Seeing voices: the role of multimodal cues in vocal learning

Varkevisser, J.M.

Seeing voices: the role of multimodal cues in vocal learning

Judith Varkevisser

Varkevisser, Judith Mirjam

Seeing voices: the role of multimodal cues in vocal learning

PhD thesis, Leiden University, the Netherlands

An electronic version of this thesis can be downloaded from:
openaccess.leidenuniv.nl

Cover design by Daniël de Muynck
Printed by Ridderprint | www.ridderprint.nl

Seeing voices:
the role of multimodal cues in vocal learning

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 20 oktober 2022
klokke 16:15 uur

door

Judith Mirjam Varkevisser
geboren te Leidschendam, Nederland
in 1992

Promotor

Prof. dr. C. J. ten Cate

Copromotor

Dr. K. Riebel

Promotiecommissie

Prof. dr. G. P. van Wezel

Prof. dr. M. Richardson

Prof. dr. C. Levelt

Prof. dr. M. Naguib (Wageningen University & Research)

Dr. J. Sakata (McGill University, Canada)

Dr. J. Hyland Bruno (Columbia University, USA)

Dit onderzoek is gefinancierd door de Human Frontier Science Program

Don't you wonder sometimes 'bout sound and vision?

David Bowie

Table of Contents

Chapter 1

General introduction

Both humans and songbirds are vocal learners that learn to produce their species-specific vocalizations early in life by exposure to the vocalizations of adult conspecifics (Doupe & Kuhl, 1999). For human speech and birdsong, better learning outcomes are often achieved with live, social, tutors than with audio-only exposure to vocalizations (speech: Bruner, 1983; Kuhl, Tsao, & Liu, 2003; Roseberry, Hirsh-Pasek, & Golinkoff, 2014, birdsong: reviewed in Baptista & Gaunt, 1997; Soma, 2011). Many researchers have argued that this is because social interactions between tutors and tutees are important in the vocal learning process (e.g. Beecher & Burt, 2004; Goldstein, King, & West, 2003; Kuhl, 2003, 2007). An open question, however, is whether and to what extent vocal learning from live tutors is also improved because live tutors enable tutees to both hear and see a tutor, instead of only hear a tutor (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). Live tutoring, in other words, results in multimodal exposure to a tutor i.e. stimulation of multiple sensory modalities, while audio-only tutoring results in unimodal tutor exposure with stimulation of a single modality.

The simultaneous presentation of two stimuli in different modalities can improve signal perception compared to the presentation of one stimulus, as has been demonstrated in laboratory experiments in many taxonomic groups (reviewed in Rowe, 1999). Improved signal processing can occur when both stimuli are informative, but also when only one stimulus is relevant to the receiver, while the other is task-irrelevant and uninformative, but can draw the receiver's attention to the relevant stimulus (Alais, Newell, & Mamassian, 2010; Feenders, Kato, Borzeszkowski, & Klump, 2017; Rowe, 1999). The production of birdsong and speech are accompanied by sound-specific visual cues, such as songbirds' beak movements and human mouth movements. This makes speech and birdsong multimodal signals, i.e. signals that can be perceived through more than one sensory modality (Halfwerk et al., 2019; Higham & Hebets, 2013; Partan & Marler, 1999). Multi- compared to unimodal signalling can be beneficial for communication. For instance, if there is noise in one channel, information conveyed in the other channel can help receivers to identify the signal correctly (Partan & Marler, 2005). Additionally, multimodal signals are more likely to be detected by receivers than unimodal signals and receivers learn to recognize signals which contain multiple components (in one or multiple modalities) faster than single component signals (reviewed in Hebets & Papaj, 2005; Rowe, 1999). For example, big brown bats (*Eptesicus fuscus*) learn to avoid noxious fireflies faster with multi- than with unimodal warning signals (Leavell et al.,

2018). Exposure to the sound-specific visual cues accompanying vocalizations might have a faciliting effect on vocal learning, for instance by improving receivers' attention to the auditory signal or by directly facilitating learning of the underlying motor program to produce these vocalizations. The idea that audio-visual compared to audio-only exposure to vocalizing tutors could have a facilitating effect on vocal learning is supported by multiple lines of evidence from human and non-human animals.

