at least 11 correct responses in the preparatory task or for not reporting hearing the auditory stimuli as well-defined spoken syllable sequences. Data from four participants of the *non-speech* group were excluded because they reported hearing auditory stimuli as well-defined spoken syllables.

### Implicit association task (IAT)

Mixed-effects models with Group (speech mode  $\times$  non-speech), Congruency (congruent  $\times$  incongruent) and Stimulus Modality (auditory  $\times$  visual) as fixed effects were used to analyze both accuracy and response-time data. By-participant random intercepts and random slopes for Congruency, Modality, and their interaction were specified. To account for learning/fatigue effects, random intercepts for block (1, 2, ... 12) and trial (1, 2, ... 16) nested within block were also included. Response times (for correct responses)

between 300 and 3,000 ms were ln-transformed and entered into a linear model. Accuracy was entered as a binary response variable into a logit model (Jaeger, 2008). The main interest was to test for the Group  $\times$  Congruency interaction, since the tested hypothesis was that cross-modal correspondences, as assessed by the effect of Congruency, would be stronger in the *speech-mode* compared with the non-speech group. Fixed-effect coefficients in the “response time” and “accuracy” models are shown in Tables 1 and 2, respectively. Mean response times and accuracy for the two groups in congruent and incongruent blocks are depicted in Fig. 2a and b. To calculate p-values for fixed effects, restricted models – each of which omitted one model term – were tested against the full model. For the “response time” linear model, this was done using a Type-III ANOVA with Satterthwaite’s approximation for degrees of freedom. For the “accuracy” logit model, likelihood-ratio tests were performed.

The ANOVA revealed that, as in Parise and Spence (2012), responses were faster in congruent compared to incongruent blocks ( $F_{(1, 44.3)} = 28.42$ ;  $p < .001$ ) and in visual compared to auditory trials ( $F_{(1, 43.5)} = 115.8$ ;  $p < .001$ ). The lack of significant interactions involving Congruency and Group ( $F < 1$ ) indicates that the *speech* and *non-speech* groups were similar in terms of the response time advantage in the congruent blocks – and, therefore, did not support the tested hypothesis.

Likelihood-ratio tests revealed significant Congruency ( $\chi^2_{(1)} = 13.45$ ;  $p < .001$ ) and Modality ( $\chi^2_{(1)} = 62.96$ ;  $p < .001$ ) effects, but no significant Group  $\times$  Congruency ( $\chi^2_{(1)} = 0.00$ ;  $p = .99$ ) nor Group  $\times$  Congruency  $\times$  Modality ( $\chi^2_{(1)} = 0.55$ ;  $p = .46$ ) interaction, indicating that the congruency effect was similar between groups. Importantly, as one can see in Fig. 2a and b, response-time and accuracy measures across conditions are not even numerically consistent with the tested hypothesis, indicating that the present negative results are not due to lack of statistical power.

As pointed out by an anonymous reviewer, examining the speed-accuracy relation is relevant for the interpretation of the IAT results. None of the groups showed speed-accuracy

tradeoffs. Rather, accuracy decreased with response time (Fig. 2c). After centering and scaling ln-transformed response times to unit variance, we added a “Response Time” fixed effect and the corresponding by-participant random slopes in the above-described “accuracy” logit model (see Davidson & Martin, 2013). Likelihood ratio tests revealed significant main effects of Response Time ( $\chi^2_{(1)} = 49.00$ ;  $p < .001$ ), Congruency ( $\chi^2_{(1)} = 8.91$ ;  $p = .003$ ), and Modality ( $\chi^2_{(1)} = 25.07$ ;  $p < .001$ ). No other significant or marginally significant model term was revealed. Particularly, the absence of a significant Response Time  $\times$  Group interaction ( $\chi^2_{(1)} = 0.13$ ;  $p = .72$ ) suggests that the decrease in accuracy with response time was similar between groups.

