Seeing voices: the role of multimodal cues in vocal learning

Varkevisser, J.M.

Chapter 5

Song learning from a singing robotic bird versus from audio-only song playback in young zebra finches

Judith Varkevisser, Ralph Simon, Ezequiel Mendoza, Constance Scharff, Wouter Halfwerk & Katharina Riebel

Abstract
Bird song is one of the best-studied examples of a vocally learned signal in non-human animals. In several songbird species, song learning success is lower in tutees exposed to playback of tutor song via loudspeakers ('tape tutoring') than in tutees raised with a singing conspecific ('live tutoring'). This is generally hypothesized to result from a lack of social interactions between tutor and tutee in the tape tutoring setting. However, tape tutoring only offers unimodal, auditory song exposure whereas birdsong is a multimodal signal. Song production is accompanied by visual cues such as head, beak and throat movements. The aim of the present study was to test whether song-specific visual cues (rather than social cues) have a facilitating effect on song learning in zebra finches, *Taenyiopygia guttata*. We investigated song learning in tutees raised with audio playback only (while housed alone or with a female companion) or with audio playback combined with a robotic zebra finch (RoboFinch) that in one group produced synchronized beak and head movements and in another group only moved after playbacks, so that its movements were decoupled from the playback. We used three different similarity assessment methods to determine the similarity between tutor and tutee song. However, none of these methods detected a significant treatment effect on song similarity. We thus did not find a facilitating effect of multimodal cues presented through a RoboFinch on zebra finch song copying. When comparing adult song, we found that tutees that had only auditorily been exposed to tutor song while housed with a social companion sang with a higher between-motif stereotypy than the tutees that had been housed solitarily throughout song tutoring, suggesting that having a social companion positively affects song development. Future studies should investigate how exposure frequency and level of interaction are potential additional modifiers on song development and song learning from the RoboFinch and investigate whether the improved performance in socially-raised tutees results from higher motivation to sing.

Introduction
Human speech and birdsong are communication signals that individuals learn early in life by exposure to the vocalizations of adult conspecifics (Bolhuis, Okanoya, & Scharff, 2010; Doupe & Kuhl, 1999). For both speech and birdsong it is unclear whether, and to what extent, learning is improved if individuals are exposed to the visual cues accompanying the production of vocalizations, such as lip movements in speech and beak movements in bird song (speech: Kuhl & Meltzoff 1982; Lewkowicz & Hansen-Tift 2012; Teinonen, Aslin, Alku, & Csibra 2008; Tenenbaum, Sobel, Sheinkopf, Malle, & Morgan 2015, birdsong: Beecher & Burt 2004; Derégnaucourt 2011; Slater, Eales, & Clayton 1988). Sev-

eral observational studies in humans suggest that exposure to these visual cues might affect vocal learning (e.g. Lewkowicz & Hansen-Tift, 2012; Young, Merin, Rogers, & Ozonoff, 2009). Birdsong development provides a model system that can be used to experimentally investigate the effect of exposure to production-related visual cues on the vocal learning process (Brainard & Doupe, 2002; Doupe & Kuhl, 1999; Goldstein et al., 2003).

A popular experimental tutoring method in the study of bird song learning is tape tutoring, i.e. playing pre-recorded tutor song to young birds via loudspeakers, as tape tutoring enables researchers to standardize and control the song that birds are exposed to (Catchpole & Slater, 2003). Tutees that are tape tutored, however, are only exposed to song auditorily, so unimodally, while birdsong is actually a multimodal signal, because the production of birdsong is accompanied by visual components such as beak, head and body movements. These visual cues might play a role in the song learning process (see Chapter 2), as signals with components in multiple modalities are easier to detect and remember than single component signals (reviewed in Rowe, 1999). Moreover, improved learning of auditory signals if they are paired with visual stimulation has been demonstrated in several bird species: in chicks in the context of filial imprinting (van Kampen & Bolhuis, 1991; van Kampen & Bolhuis, 1993) and in nightingales in the context of song learning (Hultsch et al., 1999). There are several songbird species that copy less song from a tape tutor than from a live conspecific tutor (reviewed in Baptista & Gaunt, 1997; Soma, 2011). This difference in song copying success is usually ascribed to a lack of social interaction with the tutor in the tape tutoring condition (Baptista & Petrinovich, 1986; Catchpole & Slater, 1995; Slater, Eales, & Clayton, 1988). It is as yet unknown, however, whether a lack of multimodal cues also plays a role in the lower amount of song copying in tape tutoring paradigms.

To investigate whether multi- compared to unimodal song exposure has a facilitating effect on song learning in songbirds, a tutoring method is required where not only the auditory, but also the visual component of song production can be standardized and controlled. One option is to combine a tape tutor with a video recording of a singing tutor. Using videos for tutoring birds, however, can be difficult as standard video systems are designed for human vision, which differs in several dimensions from avian vision (Cuthill et al., 2000; Fleishman & Endler, 2000; Oliveira et al., 2000). In a previous study, pairing auditory playback with a video of the singing tutor led to more stimulus engagement in zebra finch tutees, but not to enhanced song learning (Varkevisser et al., 2021). Although these videos were adjusted as much as possible to the zebra finch

visual system, it might be that specific video properties, such as the brightness of the videos, or the two-dimensionality of the videos affected the salience of the visual cues accompanying song production, and thereby the effect they might have on song learning success. Using a robotic bird can overcome this problem, as it is a three-dimensional model of a tutor, where experimenters can also control and manipulate the auditory and visual channel independently. Technical advancements enable researchers to create realistic robotic animals that can produce fast movements resembling those of live animals. Previous studies have demonstrated that robotic animals are valuable tools to study animal communication (e.g. Landgraf et al., 2008; Taylor et al., 2008). Robotic birds have already been applied successfully to test the potential importance of multi- over unimodal signalling in different contexts such as territorial defence (Anderson et al., 2013; Ręk & Magrath, 2016), courtship (Patricelli et al., 2002) and spatial orientation (Butler et al., 2017). These studies show the acceptance of a robot model by adult birds, which suggests that using a robot in developmental studies might provide an effective tutoring method where both the auditory and visual component of song production can be controlled.

This study used a robotic bird to test the effect of multi- compared to unimodal song exposure on song learning success in zebra finches, *Taeniopygia guttata*. The zebra finch, the main animal model in studies on vocal learning (Griffith & Buchanan, 2010; Mello, 2014), is one of the species that copies less song from tape tutors than from live tutors (Derégnaucourt, Poirier, van der Kant, & van der Linden, 2013; reviewed in Derégnaucourt, 2011; Slater, Eales, & Clayton, 1988). The production of zebra finch song is accompanied by beak, throat and body movements (Goller, Mallinckrodt, & Torti, 2004; Ullrich, Norton, & Scharff, 2016; Williams, 2001). It is as yet unclear whether exposure to these movements has an effect on zebra finch song learning. Previous studies that presented a visual stimulus before, during or after the auditory presentation of tutor song did not find an effect of the visual stimulation on zebra finch song learning (Bolhuis, van Mil, & Houx, 1999; Houx & ten Cate, 1999). The visual stimulus used in these studies was a non-moving taxidermic mount of an adult zebra finch. Visual stimulation that moves in synchrony with the presented sound, however, might be more salient than non-moving visual stimulation (Bolhuis et al., 1999). This poses quite a challenge, as zebra finches produce rapid changes in beak aperture during song production (Goller et al., 2004; Williams, 2001). Recent technological advancements, however, make it possible to create a realistically moving robotic model of a singing zebra finch (Simon et al., 2019).

In this study, we used a specifically developed 3-D printed robotic zebra finch with exact beak movements (RoboFinch: Simon et al., 2019). We compared song learning in young zebra finches that had either only auditory tutor song exposure or auditory exposure accompanied by the RoboFinch that produced time aligned beak and head movements corresponding to the auditorily presented song. To control for any effect that having a moving RoboFinch next to the cage might have on song learning, we also raised birds in a control condition with a complete mismatch between the auditory and visual stimulus. In this condition, the beak and head of the RoboFinch started to move after auditory song presentation had finished. In previous studies, tape tutored birds were often raised in social isolation, which might have negatively affected the juvenile's welfare and motivation for song learning and might also have contributed to the difference in song learning success between live tutored and tape tutored birds (Chapter 2). To find out how growing up in social isolation versus with a social companion affects song learning, we also included a condition in which tutees received auditory tutor song exposure only, but were housed together with an unrelated female peer. We hypothesized that the visual cues produced by the RoboFinch and presented time aligned with the auditory song playback would facilitate song learning and lead to a higher amount of tutor song copying than the other tutoring conditions.

Methods

Subjects and housing

Subjects for this study were 45 juvenile males and 9 juvenile females from the domesticated wild-type zebra finches breeding colony at Leiden University. Birds were raised and housed in breeding cages (100 x 50 x 40 cm) with their parents and siblings until 20 days post-hatching (dph, age calculated as the median hatching date of all chicks in the nest) when the father was removed. Subjects stayed with their mother and siblings from 20 to 35 dph in their home cage. All breeding cages were located in a large breeding room with multiple pairs breeding in two long stacks of cages along the two long walls. At all times, other birds could be heard and birds 2.40 m across on the opposite side of the aisle could also be seen. At 35 dph, tutees were moved into cages in sound attenuated rooms (125 x 300 x 240 cm) for song tutoring (see details below). The sound-attenuated rooms had one-way mirrors in the door, which made observation and daily welfare monitoring possible without disturbing the young birds. When the tutees reached 65 dph, they were moved to a recording cage (see below). After recording at 65 dph, tutees were housed in an individual cage or with their female companion (if they had been raised in the audio+female treatment, see below) in separate cages (150 x 40 x 50 cm) located in a room

with multiple birds, until song of the male tutees was recorded after 100 dph (see below). Throughout, birds were housed on a 13.5/10.5h light/dark cycle (with 30 minute dusk and dawn simulations), at 20-22 ºC and 45-65 % humidity. Birds had *ad libitum* access to a commercial tropical seed mixture (Beyers, Belgium), cuttlebone, grit and drinking water. This diet was supplemented three times a week with hardboiled eggs and once a week with germinated tropical seeds, vegetables and fruit.