In humans, visual information can affect speech perception in adults and infants. Visual exposure to a speaker's mouth and facial movements contributes to speech intelligibility, especially in noisy environments (Middelweerd & Plomp, 1987; Sumby & Pollack, 1954). Human infants of two months old already associate auditory and visually presented phonemes (Kuhl & Meltzoff, 1982). Besides, infants of around four months of age already perceive a different speech sound if auditory and visual speech cues are experimentally mismatched than they would perceive if the auditory and visual cue were presented separately (Burnham & Dodd, 2004). This is known as the McGurk effect and indicates that auditory and visual information are integrated into a multimodal percept (McGurk & Macdonald, 1976). Observational studies suggest that visual exposure to a speaking adult might play a role in early speech development. For instance, infants that fixate their gaze more on their mother's mouth during interaction at 6 months, show higher levels of expressive language (e.g. repeating sentences or naming objects) at age 2 (Young, Merin, Rogers, & Ozonoff, 2009). In addition, visual speech enhances learning of phoneme contrasts in 6-month-olds (Teinonen et al., 2008), and 12-month-olds pay more attention to a speaker's mouth when hearing a foreign language compared to their native language (Lewkowicz & Hansen-Tift, 2012). Infants that are born blind never experience visual exposure to speech. Although they acquire a speech system that seems comparable to that of sighted individuals, differences in the pronunciation of certain phonemes by blind and sighted individuals have been demonstrated (Ménard, Dupont, Baum, & Aubin, 2009). Moreover, for second language learning in adults, audio-visual training (with a speaker's mouth movements presented through videos of the speaker or through animation of a virtual head) improves the perception and production of unfamiliar speech contrasts more than audio-only training (e.g. Badin, Tarabalka, Elisei, & Bailly, 2010; Hazan, Sennema, Iba, & Faulkner, 2005; Hirata & Kelly, 2010; Liu, Massaro, Chen, Chan, & Perfetti, 2007; Wang, Hueber, & Badin, 2014). These studies suggest that it is worthwhile to experimentally investigate how early vocal development is affected by visual exposure to a tutor.

Like in humans, there are several studies in songbirds suggesting that visual stimulation might affect the perception and learning of vocalizations. Starlings (*Sturnus vulgaris*), for instance, show enhanced performance on a temporal order judgement task when visual cues were flanked by auditory stimuli (Feenders, Kato, Borzeszkowski, & Klump, 2017). This demonstrates that concurrent auditory and visual stimulation can influence stimulus perception. In nightingales (*Luscinia megarhynchos*), song presentation paired with stroboscope light flashes improved song learning compared to song presentation with no additional visual stimulation (Hultsch, Schleuss, & Todt, 1999). In the context of filial imprinting, young birds showed enhanced learning of an auditory stimulus when it was paired with a visual stimulus (van Kampen & Bolhuis, 1991; van Kampen & Bolhuis, 1993). These last two studies showed an effect of non-social and non-sound-specific visual stimulation on learning an auditory signal. It might be that any visual stimulation in addition to an auditory signal improves vocal learning of that signal equally, in which case visual exposure to a tutor would facilitate vocal learning to the same degree as non-social and non-sound-specific visual stimulation. It might also be that visual exposure to sound-specific movements has an additional facilitating effect on song learning, in which case seeing a singing tutor in addition to hearing song would facilitate song learning more than exposure to non-sound-specific visual stimulation.