### Cross-modal matching (CMM)

Responses were coded as values from -1 to +1 such that positive and negative values represent, respectively, “spiky” and “curved” responses. Boxplots in Fig. 3 represent responses of the two groups to the two pseudowords. For both groups, SWS pseudowords *taketa* and *maluma* seem consistently associated with spiky and curvy shapes, respectively. However, the separation between responses to *taketa* and to *maluma* is clearer in the *speech-mode* than in the *non-speech* group – suggesting a stronger sound-shape association in the former group. Since the response variable was bounded, a nonparametric ANOVA was conducted on aligned rank transformed data (Wobbrock et al., 2011) with Pseudoword (*maluma*  $\times$  *taketa*) and Group as fixed effects, and Participant as random effect. A significant interaction was found between Group and Pseudoword ( $F_{(1, 228)} = 18.11$ ;  $p < .001$ ), reflecting the stronger sound-shape association in the *speech-mode* group and, therefore, supporting the tested hypothesis. Separate (Bonferroni corrected) ANOVAs for each group revealed highly significant effects of Pseudoword for both the *non-speech* ( $F_{(1, 114)} = 66.74$ ;  $p < .001$ ) and the *speech-mode* group ( $F_{(1, 114)} = 209.71$ ;  $p < .001$ ).

**Table 1** Summary of the fixed effects in the mixed linear model (fitted with REML) on log response times in the implicit association task (7910 observations)

Predictor	Coefficient	SE	df	t value	p
Intercept	6.836	0.027	40.3	249.27	< .001
Congruency ( <i>congr</i> = 1; <i>incongr</i> = -1)	-0.048	0.009	44.3	-5.33	< .001
Group ( <i>speech</i> = 1; <i>non-speech</i> = -1)	0.009	0.021	44.0	0.41	.68
Modality ( <i>auditory</i> = 1; <i>visual</i> = -1)	0.083	0.008	43.5	10.76	< .001
Congruency $\times$ Group	-0.003	0.009	43.9	-0.35	.73
Congruency $\times$ Modality	0.005	0.004	288.9	1.24	.22
Group $\times$ Modality	0.006	0.008	43.5	0.81	.42
Congruency $\times$ Group $\times$ Modality	0.003	0.004	288.7	0.65	.51

Contrast coding is shown in parentheses

**Table 2** Summary of the fixed effects in the mixed logit model on proportion correct data in the implicit association task (8,636 observations)

Predictor	Coefficient	SE	Wald Z	<i>p</i>
Intercept	2.720	0.165	16.50	< .001
Congruency ( <i>congr.</i> = 1; <i>incongr.</i> = -1)	0.231	0.057	4.03	< .001
Group ( <i>speech</i> = 1; <i>non-speech</i> = -1)	-0.079	0.117	0.68	.50
Modality ( <i>auditory</i> = 1; <i>visual</i> = -1)	-0.328	0.042	-7.91	< .001
Congruency × Group	0.001	0.052	0.02	.99
Congruency × Modality	0.012	0.042	0.26	.80
Group × Modality	0.064	0.041	1.53	.13
Congruency × Group × Modality	0.031	0.041	0.75	.45