Song tutoring

For this study, a song was defined as one or several motifs separated from other sounds by more than two seconds of silence or when a motif was starting with additional introductory notes (Sossinka & Böhner, 1980). Motifs were defined as the individual-specific repeated syllable sequence in a song, and syllables as sounds separated from other sounds by at least 5 milliseconds of silence.

Male tutees were tutored in one of four different tutoring treatments (see Figure 1): (1) song playback and a RoboFinch (robotic zebra finch, Simon et al., 2019) positioned next to the cage that produced beak and head movements time-aligned with the presented sound ("Robot"), (2) song playback and a RoboFinch positioned next to the cage that only started moving after the auditory song presentation session had finished ("Robot mismatch"), (3) song playback only ("audio"), (4) song playback and an unrelated age-matched female housed in the same cage as the male tutee ("audio+female").

The same tutor song was presented to four male tutees, each in a different tutoring treatment (Robot, Robot mismatch, audio and audio+female). Together, these treatments formed one 'tutor group'. We used song from six different tutors, and each tutor was used for two different tutor groups. Due to the limited number of nine experimental set-ups available per round, tutees were tutored in five consecutive rounds. In the first two rounds, no birds were tutored in the audio+female treatment, so per round we tutored three tutor groups with three different treatments (Robot, Robot mismatch and audio) at the same time. In the last three rounds, per round we raised two tutor groups with all four different treatments as well as a tutee in the audio+female treatment belonging to one of the tutor groups tutored during the first two rounds. In the end, a total of 9 tutees had been raised in the audio+female treatment and 12 tutees in the other three treatments. Within one tutor group, wherever possible, all male tutees originated from the same nest (all 4 male siblings: 4/12 tutor groups, all 3 male siblings: 2/12 tutor groups, 3 male siblings and one additional male: 3/12 tutor groups, 2 male siblings and 2 additional males: 2/12 tutor groups, 2 male

siblings and 1 additional male: 1/12 tutor groups). If it was not possible to only have tutees from the same nest in one tutor group, we used unrelated chicks and made sure that the treatment that the unrelated chicks received differed across tutor groups.

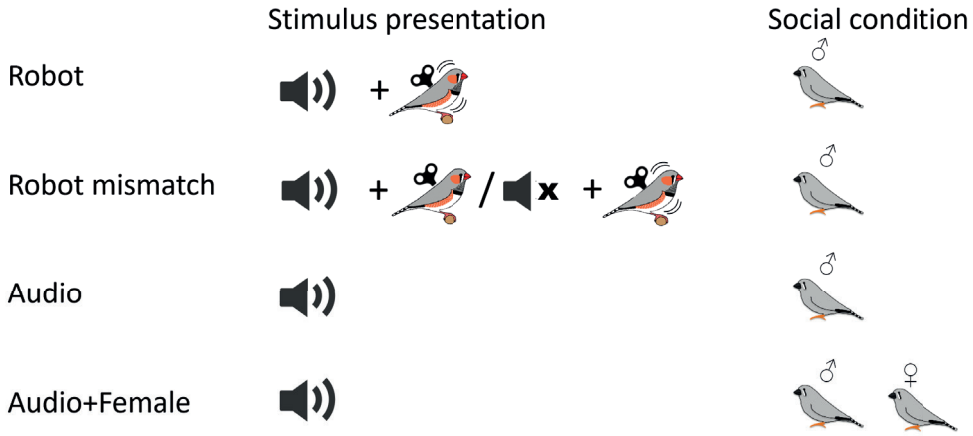

Figure 1. Schematic overview of the four different tutoring treatments. The normal loudspeaker symbol represents auditory song playback, while the loudspeaker symbol with a cross represents a situation without auditory song playback. Lines next to the RoboFinch icon (zebra finch with winding key on its back) indicate beak and head movements of the RoboFinch, In the 'social condition' column, the juvenile male icon indicates that male tutees were housed solitarily, while a male and female icon indicates that male tutees had an unrelated female peer as a social companion.

For 30 days, tutees received 6 tutoring sessions daily at 8:15 (half an hour after lights on), 9:15, 10:15, 12:15, 14:15 and 16:15. Each tutoring session lasted 30 minutes. During a tutoring session, three different types of files were played: songs, calls and head movements. The song files consisted of undirected tutor song of between 3 and 5 motifs. For each tutor, there were 3 different song files, each accompanied by the specific corresponding head and beak movements (see stimulus preparation). The call files consisted of one or two calls produced by the tutor, accompanied by the corresponding beak and head movements (see stimulus preparation). There were two different call files for each tutor and these files lasted 4 seconds. The head movement files did not contain sound, but just consisted of head movements of the RoboFinch. There were two different head movement files for each tutor and these files lasted 10 seconds. During a tutoring session, these three type of files were presented according to a pre-programmed daily schedule in which we made sure that birds were exposed to 16 songs during each of the morning sessions (8:15, 9:15 and 10:15), and 7 songs during each of the afternoon sessions (12:15, 14:15 and 16:15),

with a total of 207 to 345 motifs presented daily. This was based on song rates reported for live tutors (Böhner, 1983). In the schedule, songs often occurred in bouts of between 2 and 4 songs. In between song presentations, we randomly added head movement and call files to the schedule (the schedule that was used can be found in the appendix, Table A1). In the Robot mismatch condition, we created a complete mismatch between the auditory and visual stimulation (movement of the RoboFinch) to avoid the possibility of multisensory temporal integration or alerting effects (demonstrated in starlings: Feenders, Kato, Borzeszkowski, & Klump, 2017). In this treatment, audio files were played during the tutoring session, followed by half an hour of exposure to the movements corresponding to the sounds played during the tutoring session.

Stimulus preparation

Stimuli consisted of undirected song recordings of six adult male zebra finches from the colony (3 songs per tutor, 18 songs in total). For these recordings, a male was placed singly in a recording cage (76 x 45 x 45 cm) placed on a table in a sound-attenuated room in the afternoon of the day before recording for acclimation. The next morning, the male was recorded between 08:00 and 11:00, or until we had three song recordings. After this, the male was returned to its home cage. The recording cage had a clear Plexiglas window in the middle of the front side of the cage. This cage was placed on a table in a sound attenuated room. Only one cross perch was placed in the middle of the cage so that the bird would always be in focus on the camera. Audio recordings were made with a Sennheiser MKH40 microphone (Wedemark, Germany) hanging 50 cm above the perch in the recording cage. The microphone was connected to a TASCAM DR-100MKiii recorder (TEAC Corp., Los Angeles, USA). Audio was recorded with a sampling rate of 96 kHz and 16-bit resolution. Video recordings were made with a Casio high speed camera (EX-ZR3600, 120 fps, 12x optical zoom, Tokyo, Japan) through Plexiglas in the door of the sound attenuated room. A signal bell (70027 Heidemann, Willich, Germany), which was sound attenuated to not disturb the birds, was attached to the front side of the recording cage above the Plexiglas window and could be triggered from outside the sound attenuated room. The bell produced a short, impulse like audio signal and it was clearly visible on the video when the clapper touched the bell, which was later used to synchronize the audio and video recordings during stimulus preparations. The camera could record 120 fps videos for up to 12 minutes and at the start of each recording, we triggered the bell. Audio files were filtered with a band-stop filter from 0 to 425 Hz using Praat (version 6.0.19, Boersma & Weenink, 2008). Audio and video were synchronized with Vegas Pro (version 14.0, Magix, Berlin, Germany). For each tutor, three songs with introducto-

ry notes followed by 3 to 5 motifs were cut out of the recordings (mean song duration ± SD = 4.2 ± 1.2 seconds, mean number of motive repetitions ± SD = 3.9 ± 0.8).

We used the software Tracker (open source physics, physlets.org) to deduce movement files of the birds from the 120 fps videos. In the program, we marked forehead, the tip of the upper beak and the tip of the lower beak to analyse head movement and beak opening over time. Using this data, we created head and beak movement files which could be used to move the robots' beaks and heads. As the movements of the RobotFinch caused some clicking sounds that might have slightly interfered with the song presentation, we recorded the clicking sounds occurring with each of the tutor songs, synchronized and mixed these into the audio files. We used these files for the conditions where there was no moving robot during song presentation, so where otherwise there would not have been mechanical sounds during song presentation (i.e. Robot mismatch, audio and audio+female conditions). As we only realised that the robot made these sounds when the experiments had already started, we only corrected for the clicking sounds by presenting these edited audio files with the extra mechanical sounds for half of the tutor groups. Therefore, each tutor song was presented to one tutor group without clicking sounds and to one tutor group with clicking sounds. After creating the audio stimuli, we played them back through the loudspeaker next to the experimental set-up (see below) and recorded them with a microphone (MKH40, Sennheiser, Wedemark, Germany) positioned inside the cage. Using Praat software, we visually compared the power spectra (Fast Fourier transform) of these recordings with the power spectra of the original stimuli and did not observe any systematic differences.