Birdsong development provides a model system that can be used to experimentally investigate the effect of audio-visual compared to audio-only exposure to a tutor on vocal development. In this thesis, this question will be addressed investigating song development in zebra finches (*Taeniopygia guttata*), the primary experimental animal model for studies on vocal learning (Griffith & Buchanan, 2010; Mello, 2014). Zebra finch song production is accompanied by specific beak and body movements (Franz & Goller, 2002; Ohms, Snelderwaard, ten Cate, & Beckers, 2010; Ullrich, Norton, & Scharff, 2016; Williams, 2001). Individual zebra finches show stereotyped patterns of beak movements during song renditions (Goller, Mallinckrodt, & Torti, 2004; Williams, 2001). Changes in beak aperture are correlated with changes in song amplitude and frequency, and rapid changes in beak aperture occur mainly just before the onset of sound production and at rapid acoustic transitions during song (Goller et al., 2004; Ohms et al., 2010; Williams, 2001). A correlation between beak aperture and song frequency has been demonstrated in other songbird species as well (e.g. Podos, Southall, & Rossi-Santos, 2004; Westneat, Long, Hoese, & Nowicki, 1993). Zebra finches mainly combine singing with body movements as part of the courtship display performed in the presence of female conspe-

cifics (Ullrich et al., 2016; Williams, 2001). Audio-visual exposure to a singing tutor might affect song learning in zebra finches because it enables them to see these song-specific movements in addition to hearing the song.

The possibility that visual exposure to a singing tutor might (partially) explain improved song learning from live tutors has not been systematically studied yet. However, several observations suggest that young birds may attend to both auditory and visual information during song learning. In zebra finches, for instance, the beak movements of pupils show high similarity with those of their tutors compared to unfamiliar males (Williams, 2001), visual cues guide tutor choice (Mann & Slater, 1995; Mann, Slater, Eales, & Richards, 1991), and visual stimulation contingent with immature song production improves song learning in juvenile zebra finches (Carouso-Peck and Goldstein 2019). Moreover, zebra finches copy more song from a visible conspecific than from the playback of pre-recorded tutor song (Derégnaucourt, Poirier, van der Kant, & van der Linden, 2013) or from direct passive auditory exposure to a tutor through an opaque screen (Eales, 1989) or loudspeaker (Chen, Matheson, & Sakata, 2016). Although this suggests that seeing a tutor improves song learning, in these studies multimodal and social tutoring were confounded: in the tutoring treatments in which tutees could see their tutor, they could also visually interact with it. This makes it difficult to disentangle the effect of social and multimodal tutoring on zebra finch song learning. In this thesis, I therefore investigated the effect of multimodal tutoring on song learning success by using different tutoring methods where tutees could see the visual component of song production, without being able to visually interact with a tutor. To this end, I could utilise a robotic zebra finch that was jointly developed with collaborators from the Vrije Universiteit Amsterdam and that allowed standardized and controlled presentation of the auditory and visual component of song.

Thesis outline

To answer the question whether multi- compared to unimodal exposure to a tutor affects zebra finch song learning, this thesis first revisits the literature on zebra finch song learning experiments from the perspective of multi- versus unimodal tutoring and then describes three different song tutoring experiments.

In Chapter 2, the literature on zebra finch song learning under different tutoring treatments was reviewed to find out whether it supports the hypothesis that multi- compared to unimodal tutoring facilitates zebra finch song learning. Zebra finches copy more song from a live tutor than from auditory only expo-

sure to tutor song (Chen et al., 2016; Derégnaucourt et al., 2013; Eales, 1989). Several stimulus dimensions that differ between live and audio only tutoring have been experimentally tested for their effect on song learning, but it is as yet unclear what exactly the key facilitating factor of a live tutor is. The most favoured hypothesis for this difference is that a lack of social interaction with the tutor leads to poorer song copying from audio only playback than from a live tutor (Chen et al., 2016; Derégnaucourt et al., 2013; Slater, Eales, & Clayton, 1988). In this review, I investigated whether previous song learning studies have systematically controlled for multi- versus unimodal tutoring and whether their outcomes are in line with multi- compared to unimodal tutoring having an effect on the song learning process.