Contrast coding is shown in parentheses

## Discussion

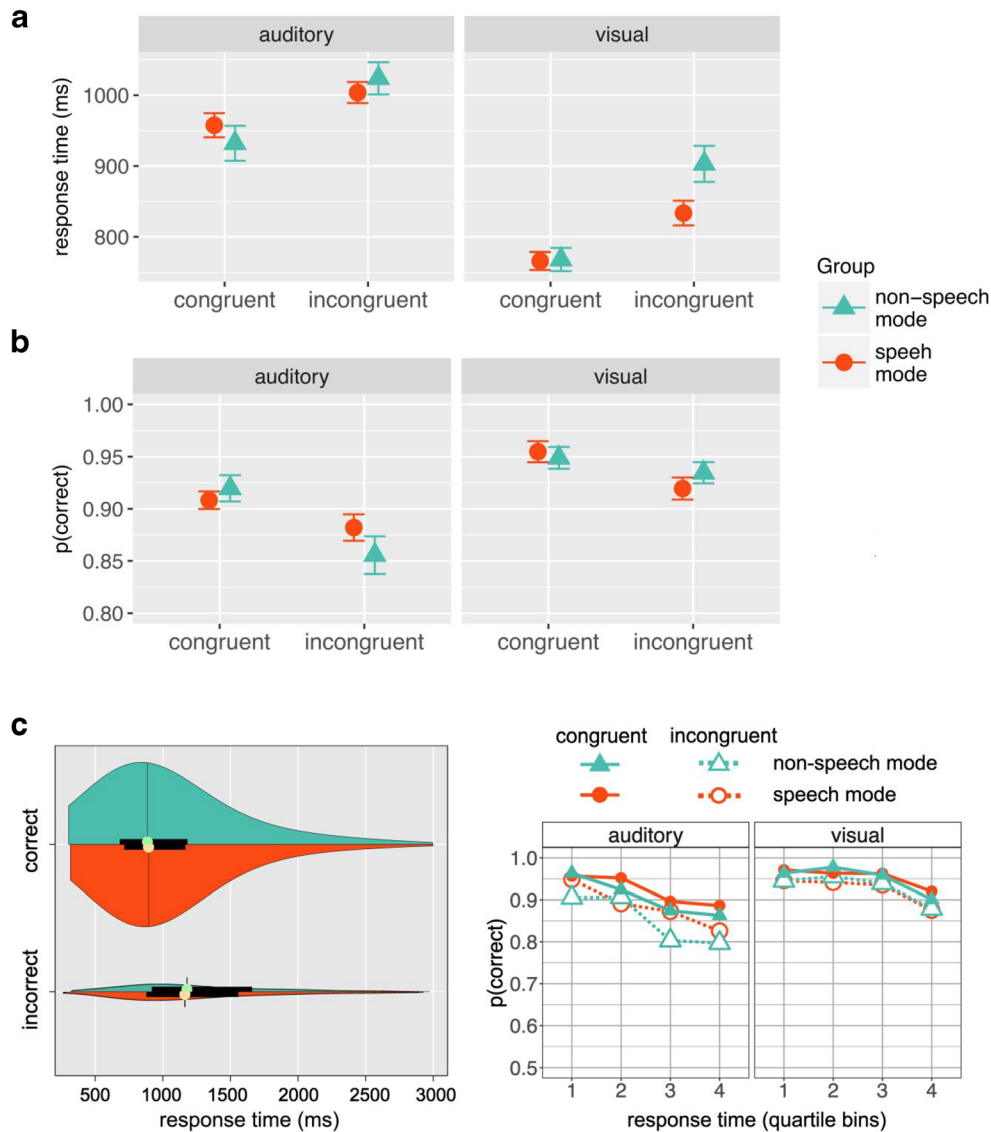
Both groups showed the expected sound-shape correspondence effects in both tasks, confirming that the boubu-kiki effect does not require that listeners perceive sounds as speech. In the IAT, correspondence effects were not affected by whether listeners were in the non-speech or in the speech-mode group. This provides no support for the hypothesis of an important role of speech-specific processing in the boubu-kiki effect. In the explicit CMM task, the correspondence effect was significant for both groups, but stronger for the speech-mode group, indicating that speech-specific processing did affect sound-shape correspondences. Thus, the latter task seems to tap into aspects of shape-sound correspondences that are not reflected in IAT performance. However, no specific hypothesis had been advanced regarding differences between tasks. Particularly, the present study was not designed to compare implicit versus explicit processing. Thus, this aspect of our results requires cautious interpretation.