Experimental set-up

The experimental set-up consisted of a cage (70 x 60 x 45 cm, the same cage as used in Varkevisser et al. (2021)) placed on a table in a sound attenuated room. The cage had three sides of meshed wire and one side of black plastic. A window (20 x 15 cm) was cut out of the plastic and covered with meshed wire. A loudspeaker (Blaupunkt, CB4500, Hildesheim, Germany) was positioned behind the meshed wire window at 18 cm distance. In front of the loudspeaker, a panel covered in black loudspeaker cloth was positioned so that the loudspeaker was not visible for the tutee birds. Sound was played-back with a peak amplitude of 74 dB (Fast, A, re 20 µPa, SL-451, Voltcraft, Conrad Electronic SE, Hirschau, Germany) at the perch closest to the meshed wire window. A webcam (Renkforce RF-4805778, Conrad, Hirschau, Germany) was installed next to the cage to record the tutees' behaviour in the cage. In the two robot

conditions (Robot and Robot mismatch), a RoboFinch (Simon et al., 2019) was positioned in front of this panel (see Figure 2).

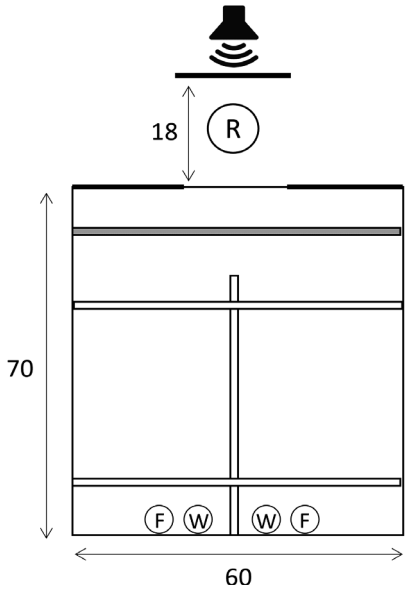

Figure 2. Schematic top view of the experimental set-up with perches. R = location of the RoboFinch in the Robot and Robot mismatch treatment, F = food, W = water. The loudspeaker was placed behind a panel covered in black loudspeaker cloth. All measurements are in cm.

RoboFinch

The RoboFinch is a realistic 3D printed, coloured, plastic model of a zebra finch (for details, see Simon et al., in prep). The beak and head of the Robo-Finch can move and the body can rotate. The latter was not used in the current study. The shape of the robotic finch was based on a 3D scan (handheld 3D scanner Eva, Artec3D, Luxembourg, Luxembourg) of a taxidermic model of an adult male zebra finch. The beak was scanned (ATOS 5X, gom, Braunschweig, Germany) with high resolution from a prepared skull. These scans were combined in the program Catia V5R20 (Dassault Systèmes), which was also used for the implementation of the inner mechanics. We printed the RoboFinch with stereolithography 3D printing (Form 2, Formlabs, Somerville, Massachusetts, US), which uses a laser to cure solid isotropic parts from a liquid photopolymer resin (Grey Pro, Formlabs Resin). The movement of the head and beak was controlled by coils from dismantled DigiBirds (Silverlit Toys Manufactory, Hongkong, China). These coils are cost-effective, small and allow fast movements up to 100 Hz. The coils were controlled via a custom build controller board, which was based on an Arduino board (Adafruit 3405,

Mouser electronics, Germany). The stepping motor (Nema 17 Bipolar Stepper Motor) was controlled via a Pololu stepping motor control. The Arduino and the stepping motor control were connected to a small desktop PC (Intel NUC i5) and controlled via a custom build LabView (National Instruments, Austin, Texas, US) Program. The program also scheduled stimulus presentation including audio playback.

The 3D-printed models were hand painted with acrylic paints (Citadel Colours Games Workshop, London, England, see Simon et al., in prep.). We found life-like colours by comparing the paints or paints mixtures with the results of spectrometer measurements of the plumage of the zebra finches. We focused on 4 colour patches: the red beak, the orange/red cheeks, the brownish pattern on the side beneath the wings and the greyish back side. We measured these patches for six male zebra finches using dead birds that were directly frozen after they had been sacrificed for other purposes. For each bird we took 6 measurements of the relative radiance of each colour patch with a Flame spectrometer (QR400-7-SR-BX reflection probe and a DH-2000-BAL balanced UV-VIS light source, spectralon white standard, all from Ocean Insight (Orlando, FL, USA)) and compared the spectra to the ones of the colored 3D models. See Appendix, Fig A1 for a comparison.

Song recording tutees

Song recordings of the male tutees took place in a recording cage (76 x 45 x 45 cm) in a sound-attenuated recording room (125 x 300 x 240 cm) following the methods described in Varkevisser et al. (2021). Recordings were made continuously during the next morning with a Sennheiser MKH40 microphone (Wedemark, Germany) connected to a TASCAM DR-100MKiii recorder (96 kHz sampling rate, 16-bit resolution), hanging at 50 cm above the perch in the recording cage. After a recording session, birds were moved back to the experimental set-up. Tutees were recorded twice: once at 65 dph (X ± SE: 64.9 ± 0.9) and once as young adults after 100 dph (X ± SE: 116.1 ± 10.8). In many tutees, the song that was recorded at 65 days post-hatching was still too variable to recognize syllables and motifs. All analyses were therefore conducted on the song recordings made after 100 dph. One male tutee died before we could record his song after 100 dph, leaving song of 44 male tutees for the song analysis.

Song analysis

The song analysis method and parameters are identical to Varkevisser et al. (2021, see Table 1 in Chapter 4). Briefly, for song selection and sound editing,

we used spectrograms calculated with the Praat-software (fast Fourier transformations with 1000 time and 250 frequency steps, 0.005s window length, dynamic range 55 dB, Gaussian window, Praat v. 6.0.19, Boersma & Weenink, 2008). All songs from the recording sessions' audio files were edited into single files and saved into one folder per male. From this folder, we randomly selected twenty songs with custom-written software by Niklas J. Tralles and used this sample to calculate sequence linearity and consistency (Scharff & Nottebohm 1991). Sequence linearity was calculated by dividing the number of different syllables by the number of different transitions between syllables in a song. This indicates how stereotyped syllables are ordered in a song, with more stereotyped songs yielding higher scores. Consistency was determined by first noting all transitions in the twenty songs. For each syllable, the typical transition was then determined by looking at the most frequently encountered transition from this syllable. The total number of occurrences of typical transitions was then divided by the total number of transitions encountered in the twenty randomly selected songs. Again, more stereotyped songs receive a higher score.

We also used the sample of twenty songs to identify a tutee's 'typical' (most frequently observed) and 'full' motif (the motif with the highest number of different syllables) within this sample. We determined the number of unique syllables in the typical motif by visually inspecting the spectrograms in Praat. The full motifs were used for the human observer similarity scoring and to determine the total number of syllables in the tutee's repertoire. For each tutee, we labelled different syllables with different letters (see Figure 3). From the twenty songs, we also randomly selected a new smaller subsample consisting of ten songs. We used a random number generator (http://www.random.org) to randomly select one motif from each of these ten songs. We cut these motifs from the recordings, band stop filtered them (0 to 420 Hz) and normalised them (with the 'scale peak' function in Praat). Introductory notes that did not occur with every repetition of the motif were not considered part of the motif and were cut off before proceeding with the analyses. These ten motifs were used for the SAP and Luscinia similarity and stereotypy scores (see below).

To allow comparison with earlier studies of zebra finch learning that mostly used either human observers (e.g. Bolhuis et al., 1999; Houx & ten Cate, 1999a) or automated methods such as Sound Analysis Pro (SAP, Tchernichovski, Nottebohm, Ho, Pesaran, & Mitra 2000) or Luscinia (Lachlan et al., 2010) and with our two other experiments comparing uni- versus multimodal tutoring, song similarity was assessed in exactly the same way as described in Varkevisser et al. (2021), and in Chapter 3 and 4. Briefly, for the human ratings of similarity,

three independent observers were asked to indicate for each syllable in the tutee's repertoire, which syllable of a model's motif it resembled most and to then indicate the degree of similarity on a four-step scale (0 = 'no similarity at all', 1 = 'slight similarity', 2 = 'moderate similarity' and 3 = 'very strong similarity'). Each tutee was compared with two models: the actual tutor and an unfamiliar control model, which was the tutor of another group. Observers were blind to the treatment groups that the tutees belonged to and to which model song belonged to the tutor and which model song belonged to the control male. We calculated repeatability with a one-way ANOVA (following Lessells & Boag 1987) with the similarity score as the dependent variable and tutee ID as factor. The repeatability estimates of the normalized scores of the three observers was high (Tutor-Tutee: $F_{3,41} = 10.16$, $p < 0.01$, $r \pm SE = 0.75 \pm 0.06$, Tutee-Tutor: $F_{3,41} = 8.00$, $p < 0.01$, $r \pm SE = 0.70 \pm 0.05$). For the analyses, we used the total sums of similarity scores of all three observers in relation to the potential maximum score a bird could have received from three observers. We assessed similarity in two ways: (1) the proportion and similarity of the model's syllables copied by the tutee ("similarity score model-tutee") and (2) the proportion and similarity of the tutee's syllables shared with the model ("similarity score tutee-model"). For the model-tutee comparison, for each model syllable, the ID and similarity score of the tutee syllable that received the highest score was noted, and these scores were summed.