The observed correspondence effects for SWS stimuli that were not heard as speech can be explained based on associations of visual contours with auditory stimulus attributes, rather than with speech-specific representations of articulatory gestures or abstract phonological units. Consistent with findings on non-verbal sound-shape correspondences (Adeli, Rouat, & Molotchnikoff, 2014; Liew, Lindborg, Rodrigues, & Styles, 2018; Marks, 1987; O’Boyle & Tarte, 1980; Parise & Spence, 2012; Walker et al., 2010), it has been conjectured that differences in the frequency content and waveform envelope are key to the boubu-kiki effect (Fort et al., 2015; Nielsen & Rendall, 2011). Regarding the *taketa-maluma* pair, vowel [e] sounds brighter than [u] due to its higher formant frequencies – the second formant frequency has been identified as a major contributor to sound-shape correspondences (Knoeferle et al., 2017); energy changes associated with voiceless obstruents [t, k] are sharper than those associated with sonorants [m, l] (see Fig. 1b and c). Indeed, without having syllables to speak about the stimuli, participants in the non-speech group reported hearing “taketa” as treble relative to

“maluma”. Of note, the present findings show that speech stimuli reduced to the three lower formant contours, which excludes the fundamental frequency and noise-related features, are sufficient for the boubu-kiki effect to occur. Further experimental research manipulating and controlling for multiple acoustic and visual features (as recently reported for sound-color correspondences; Anikin & Johansson, 2019) are necessary to refine the picture of the sensory dimensions involved in sound-shape correspondences.

Hearing SWS stimuli as speech did not increase the correspondence effect in the IAT, suggesting that speech-specific processing plays little or no role in sound-shape correspondences as assessed by this task. This is consistent with Parise and Spence’s (2012) interpretation of IAT results on five types of audiovisual correspondences, involving both speech and non-speech sounds, as reflecting a single automatic mechanism that deals with associations between auditory and visual features. Of course, one could entertain the less parsimonious possibility that distinct mechanisms underlie the statistically equivalent correspondence effects for the non-speech and speech-mode groups, and that a putative language-related mechanism alone could account for the effect in the latter group. Although this possibility cannot be ruled out, it finds no support in IAT results – no main effect or any interaction involving the factor Group was found. Not only did the groups not differ significantly in response time and accuracy, but they also both showed a similar decrease in accuracy with increasing response time. This also indicates that IAT performance was not appreciably affected by whether the participant had previously performed the sound localization task (non-speech group) or the pseudoword identification task (speech-mode group), and hence that SWS stimuli were as distinguishable for participants of one group as for participants of the other.

In the CMM task, hearing stimuli in speech mode seems to strengthen the grasp participants have on cross-modal compatibilities. Thus, the two tasks employed here seem to be differentially sensitive to distinct mechanisms underlying the boubu-kiki phenomenon. In the related but distinct field of intersensory integration, studies indicate that listening to

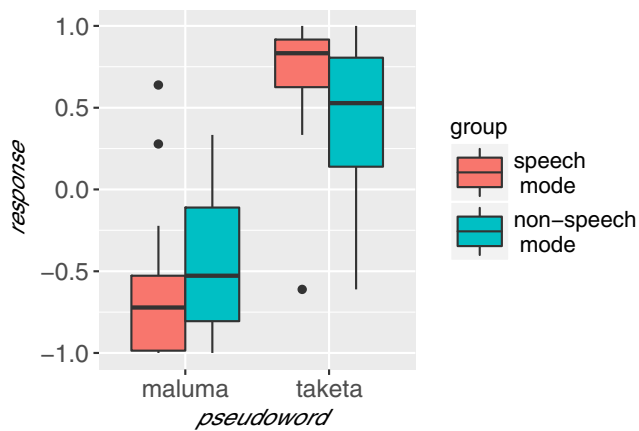


**Fig. 2** Implicit association task (IAT) performance for auditory and visual stimuli in congruent and incongruent blocks. **(a)** Mean response times. **(b)** Proportions of correct responses. The non-speech (green triangles) and the speech-mode (red circles) groups showed similar performance improvements in congruent compared to incongruent blocks. Bars represent standard errors after adjusting values for within-participant designs (Morey, 2008). **(c)** *Left panel*: violin plots showing kernel estimates of response time distributions for correct and incorrect responses in the non-

speech (green) and speech-mode (red) groups. Boxplots are shown in black. Circles represent the median. *Right panel*: proportion correct as a function of response time. For illustration purposes, response times were binned. For each participant, modality, and congruency level, bin 1 contains the 25% fastest responses; bin 2 contains the next 25% fastest responses, and so on. The left panel in (c) was created using the violplot R package (Adler, 2019). All other plots were generated via ggplot2 (Wickham, 2009)

SWS stimuli as speech is crucial for the integration of visual (lipread) and auditory information in sound identification tasks, but has no effect on visually enhanced detection of SWS in noise (Eskelund, Tuomainen, & Andersen, 2011) nor on judgments of synchrony and temporal order between SWS and lipread stimuli (Vroomen & Stekelenburg, 2011). This has been interpreted as evidence for two distinct mechanisms that improve speech intelligibility via audiovisual integration: one based on speech-specific content in auditory and visual signals, and a more general mechanism that exploits cross-modal covariation to enhance auditory signal-to-noise

ratio. Something analogous may occur with cross-modal correspondences between speech sounds and shapes – i.e., both speech-specific and general perceptual mechanisms may contribute to them. Possibly, while the IAT taps auditory-visual associations shared by cross-modal correspondences involving speech and non-speech sounds, CMM results reflect the latter associations combined with mappings between shapes and higher-level, speech-specific representations, which might include articulatory features (Maurer et al., 2006; Ramachandran & Hubbard, 2001), language-specific phonological units (Shang & Styles, 2017; Styles & Gawne, 2017),



**Fig. 3** Boxplots for responses of the speech-mode and non-speech groups in the explicit task. Positive values represent “spiky” responses; negative values represent “curved” responses. Both groups consistently associated pseudowords *maluma* and *taketa* with curved and spiky shapes, respectively. However, this association was more pronounced for the speech-mode group

and the corresponding orthographic forms (Cuskley, Simner, & Kirby, 2017). This is consistent with the idea that sound-symbolism effects in adults result from both pre-linguistic and language-related biases (Ozturk et al., 2013).

While the IAT provides indirect performance measures of task-irrelevant associations reflecting presumably automatic processes (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Parise & Spence, 2012), CMM relies on introspective reports. We speculate that the requirement to judge how well pseudowords and shapes match led participants of the speech-mode group to consider similarities based on speech-specific representations in addition to basic auditory qualities. However, CMM and IAT differ in many other ways. The IAT requires participants to keep the stimulus-response mapping for the current block in memory and respond correctly and quickly to each unimodal (either visual or auditory) stimulus. Feedback was provided to keep performance at reasonable levels and multiple trial blocks were necessary to assess correspondence effects. The CMM task is much less effortful; correspondences can be accessed directly in a few trials, each of which contain both visual and auditory stimuli; responses are not speeded and there is no sense in classifying them as correct/incorrect. Moreover, it is not clear how performing the IAT could affect CMM, which was always the last task to be performed in order to avoid directing attention to sound-shape correspondences before the IAT.

The present findings warrant future studies designed specifically to investigate which features of different tasks are associated with their sensitivity to different mechanisms underlying sound symbolism. To test whether the contribution of speech-specific processing is indeed contingent on the explicit assessment of correspondences, one should conceive a more comparable “implicit-explicit” pair of tasks. Also of interest would be a replication of the present study using the implicit,

speeded classification task used by Evans and Treisman (2010), which detects correspondence effects at the perceptual level, rather than at the level of response selection as does the IAT (Parise & Spence, 2012). Testing for different instances of sound symbolism under different task requirements can be of great help in elucidating the nature of the involved representations and processes.