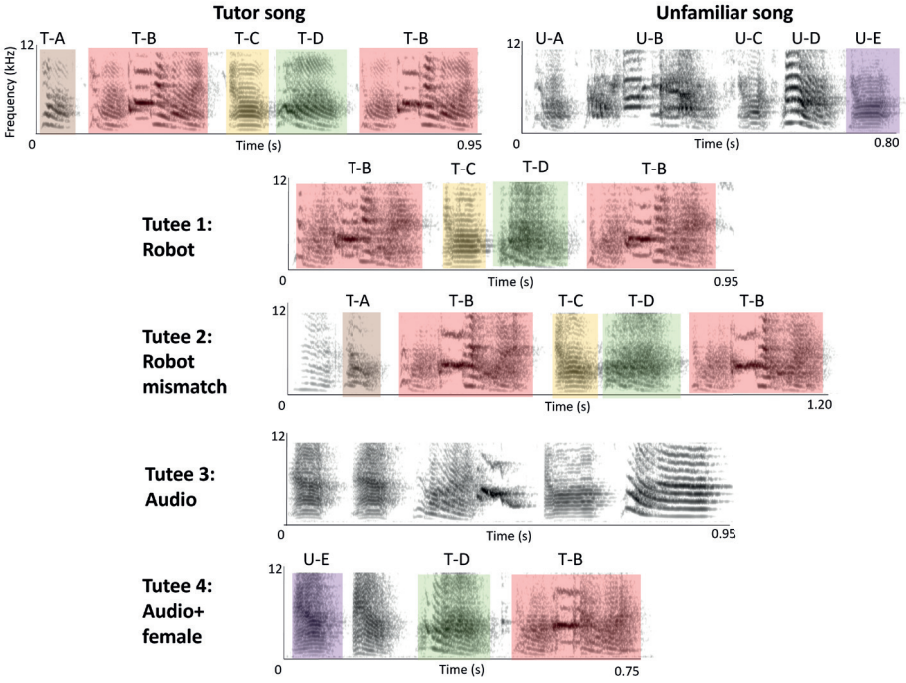

Figure 3. Spectrograms of full motif of tutor, unfamiliar full motif of another adult male and three tutees from one tutor group. Letters above spectrograms of tutor and unfamiliar song indicate how syllables were labelled with letters for further analyses. Human observer similarity between tutors and tutees was scored on a scale from 0 to 3. Syllables marked with the same colour had a total similarity score of 4 or higher when the similarity scores of all three observers for this comparison were summed up.

For the automated song comparisons, we compared each of the 10 randomly selected motifs of a tutee to each of 10 randomly selected motifs of its tutor using both Luscinia (Lachlan, Verhagen, Peters, & ten Cate, 2010) and Sound Analysis Pro (MxN comparison, default settings tuned for zebra finch, per tutor-tutee pair amplitude thresholds were adjusted for correct syllable segmentation, version 2011.104). In Sound Analysis Pro, for each possible comparison, we calculated the asymmetric similarity score for the tutor to tutee comparison (SAP similarity score tutor-tutee), which indicates the percent of sounds in the tutor's song that are observed in the tutee's song, as well as for the tutee to tutor comparison (SAP similarity score tutee-tutor), which indicates the percent of sounds in the tutee's song observed in the tutor's song. We used the median value of these scores as the measure of similarity (henceforth 'SAP similarity score'). In Luscinia, we chose the features 'mean frequency', 'fundamental frequency', 'fundamental frequency change' and 'time' for the acoustic distance calculations (following Lachlan, van Heijningen, ter Haar, & ten Cate 2016). The output of the DTW analysis is a distance measure between 0 and 1 for all possible pairs of motifs. In contrast to the human observer and SAP similarity scores, this is a symmetric score, so there is no difference between a model to tutee or tutee to model comparison. We used the median distance score for each tutee-model pair, and transformed it into a similarity score by calculating 1-distance score (henceforth 'Luscinia similarity score'), so that, like with the other scores, a higher score indicates a higher similarity. As a measure of song stereotypy, we also compared the ten tutee motifs to each other in Sound Analysis Pro and Luscinia. We used the same settings for this comparison as for the tutor to tutee comparisons. In Sound Analysis Pro, we calculated the median of the symmetric similarity score for the comparison of the ten tutee motifs. This will be referred to as the 'SAP stereotypy score'. In Luscinia, we used the median distance score for the comparison of the ten tutee motifs and then calculated 1- this distance score, again so that a higher score indicates a higher similarity. This score will be referred to as the 'Luscinia stereotypy score'.

Statistical analysis

RStudio (R: version 3.5.1) was used for all statistical analyses. We used linear mixed-effects models (LMMs) to test whether treatment groups differed in

linearity, consistency, the human observer, SAP and Luscinia scores and the number of unique syllables in the tutee's motif. Human observer, SAP and Luscinia scores were arcsine square root transformed prior to this analysis to meet model assumptions. To test whether treatment groups differed in the total number of syllables in the tutee's motif, generalized linear mixed-effect models (GLMMs) with a Poisson distribution and log-link function were used (package *lme4*: Bates, Mächler, Bolker, & Walker, 2014). 'Tutor' (the 6 different tutor IDs) was included as random factor in all models. We used ANOVA's to compare the null model with only the random factor to the model with 'treatment' (Robot, Robot mismatch, Audio or Audio+female) as a fixed effect. To test whether tutees had a higher score for human observer similarity with the song of the tutor than with the unfamiliar song of another male, we built LMMs and tested whether adding 'song model' (tutor or unfamiliar)' as fixed factor significantly improved the null models (with 'Tutor' and 'Bird ID' as random factors). For all models, a Shapiro-Wilk test was used to test whether the models' residuals followed a normal distribution. Post-hoc tests with Tukey adjustment for multiple comparisons were performed for between treatment comparisons (package *emmeans* Lenth, Singmann, Love, Buerkner, & Herve, 2018).

Ethics statement

Following European and national law, all procedures were reviewed and approved by the Leiden University Committee for animal experimentation, Leiden University Animal Welfare Body and the Centrale Commissie voor Dierproeven (CCD) of the Netherlands (permit number AVD1060020186606).

Results

Song structure and performance

The song structure and performance parameters (total number of syllables, number of unique syllables, linearity and consistency) did not differ between the treatment groups (models including 'treatment' were not significantly better than null models, see Table 1 and 2).

Table 1. Mean values of song structure and performance parameters in the song of the tutors and tutees. The three rightmost columns give the statistical details from the ANOVA that was used to compare the null model and the model including 'treatment' as a fixed effect.

	Tutor[1]	Robot	Robot mismatch	Audio	Audio+ female	ANOVA		
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	N	χ2	p
Total # syllables	*5.8 ± 1.7*	4.6 ± 2.3	4.6 ± 2.0	4.6 ± 1.7	4.3 ± 1.9	44	0.12	0.99
# unique syllables	*5.2 ± 1.5*	4.3 ± 0.9	4.6 ± 1.9	5.3 ± 1.6	4.6 ± 1.7	44	1.19	0.76
Linearity	*0.43 ± 0.06*	0.43 ± 0.12	0.44 ± 0.06	0.41 ± 0.10	0.46 ± 0.11	44	1.66	0.65
Consistency	*0.93 ± 0.04*	0.89 ± 0.12	0.89 ± 0.11	0.83 ± 0.14	0.90 ± 0.07	44	2.46	0.48

[1] In the models, only the data from the tutees from the different tutoring treatments was compared. The tutor data was not included in the models.

Table 2. Details of models with treatment as fixed factor for the song structure and performance parameters.

Response variable	***Model term***	***Level***	***Estimate***	***SE***	***z or t***
A. Total number of syllables[1]	Intercept		1.52	0.13	11.29
	Treatment				
		Audio+female	-0.06	0.21	0.79
		Robot	0.00	0.19	0.00
		Robot mismatch	0.01	0.19	0.06
B. Number of unique syllables[1]	Intercept		1.66	0.12	13.16
	Treatment				
		Audio+female	-0.14	0.20	-0.71
		Robot	-0.19	0.19	-1.02
		Robot mismatch	-0.14	0.19	-0.76
C. Linearity[2]	Intercept		0.41	0.03	13.58
	Treatment				
		Audio+female	0.05	0.04	1.20
		Robot	0.02	0.04	0.61
		Robot mismatch	0.03	0.04	0.82

D. Consistency[2]	Intercept		0.83	0.03	24.6
	Treatment				
		Audio+female	0.06	0.05	1.24
		Robot	0.06	0.05	1.26
		Robot mismatch	0.06	0.05	1.14

[1] GLMM with a Poisson distribution and 'Tutor' as a random factor

[2] LMM with 'Tutor' as a random factor

Similarity to tutor song

Comparison different similarity assessment methods

There was a significant correlation between the human observer and the Luscinia similarity score, but not between the SAP and human observer or the SAP and Luscinia similarity scores (see Table 3), suggesting that these measures pick up on different dimensions of song similarity. It is important to note, however, that the human observer similarity scores were based on one exemplar of the typical motif, whereas the SAP and Luscinia scores were based on 10 randomly selected motifs per tutee.

Table 3. Pearson correlation coefficients for the human observer similarity scores (square-root transformed to meet assumptions of normality), the median SAP similarity scores and the median Luscinia similarity scores for the tutor to tutee comparison. Significant p-values are given in bold.

Comparison	N	r	p
Human observer sim. score – SAP sim. score	44	-0.14	0.37
Human observer sim. score – Luscinia sim. score	44	0.69	**< 0.01**
SAP sim. score - Luscinia sim. score	44	-0.14	0.37

Similarity scores for the comparison between tutor and tutee songs

To find out whether the tutees had learned from the tutor, we first checked whether their song was more similar to the song of their tutor than to the song of an unfamiliar male. The human observer similarity scores for the tutor to tutee and tutee to tutor comparison were significantly higher than the similarity scores for the comparisons with an unfamiliar song (the LMM with 'song model (tutor or unfamiliar)' was significantly better than the null LMM, 'song model' to tutee comparison: $N = 44$, $\chi^2 = 17.57$, $p < 0.01$, Table 4A, tutee to 'song model' comparison: $N = 44$, $\chi^2 = 16.12$, $p < 0.01$, Table 4B). As this meant that tutees' songs were more similar to their tutor's song than would be expected by random sharing in the colony, we assumed that the tutees had learned at least some aspects from their tutors. For all subsequent analyses, we proceeded

with comparisons between tutor and tutees only.

Table 4. Comparisons of the similarity of the model songs to the tutee songs (A) and the tutee songs to the model songs (B) by fitting lineair mixed models. Details of best model (LMM) for the arcsine square-root transformed human observer similarity scores are given.

Human observer similarity scores

Response variable	*Model term*	*Level*	*Estimate*	*SE*	*t*
A. Model-tutee[1]	Intercept		0.56	0.05	12.03
	Model				
		Unfamiliar	-0.16	0.04	-4.40
B. Tutee-model[1]	Intercept		0.65	0.05	11.86
	Model				
		Unfamiliar	-0.17	0.04	-4.20

[1]LMM with random factors 'Tutor' and 'Bird ID'.

In the comparison of the syllables in the tutor's repertoire to those in the tutee's repertoire (tutor-tutee comparison), adding 'treatment' as fixed factor did not significantly improve the null model (N = 44, $\chi^2 = 2.17$, p = 0.54). The human observer similarity scores were highest for the Robot mismatch (model estimates LMM X ± SE: 0.63 ± 0.08, Figure 4A, Table 5A) and the Robot group (X ± SE: 0.57 ± 0.07), and lowest for the audio (X ± SE: 0.53 ± 0.09) and audio+female group (X ± SE: 0.53 ± 0.08). In the comparison of the syllables in the tutee's repertoire to those in the tutor's repertoire (tutee-tutor comparison), adding 'treatment' as fixed factor also did not significantly improve the null model (N = 44, $\chi^2 = 3.91$, p = 0.27). For this comparison, human observer similarity scores were highest in the Robot group (model estimates LMM X ± SE: 0.74 ± 0.08, Figure 4B, Table 5A), followed by the Robot mismatch group (X ± SE: 0.67 ± 0.08) and were lowest in the audio+female (X ± SE: 0.61 ± 0.09) and the audio group (X ± SE: 0.59 ± 0.11).

Sound Analysis Pro similarity scores did not differ between treatment groups for the tutor-tutee or the tutee-tutor comparison (model including treatment was not significantly better than the model without treatment, tutor-tutee: N = 44, $\chi 2 = 6.20$, p = 0.10, Figure 4C, Table 5B, tutee-tutor: N = 44, $\chi^2 = 0.57$, p = 0.90, Figure 4D, Table 5B).

The Luscinia similarity score for the comparison of tutor and tutee song did not differ between treatment groups (model including treatment was not significantly better than the model without treatment, N = 44, $\chi^2 = 4.77$ p = 0.19,

Figure 4E, Table 5C).

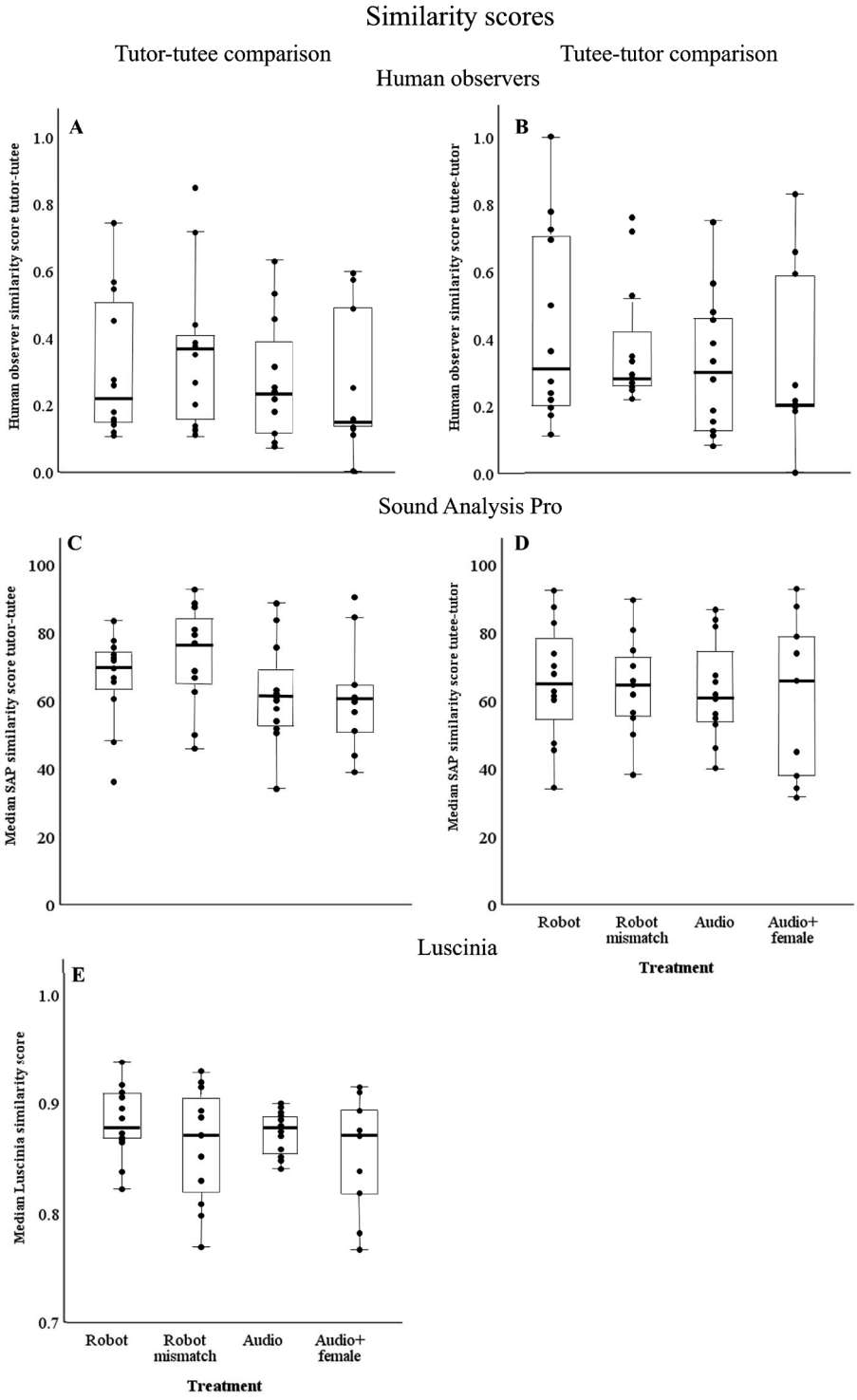

Figure 4. Graph showing the human observer similarity score for the tutor-tutee (a) and the tutee-tutor comparison (b), the SAP similarity score for the tutor-tutee (c) and the tutee-tutor (d) comparison and the Luscinia similarity score for the symmetric tutee and tutor comparison (e).

Table 5. Details of LMMs with 'Treatment' as fixed factor for the arcsine square root transformed human observer (A), SAP (B) and Luscinia (C) similarity scores for the comparison of tutor and tutee song.

			Tutor-tutee			***Tutee-tutor***		
Response variable	***Model term***	***Level***	***Estim.***	***SE***	***t***	***Estim.***	***SE***	***t***
A Human observers sim. scores[1]	Intercept		0.53	0.09	5.77	0.59	0.11	5.45
	Treatment							
		Audio +female	-0.001	0.08	-0.01	0.02	0.09	0.28
		Robot	0.04	0.07	0.60	0.15	0.08	1.83
		Robot mismatch	0.10	0.08	1.26	0.08	0.08	0.92
B SAP sim. scores[1]	Intercept		0.91	0.06	15.8	0.93	0.07	12.89
	Treatment							
		Audio +female	-0.01	0.06	-0.16	-0.01	0.06	-0.11
		Robot	0.05	0.06	0.91	0.03	0.05	0.58
		Robot mismatch	0.12	0.06	2.13	0.02	0.05	0.28
C Luscinia sim. scores[1]	Intercept		0.09	0.0008	119			
	Treatment							
		Audio +female	-0.001	0.0008	-1.25			
		Robot	0.0005	0.0007	0.67			
		Robot mismatch	-0.0007	0.0007	-0.99			

[1]LMM with random factor 'Tutor'.

SAP and Luscinia stereotypy scores

The treatment groups differed in the SAP stereotypy score: tutees from the audio+female group had a higher SAP stereotypy score than tutees from the audio group (model including 'treatment' was significantly better than null model

for the SAP stereotypy score (N = 41, χ^2 = 7.76, p = 0.05, Figure 5A, Table 6A)). There was no significant difference between the treatment groups in the Luscinia stereotypy scores (model including 'treatment' was not significantly better than null model for the Luscinia similarity score (N = 41, χ^2 = 1.62, p = 0.66, Figure 5B, Table 6B)), but for these scores, like for the SAP stereotypy scores, the estimate was lowest for the tutees from the audio group (Table 6B).