**Acknowledgements** Both authors were supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Brazil. The authors also gratefully acknowledge Professors Barry C. Smith, Sarah Garfinkel, Vincent Hayward, and André J. Athab, as well as anonymous reviewers for thoughtful comments and suggestions.

## Compliance with ethical standards

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

**Open practices statement** The data used in this research were made available to reviewers and to journal editors.

## References

- Adeli, M., Rouat, J., & Molotchnikoff, S. (2014). Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in Human Neuroscience*, 8. <https://doi.org/10.3389/fnhum.2014.00352>
- Adler, D. (2019). violot: Violin Plot (Version 0.3.0). Retrieved from <https://CRAN.R-project.org/package=violot>
- Anikin, A., & Johansson, N. (2019). Implicit associations between individual properties of color and sound. *Attention, Perception, & Psychophysics*, 81(3), 764–777. <https://doi.org/10.3758/s13414-018-01639-7>
- Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, 53, 115–121. <https://doi.org/10.1016/j.neuropsychologia.2013.11.011>
- Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39), 10818–10823. <https://doi.org/10.1073/pnas.1605782113>
- Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (Version 5.3.41)[Software].
- Bremner, A., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. and Spence, C. (2013). Bouba and Kiki in Namibia? A remote culture makes similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition* 126, 165–172. <https://doi.org/10.1016/j.cognition.2012.09.007>
- Chen, Y.-C., Huang, P.-C., Woods, A., & Spence, C. (2016). When “Bouba” equals “Kiki”: Cultural commonalities and cultural differences in sound–shape correspondences. *Scientific Reports*, 6, 26681. <https://doi.org/10.1038/srep26681>
- Chow, H. M., & Ciaramitaro, V. (2019). What makes a shape “baba”? The shape features prioritized in sound–shape correspondence change with development. *Journal of Experimental Child Psychology*, 179, 73–89. <https://doi.org/10.1016/j.jecp.2018.10.005>