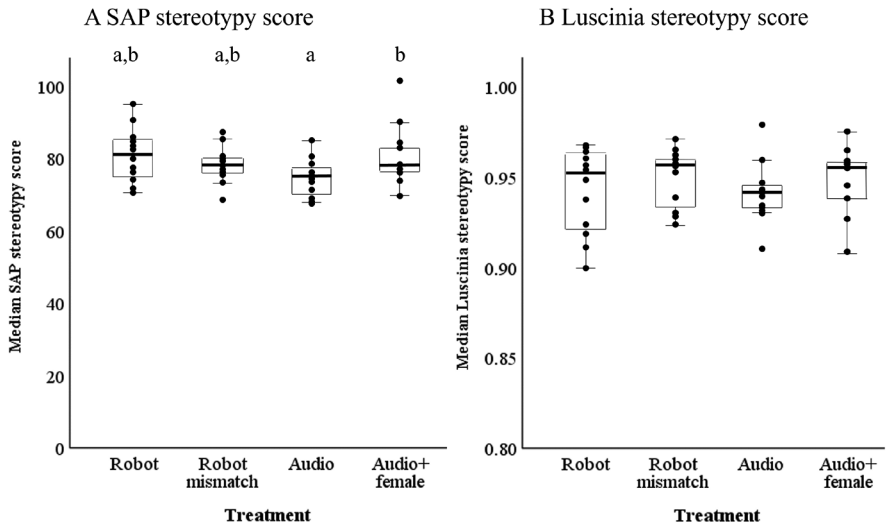

Figure 5. SAP (A) and Luscinia (B) stereotypy scores for the 10 randomly selected tutee motifs. Different letters above boxes in 5A indicate a significant difference of $p < 0.05$ according to post-hoc test, LMM see Table 6A.

Table 6. Details of best model (LMM) for the (arcsine square root transformed) SAP (A) and Luscinia (B) stereotypy scores for the comparison of ten randomly selected tutee motifs.

Response variable	***Model term***	***Level***	***Estimate***	***SE***	***z t***
A. SAP stereotypy score[1]	Intercept		1.00	0.03	28.79
	Treatment				
		Audio+female	0.13	0.05	2.63
		Robot	0.10	0.05	2.08
		Robot mismatch	0.05	0.05	1.04

B. Luscinia stereotypy score[2]	Intercept		0.94	0.006	168.46
	Treatment				
		Audio+female	0.007	0.008	0.82
		Robot	0.002	0.008	0.28
		Robot mismatch	0.009	0.008	1.09

[1] LMM with random factor 'Tutor'. Significant post-hoc comparisons: Audio vs. Audio+female: estimate: -0.13, SE: 0.05, t: -2.62, p = 0.05.
[2] LMM with random factor 'Tutor'.

Discussion and conclusions

The aim of this study was to test whether song learning from playback combined with a robotic zebra finch would lead to improved song learning if compared to audio-only playback. Tutees were raised under four different tutoring conditions: auditory tutor song playback, song playback together with synchronized visual cues produced by a RoboFinch, song playback and visual cues by a RoboFinch that started after song presentation had finished and auditory tutor song playback while the male was housed together with a female companion. Song learning success was assessed when the tutees had reached adulthood using three commonly used song similarity assessment methods. We had hypothesized that auditory stimulation combined with synchronized visual stimulation would improve song learning compared to unimodal auditory stimulation. However, contrary to our expectations, none of the similarity assessment methods detected a significant effect of tutoring treatment on the similarity between tutor and tutee song. There was, however, an effect of tutoring treatment on motif stereotypy as calculated in Sound Analysis Pro: this was lowest in the audio only group and highest in the audio+female group. The tutees raised with the robots had intermediate between-motif stereotypy and, other than the audio-only group, did not differ significantly from the audio+female group. This observation would be in line with an effect of multimodal exposure or an effect of a 'companion' arising both from a female or robot companion.

While this suggests an effect on improved motor performance via practicing, improved copying from the tutor (measured by the song similarity measures) was not found, although improved song learning in the multimodal condition had been our main prediction. The finding is however in line with previous studies presenting a visual stimulus before, during or after the playback of tutor song that also did not find an effect of the visual stimulation on zebra finch song copying success (Bolhuis, van Mil, & Houx, 1999; Houx & ten Cate, 1999). Because these studies had used a non-moving taxidermic mount of an adult zebra finch as a stationary visual stimulus, we had expected that visual

stimulation moving in synchrony with the song would be more salient and possibly have a stronger effect on song learning. Like in these previous studies and the video tutoring experiment described in Chapter 4, the tutees in the study described here were interested in the visual stimulus, as they spent a larger proportion of time close to the stimulus in the robot conditions than in conditions without a robot (Simon et al., in prep.). The tutees remained interested in the robot throughout the tutoring period. This suggests that multimodal stimulus presentation affects tutees' engagement with the stimulus, but, at least in the paradigms used for now, might not affect song learning success. It should be noted, however, that song learning and development entail more than just imitating the syllables in a tutor's song. A previous study, for instance, found no effect of rearing condition on the number of elements that tutees had copied from the tutor, but did find that adult female conspecifics discriminated and expressed different preferences for songs from tutees from the different rearing conditions (Holveck et al., 2008). This opens the possibility that the different tutoring treatments in the current study also might have affected aspects of song performance and delivery that we did not analyse here as we focussed on how much tutees learned from their tutors.

We found a difference between the solitary housed tutees raised with audio only tutor song exposure versus those raised with a female companion and audio-only song exposure. The latter group sang with a higher between-motif stereotypy than the birds that were also raised with audio only song exposure, but housed in social isolation throughout the tutoring period. This might be because the tutees housed with a female companion practiced more during motor learning than the tutees without a female companion, as zebra finches sing more while they have a social, male or female, companion, compared to socially isolated housing (Jesse & Riebel, 2012). The importance of practice on song quality has been demonstrated experimentally by temporarily pharmacologically blocking vocal motor control which disrupted vocal motor practice and resulted in impoverished adult song production (Pytte & Suthers, 2000). In young zebra finches that produce immature songs, a female conspecific can elicit songs with more mature properties, such as a higher stereotypy in the acoustic properties of syllables (Kojima & Doupe, 2011). This might mean that the tutees housed with the female companion practiced this more stereotyped version of song more often than the birds housed in social isolation, which possibly had an effect on the stereotypy in the adult song of these tutees. The lack of a live companion is a potential confound in previous studies comparing live with tape tutoring: live tutored tutees usually have the tutor as a social companion, while the tutees with audio only exposure to tutor song are normally

housed in social isolation (e.g. Chen, Matheson, & Sakata, 2016; Derégnaucourt et al., 2013; Eales, 1989). Our results suggest that being housed with or without a social companion during song development affects song learning outcomes and that future studies should aim for a comparable social environment across different tutoring conditions. The tutees in the robot conditions sang with intermediate levels of between-motif stereotypy that did not differ significantly from the other two conditions. This suggests that being housed with a RoboFinch might affect motif stereotypy to some degree. Observations of tutee behaviour during this tutoring experiment showed that tutee singing behaviour was affected by the presence of the robots (Simon et al., in prep.), which in turn might have influenced the stereotypy with which the tutees produced their song. This suggests that the robot could be a tool to identify what stimulus properties are essential for 'social interaction' (Nelson, 1997).

There are several explanations possible for the absence of an effect of the RoboFinches on tutor song similarity. One possibility is that the context in which the tutor songs were recorded was suboptimal. We recorded tutors that were housed alone and singing undirected song. However, when housed together with juveniles, zebra finch adults can produce pupil-directed song towards them (Chen et al., 2016). This differs from undirected and female-directed song in several acoustic parameters. Female-directed and undirected song differ in the accompanying body posture and movements (Sossinka & Böhner, 1980) and it is possible that specific visual components proceed, accompany or follow the production of pupil-directed song. It might therefore be that tutoring with audio or audio-visual pupil-directed song leads to better song learning outcomes compared to tutoring with undirected song. Another future avenue to explore is the role of interaction and tutee-tutor contingencies. The RoboFinch could be used to emulate the interactive properties of a live tutor. For example, the RoboFinch could present tutor song contingent with tutee behaviour, or could respond to immature tutee vocalizations. Both of these interactive processes are thought to facilitate zebra finch song learning (Adret 1993; Derégnaucourt et al. 2013, Carouso-Peck and Goldstein 2019; Carouso-Peck et al. 2020, but see Houx and ten Cate 1999b). A final possibility is that the amount of song exposure frequency was suboptimal, possibly leading to a ceiling or floor effect and thereby masking treatment effects. Song exposure frequency is a debated influence on song learning (Chen et al. 2016; Derégnaucourt et al. 2013; Tchernichovski et al. 1999). In the present study, tutees were exposed to approximately 276 motifs daily, which was based on song rates expected for live tutors (Böhner, 1983; Jesse & Riebel, 2012). Some studies, however, suggest that a high amount of song exposure might negatively affect

zebra finch song learning (Chen et al., 2016; Tchernichovski & Mitra, 2002; Tchernichovski et al., 1999), and that exposure to 40 motifs daily leads to optimal song copying (Tchernichovski et al. 1999). More research is needed to find out the optimal song exposure frequency for song tutoring using robots.

It is also possible that our sample size was too small and there was too much individual variation to be able to detect treatment effects or that the differences in song learning between the treatment groups were too subtle to be picked up by our song analysis methods. However, in order to compare our data with the classic zebra finch song learning literature as well as with the more recent song learning studies, we used the three most common and established similarity assessment methods: human observers, SAP and Luscinia. Even though the correlation between the scores obtained by the different methods was low, suggesting that the methods pick up different aspects of song similarity, none of these methods picked up a significant effect of treatment on tutor-tutee similarity. Unlike other studies that have demonstrated improved learning with multimodal stimulation (Hebets & Papaj, 2005, Rowe, 1999, Hultsch et al. 1999), the results of this study did not show a facilitating effect of multimodal exposure on zebra finch song learning. This was, however, the first study using a robotic zebra finch to study the effect of multimodal cues on song learning. More research is needed to find out how different methodological choices affect the influence of the RoboFinch on zebra finch behaviour and song learning. As the RoboFinch enables researchers to standardize and control both the auditory and visual information presented to young birds, it is an interesting tool for future research into multimodal communication.

Acknowledgements
Funding for this research was provided by a Human Frontier Science Program Grant (No RGP0046/2016). We would like to thank Jing Wei, Quanxiao Liu and Zhiyuan Ning for the visual comparison of the spectrograms and Dré Kampfraath, Rogier Elsinga, Peter Wiersma and Wesley Delmeer for their help during the development of the RoboFinch.

References

Adret, P. (1993). Operant conditioning, song learning and imprinting to taped song in the zebra finch. Animal Behaviour, 46, 149–159.

Anderson, R. C., DuBois, A. L., Piech, D. K., Searcy, W. A., & Nowicki, S. (2013). Male response to an aggressive visual signal, the wing wave display, in swamp sparrows. Behavioral Ecology and Sociobiology, 67(4), 593–600. https://doi.org/10.1007/s00265-013-1478-9

Baptista, L. F., & Gaunt, S. L. L. (1997). Social interaction and vocal development in birds. In C. T. Snowdon & M. Hausberger (Eds.), Social influences on vocal development (pp. 23–40). Cambridge, Cambridge University Press.

Baptista, L. F., & Petrinovich, L. (1986). Song development in the white-crowned sparrow: social factors and sex differences. Animal Behaviour, 34(5), 1359–1371. https://doi.org/10.1016/S0003-3472(86)80207-X

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. 67(1). https://doi.org/10.18637/jss.v067.i01

Beecher, M. D., & Burt, J. M. (2004). The role of social interaction in bird song learning. Current Directions in Psychological Science, 13(6), 224–228. https://doi.org/10.1111/j.0963-7214.2004.00313.x

Böhner, J. (1983). Song learning in the zebra finch (*Taeniopygia guttata*): Selectivity in the choice of a tutor and accuracy of song copies. Animal Behaviour, 31(1), 231–237. https://doi.org/10.1016/S0003-3472(83)80193-6

Bolhuis, J. J., Okanoya, K., & Scharff, C. (2010). Twitter evolution: Converging mechanisms in birdsong and human speech. Nature Reviews Neuroscience, 11(11), 747–759. https://doi.org/10.1038/nrn2931

Bolhuis, J., van Mil, D., & Houx, B. (1999). Song learning with audiovisual compound stimuli in zebra finches. Animal Behaviour, 58, 1285–1292. https://doi.org/10.1006/anbe.1999.1266

Brainard, M. S., & Doupe, A. J. (2002). What songbirds teach us about learning. Nature, 417, 351–358. https://doi.org/10.1038/417351a

Butler, N. E., Magrath, R. D., & Peters, R. A. (2017). Lack of alarm calls in a gregarious bird: models and videos of predators prompt alarm responses but no alarm calls by zebra finches. Behavioral Ecology and Sociobiology, 71(8). https://doi.org/10.1007/s00265-017-2343-z

Catchpole, C. K., & Slater, P. J. B. (1995). How song develops. In C. K. Catchpole & P. J. B. Slater (Eds.), Bird Song: Biological Themes and Variations (pp. 45–69). Cambridge: Cambridge University Press.

Catchpole, C. K., & Slater, P. J. B. (2003). Bird song: biological themes and variations. Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.004

Chen, Y., Matheson, L. E., & Sakata, J. T. (2016). Mechanisms underlying the social enhancement of vocal learning in songbirds. Proceedings of the National Academy of Sciences, 201522306. https://doi.org/10.1073/pnas.1522306113

Cuthill, I. C., Hart, N. S., Partridge, J. C., Bennett, A. T. D., Hunt, S., & Church, S. C. (2000). Avian colour vision and avian video playback experiments. Acta Ethologica, 3, 29–37. https://doi.org/10.1007/s102110000027

Derégnaucourt, S. (2011). Birdsong learning in the laboratory, with especial reference to the song of the zebra finch (*Taeniopygia guttata*). Interaction Studies, 12, 324–350. https://doi.org/10.1075/is.12.2.07der

Derégnaucourt, S., Poirier, C., van der Kant, A., & van der Linden, A. (2013). Comparisons of different methods to train a young zebra finch (*Taeniopygia guttata*) to learn a song. Journal of Physiology, 107, 210–218. https://doi.

org/10.1016/j.jphysparis.2012.08.003
Doupe, A. J., & Kuhl, P. K. (1999). Bird song and human speech: common themes and mechanisms. Annu. Rev. Neurosci., 22, 567–631. https://doi.org/10.1146/annurev.neuro.22.1.567
Eales, L. A. (1989). The influences of visual and vocal interaction on song learning in zebra finches. Animal Behaviour, 37, 507–508. https://doi.org/10.1016/0003-3472(89)90097-3
Feenders, G., Kato, Y., Borzeszkowski, K. M., & Klump, G. M. (2017). Temporal ventriloquism effect in european starlings: evidence for two parallel processing pathways. Behavioral Neuroscience, 131(4), 337–347. https://doi.org/10.1037/bne0000200
Fleishman, L. J., & Endler, J. A. (2000). Some comments on visual perception and the use of video playback in animal behavior studies. Acta Ethologica, 3(1), 15–27. https://doi.org/10.1007/s102110000025
Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. Proceedings of the National Academy of Sciences of the United States of America, 100(13), 8030–8035. https://doi.org/10.1073/pnas.1332441100
Goller, F., Mallinckrodt, M. J., & Torti, S. D. (2004). Beak gape dynamics, during song in the zebra finch. Journal of Neurobiology, 59(3), 289–303. https://doi.org/10.1002/neu.10327
Griffith, S. C., & Buchanan, K. L. (2010). The zebra finch : the ultimate Australian supermodel. Emu, 110, v–xii. https://doi.org/10.1071/MUv110n3_ED
Holveck, M. J., Vieira De Castro, A. C., Lachlan, R. F., ten Cate, C., & Riebel, K. (2008). Accuracy of song syntax learning and singing consistency signal early condition in zebra finches. Behavioral Ecology, 19(6), 1267–1281. https://doi.org/10.1093/beheco/arn078
Houx, B. B., & ten Cate, C. (1999a). Do stimulus-stimulus contingencies affect song learning in zebra finches (*Taeniopygia guttata*)? Journal of Comparative Psychology, 113(3), 235–242. https://doi.org/10.1037/0735-7036.113.3.235
Houx, B. B., & ten Cate, C. (1999b). Song learning from playback in zebra finches: is there an effect of operant contingency? Animal Behaviour, 57(4), 837–845. https://doi.org/10.1006/anbe.1998.1046
Hultsch, H., Schleuss, F., & Todt, D. (1999). Auditory-visual stimulus pairing enhances perceptual learning in a songbird. Animal Behaviour, 58, 143–149. https://doi.org/10.1006/anbe.1999.1120
Jesse, F., & Riebel, K. (2012). Social facilitation of male song by male and female conspecifics in the zebra finch, *Taeniopygia guttata*. Behavioural Processes, 91(3), 262–266. https://doi.org/10.1016/j.beproc.2012.09.006
Kojima, S., & Doupe, A. J. (2011). Social performance reveals unexpected vocal competency in young songbirds. Proceedings of the National Academy of Sciences of the United States of America, 108(4), 1687–1692. https://doi.org/10.1073/pnas.1010502108
Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy.