- Cuskley, C., Simner, J., & Kirby, S. (2017). Phonological and orthographic influences in the bouba-kiki effect. *Psychological Research*, 81(1), 119–130. <https://doi.org/10.1007/s00426-015-0709-2>
- D'Onofrio, A. (2014). Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm. *Language and Speech*, 57(3), 367–393. <https://doi.org/10.1177/0023830913507694>
- Darwin, C. (2003). Sine-wave speech produced automatically using a script for the PRAAT program. [http://www.lifesci.sussex.ac.uk/home/Chris\\_Darwin/SWS/](http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/SWS/) (Last viewed July, 2017)
- Davidson, D. J., & Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta Psychologica*, 144(1), 83–96. <https://doi.org/10.1016/j.actpsy.2013.04.016>
- Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *Neuroimage*, 24(1), 21–33. <https://doi.org/10.1016/j.neuroimage.2004.09.039>
- Deroy, O., & Spence, C. (2013). Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review*, 20(4), 643–664. <https://doi.org/10.3758/s13423-013-0387-2>
- Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: Dissociating identification and detection. *Experimental Brain Research*, 208(3), 447–457. <https://doi.org/10.1007/s00221-010-2495-9>
- Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, 10(1), 1–12. <https://doi.org/10.1167/10.1.6>
- Fort, M., Martin, A., & Peperkamp, S. (2015). Consonants are more important than vowels in the Bouba-kiki effect. *Language and Speech*, 58(2), 247–266. <https://doi.org/10.1177/0023830914534951>
- Fort, M., Weiß, A., Matin, A., & Peperkamp, S. (2013). Looking for the bouba-kiki effect in prelexical infants. In S. Ouni, F. Berthommier, & A. Jesse (Eds.), *Proceedings of the 12th international conference on auditory-visual speech processing* (pp. 71–76). Annecy France: Inria.
- Godoy, M., Duarte, A. C. V., Silva, F. L. F., Albano, G. F., Souza, R. J. P., & Silva, Y. U. A. P. M. (2018). The replication of takete-maluma effect in Brazilian Portuguese. *Revista do GELNE*, 20(1), 87–100. <https://doi.org/10.21680/1517-7874.2018v20n1ID13331>
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17–41. <https://doi.org/10.1037/a0015575>
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), 393–402. <https://doi.org/10.1038/nrn2113>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>
- Khoshkhou, S., Leonard, M. K., Mesgarani, N., & Chang, E. F. (2018). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. *Brain and Language*, 187, 83–91. <https://doi.org/10.1016/j.bandl.2018.01.007>
- Knoeferle, K., Li, J., Maggioni, E., & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, 7(1), 5562. <https://doi.org/10.1038/s41598-017-05965-y>
- Kohler, W. (1947). *Gestalt psychology* (1929). New York, NY: Liveright.
- Liew, K., Lindborg, P., Rodrigues, R., & Styles, S. J. (2018). Cross-modal perception of noise-in-music: Audiences generate spiky shapes in response to auditory roughness in a novel electroacoustic concert setting. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.00178>
- Lockwood, G., & Dingemans, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01246>
- Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3), 384. <https://doi.org/10.1037/0096-1523.13.3.384>
- Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science*, 9(3), 316–322. <https://doi.org/10.1111/j.1467-7687.2006.00495.x>
- McCormick, K., Kim, J., List, S., & Nygaard, L. C. (2015). Sound to Meaning Mappings in the Bouba-Kiki Effect. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society*. Austin, TX: Cognitive Science Society.
- Monahan, P. J. (2018). Phonological knowledge and speech comprehension. *Annual Review of Linguistics*, 4(1), 21–47. <https://doi.org/10.1146/annurev-linguistics-011817-045537>
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, 4(2), 61–64. <https://doi.org/10.20982/tqmp.04.2.p061>
- Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Experimentale*, 65(2), 115–124. <https://doi.org/10.1037/a0022268>
- Nielsen, A. K., & Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(2), 153–163. <https://doi.org/10.1037/a0030553>
- 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(6), 535–544. <https://doi.org/10.1007/BF01068115>
- Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound-shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114(2), 173–186. <https://doi.org/10.1016/j.jecp.2012.05.004>
- Parise, C., Spence, C. (2012). Audiovisual cross-modal correspondences and sound symbolism: A study using the implicit association task. *Experimental Brain Research*, 220 (3–4), 319–333. <https://doi.org/10.1007/s00221-012-3140-6>
- Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies*, 8(12), 3–34.
- Remez, R. E., Rubin, P. E., & Pisoni, D. B. (1983). Coding of the speech spectrum in three time-varying sinusoids. *Annals of the New York Academy of Sciences*, 405(1), 485–489. <https://doi.org/10.1111/j.1749-6632.1983.tb31663.x>
- Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212(4497), 947–949. <https://doi.org/10.1126/science.7233191>
- Remez, R. E., & Thomas, E. F. (2013). Early recognition of speech. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(2), 213–223. <https://doi.org/10.1002/wcs.1213>
- Schellinger, S. K., Munson, B., & Edwards, J. (2017). Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods. *Clinical Linguistics & Phonetics*, 31(1), 80–103. <https://doi.org/10.1080/02699206.2016.1205665>
- Shang, N., & Styles, S. J. (2017). Is a high tone pointy? Speakers of different languages match Mandarin Chinese tones to visual shapes differently. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.02139>

- Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, 25(5), 1619–1643. <https://doi.org/10.3758/s13423-017-1361-1>
- Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4), 971–995. <https://doi.org/10.3758/s13414-010-0073-7>
- Styles, S. J., & Gawne, L. (2017). When does Maluma/Takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *I-Perception*, 8(4), 2041669517724807. <https://doi.org/10.1177/2041669517724807>
- Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, 118(1), 75–83. <https://doi.org/10.1016/j.cognition.2010.10.002>
- 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(1), 21–25. <https://doi.org/10.1177/0956797609354734>
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.
- Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143–146). ACM.
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.