Science, 218, 1138–1141. https://doi.org/10.1126/science.7146899
Lachlan, R. F., van Heijningen, C. A. A., ter Haar, S. M., & ten Cate, C. (2016). Zebra finch song phonology and syntactical structure across populations and continents-a computational comparison. Frontiers in Psychology, 7, 1–19. https://doi.org/10.3389/fpsyg.2016.00980
Lachlan, R. F., Verhagen, L., Peters, S., & ten Cate, C. (2010). Are there species-universal categories in bird song phonology and syntax? A comparative study of chaffinches (*Fringilla coelebs*), zebra finches (*Taenopygia guttata*), and swamp sparrows (*Melospiza georgiana*). Journal of Comparative Psychology, 124(1), 92–108. https://doi.org/10.1037/a0016996
Landgraf, T., Moballegh, H., & Rojas, R. (2008). Design and development of a robotic bee for the analysis of honeybee dance communication. Applied Bionics and Biomechanics, 5(3), 157–164. https://doi.org/10.1080/11762320802617552
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: estimated marginal means, aka least-squares means.
Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. Proceedings of the National Academy of Sciences of the United States of America, 109(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109
Mello, C. V. (2014). The zebra finch, *Taeniopygia guttata*: An avian model for investigating the neurobiological basis of vocal learning. Cold Spring Harbor Protocols, 2014(12), 1237–1242. https://doi.org/10.1101/pdb.emo084574
Nelson, D. (1997). Social interaction and sensitive phases for song learning: A critical review. In C. T. Snowdon & M. Hausberger (Eds.), Social influences on vocal development (pp. 7–22). Cambridge, Cambridge University Press.
Oliveira, R. F., Rosenthal, G. G., Schlupp, I., McGregor, P. K., Cuthill, I. C., Endler, J. A., Fleishman, L. J., Zeil, J., Barata, E., Burford, F., Gonzalves, D., Haley, M., Jakobsson, S., Jennions, M. D., Koner, K. E., Lindstrom, L., Peake, T., Pilastro, A., Pope, D. S., … Waas, J. R. (2000). Considerations on the use of video playbacks as visual stimuli: the Lisbon workshop consensus. Acta Ethologica, 3(1), 61–65. https://doi.org/10.1007/s102110000019
Patricelli, G. L., Uy, J. A. C., Walsh, G., & Borgia, G. (2002). Male displays adjusted to female's response. Nature, 415(6869), 279–280. https://doi.org/10.1038/415279a
Pytte, C. L., & Suthers, R. A. (2000). Sensitive period for sensorimotor integration during vocal motor learning. Journal of Neurobiology, 42(2), 172–189. https://doi.org/10.1002/(SICI)1097-4695(20000205)42:2<172::AID-NEU2>3.0.CO;2-I
Ręk, P., & Magrath, R. D. (2016). Multimodal duetting in magpie-larks: how do vocal and visual components contribute to a cooperative signal's function? Animal Behaviour, 117, 35–42. https://doi.org/10.1016/j.anbehav.2016.04.024
Rowe, C. (1999). Receiver psychology and evolution of multicomponent signals. Animal Behaviour, 58, 921–931. https://doi.org/10.1006/anbe.1999.1242
Simon, R., Varkevisser, J., Mendoza, E., Hochradel, K., Scharff, C., Riebel, K., & Half-

werk, W. (2019). Development and application of a robotic zebra finch (RoboFinch) to study multimodal cues in vocal communication. PeerJ Preprints 7:E28004v3. https://doi.org/10.7287/peerj.preprints.28004v1

Slater, P. J. B., Eales, L. A., & Clayton, N. S. (1988). Song learning in zebra finches (*Taeniopygia guttata*): progress and prospects. Advances in the Study of Behaviour, 18, 1–34. https://doi.org/10.1016/S0065-3454(08)60308-3

Soma, M. F. (2011). Social factors in song learning: a review of Estrildid finch research. Ornithological Science, 10(2), 89–100. https://doi.org/10.2326/osj.10.89

Sossinka, R., & Böhner, J. (1980). Song types in the zebra finch. Zeitschrift Für Tierpsychologie, 53, 123–132. https://doi.org/10.1111/j.1439-0310.1980.tb01044.x

Taylor, R. C., Klein, B. A., Stein, J., & Ryan, M. J. (2008). Faux frogs: multimodal signalling and the value of robotics in animal behaviour. Animal Behaviour, 76(3), 1089–1097. https://doi.org/10.1016/j.anbehav.2008.01.031

Tchernichovski, O, & Mitra, P. P. (2002). Towards quantification of vocal imitation in the zebra finch. Journal of Comparative Physiology A, 188(11–12), 867–878. https://doi.org/10.1007/s00359-002-0352-4

Tchernichovski, O., Lints, T., Mitra, P. P., & Nottebohm, F. (1999). Vocal imitation in zebra finches is inversely related to model abundance. Proceedings of the National Academy of Sciences of the United States of America, 96(22), 12901–12904. https://doi.org/10.1073/pnas.96.22.12901

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., & Mitra, P. P. (2000). A procedure for an automated measurement of song similarity. Animal Behaviour, 59(6), 1167–1176. https://doi.org/10.1006/anbe.1999.1416

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. Cognition, 108(3), 850–855. https://doi.org/10.1016/j.cognition.2008.05.009

Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. Journal of Child Language, 42(06), 1173–1190. https://doi.org/10.1017/S0305000914000725

Ullrich, R., Norton, P., & Scharff, C. (2016). Waltzing Taeniopygia: integration of courtship song and dance in the domesticated Australian zebra finch. Animal Behaviour, 112, 285–300. https://doi.org/10.1016/j.anbehav.2015.11.012

van Kampen, H. S., & Bolhuis, J. J. (1991). Auditory learning and filial imprinting in the chick. Behaviour, 117, 303–319. https://doi.org/10.1163/156853991X00607

van Kampen, H. S., & Bolhuis, J. J. (1993). Interaction between auditory and visual learning during filial imprinting. Animal Behaviour, 45, 623–625. https://doi.org/10.1006/anbe.1993.1074

Varkevisser, J. M., Simon, R., Mendoza, E., How, M., van Hijlkema, I., Jin, R., Liang, Q., Scharff, C., Halfwerk, W. H., & Riebel, K. (2021). Adding colour-realistic video images to audio playbacks increases stimulus engagement but does not enhance vocal learning in zebra finches. Animal Cognition. https://doi.org/10.1007/s10071-021-01547-8

Williams, H. (2001). Choreography of song, dance and beak movements in the zebra finch (*Taeniopygia guttata*). The Journal of Experimental Biology, 204, 3497–3506.

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: Predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. Developmental Science, 12(5), 798–814. https://doi.org/10.1111/j.1467-7687.2009.00833.x

Appendix

Table A1. File presentation schedule used during tutoring sessions. 'Time' indicates the time at which the playback started.

Time	File	# playbacks	Time	File	# playbacks
8:15	song1	4	10:40	head movement1	2
8:17	head movement1	2	10:42	song2	4
8:19	call1	2	10:44	call2	2
8:21	head movement2	4	12:15	head movement2	3
8:23	song2	2	12:16	head movement1	2
8:25	call2	4	12:19	call2	1
8:27	song1	2	12:20	song3	2
8:30	head movement1	3	12:23	head movement1	4
8:31	head movement2	2	12:26	song3	3
8:32	song3	2	12:28	call2	3
8:34	call2	3	12:32	call1	4
8:36	head movement2	2	12:35	head movement2	3
8:38	song3	4	12:38	song2	2
8:41	head movement1	3	12:40	head movement1	4
8:43	call1	1	12:43	call1	2
8:44	song2	2	14:15	song1	3
9:15	call1	4	14:16	song2	2
9:17	song1	3	14:20	call2	3
9:20	head movement1	3	14:23	head movement2	4
9:23	call1	2	14:26	call1	1
9:26	song1	3	14:27	head movement1	2
9:29	head movement2	4	14:30	song3	2
9:32	song3	2	14:34	call2	4
9:34	call2	3	14:36	head movement1	1
9:36	song2	2	14:37	call1	3
9:38	call2	3	14:40	call2	2
9:40	song2	3	14:43	head movement2	3
9:42	head movement1	2	16:15	head movement2	3
9:43	song1	3	16:17	head movement1	2
10:15	song3	4	16:19	call1	5
10:16	song2	1	16:22	song2	3

10:20	call2	2	16:25	call2	5
10:22	head movement2	4	16:30	song1	2
10:25	call2	5	16:33	head movement1	2
10:27	song3	3	16:36	song3	2
10:30	head movement1	2	16:38	call1	1
10:32	call1	1	16:40	call2	4
10:33	head movement1	2	16:43	head movement1	2
10:35	song1	4	16:44	head movement2	1
10:37	call1	5			

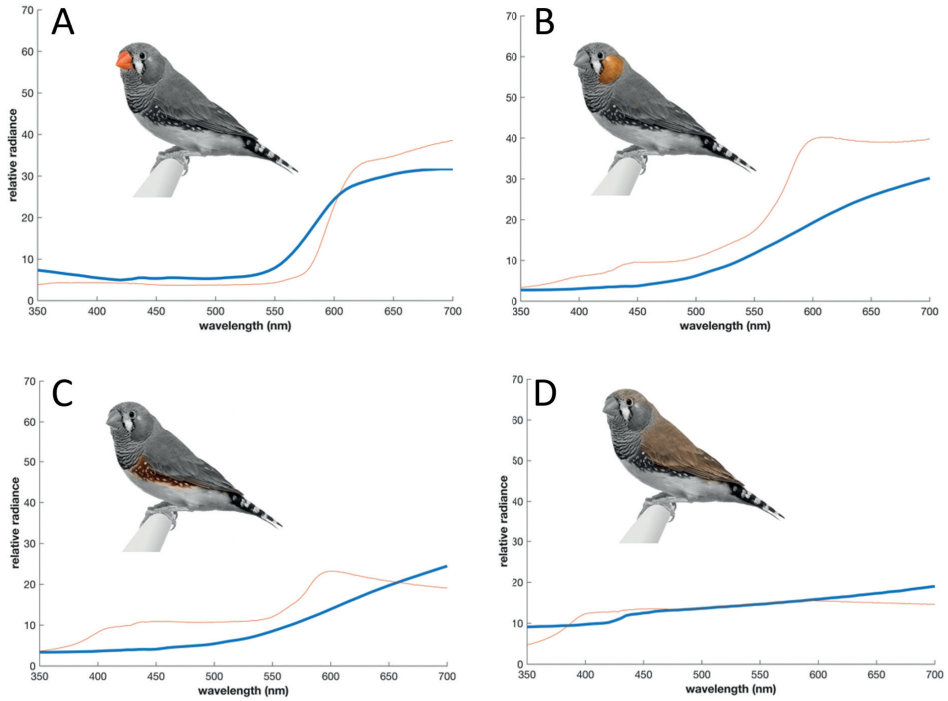

Figure A1. Colouring of the RoboFinch (red line) in comparison to real feathers/beaks of zebra finches (blue line). (a) beak, (b) cheeks, (c) sides, (d) back.