Chapter 3 describes a song tutoring experiment aimed at testing whether multi- compared to unimodal exposure to a live tutor facilitates zebra finch song learning. I investigated song learning in tutees that had visual exposure to an adult conspecific (the tutor) through a one-way mirror. These tutees thus had multimodal tutor exposure, but as the tutor could not see them, there was no visual social tutor-tutee interaction possible. I compared song learning in these tutees to that in tutees that did not have visual, but only auditory and therefore unimodal exposure to the tutor. I also investigated song learning in tutees that were raised in the same cage as the tutor, and that thus had multimodal tutor exposure, as well as the opportunity to visually and physically interact with the tutor. Tutees from all treatments could vocally interact with each other and the tutor and all tutees were housed with a female companion to avoid social isolation in the tutees that were not housed in the same cage as the tutor. If visual cues play a role in song learning, the tutees with multimodal tutor exposure would show improved tutor song copying compared to the tutees with unimodal tutor exposure. These results could be interpreted as support for the hypothesis that multimodal tutor exposure facilitates song learning, but an alternative, non-mutually exclusive, explanation could be that the tutor had facilitated song learning by providing visual feedback in response to the tutees' vocalizations. To prevent the possibility of the tutor providing visual feedback to tutees, I used artificial, instead of live, tutors in Chapter 4 and 5.

The studies described in Chapter 4 and 5 tested whether learning from passive, pre-recorded tutor song would be facilitated if tutees would at the same time be exposed to the visual cues accompanying the production of this song. In Chapter 4, these visual cues were presented through videos that had been adjusted for the zebra finch visual system using colour realistic imaginary and high speed video recordings and displays. I investigated song learning in tutees

that were exposed to a time aligned video of a tutor singing the song that they were at the same time auditorily exposed to. I compared this to learning in tutees that only heard this song and tutees that heard this song while they were exposed to the tutor video, but here the pixels were randomized and the frames were played in reversed order. The tutees that were presented to the original tutor video in addition to auditory song exposure were expected to show improved song learning compared to the other two tutoring conditions. While the realistic imaging techniques thus ensured a high fidelity audio-visual recording of a singing male, a video is only two-dimensional and lacking the depth of a real bird. This issue was addressed in the study described in Chapter 5.

In the study in Chapter 5, visual cues were presented by means of a three-dimensional robotic zebra finch producing beak and head movements time-aligned with the tutor song (RoboFinch, Simon et al., 2019). Tutees were exposed to the RoboFinch and their song learning was compared to that in two control groups: tutees exposed to the same tutor song without the robotic zebra finch present and tutees exposed to a robotic zebra finch that only started moving after auditory song presentation had finished. In this experiment, I also included a condition in which tutees were housed with a female companion while being exposed to song auditorily only, to find out whether the social isolation of the other tutees would negatively affect their song learning success. I expected the visual cues produced by the Robofinch and presented synchronized with the auditory song playback to facilitate song learning and to lead to a higher amount of tutor song copying than the other tutoring treatments.

Chapter 6 discusses the main conclusions with respect to the effect of audio-visual tutor exposure on song learning and discusses the results of this thesis in a broader perspective.

References

Alais, D., Newell, F. N., & Mamassian, P. (2010). Multisensory processing in review: From physiology to behaviour. Seeing and Perceiving (Vol. 23). https://doi.org/10.1163/187847510X488603

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you "read" tongue movements? Evaluation of the contribution of tongue display to speech understanding. Speech Communication, 52(6), 493–503. https://doi.org/10.1016/j.specom.2010.03.002

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.

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

Bruner, J. (1983). Child's talk: Learning to use language. New York: W.W. Norton.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. Developmental Psychobiology, 45(4), 204–220. https://doi.org/10.1002/dev.20032

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

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. (1987). Do zebra finch males that have been raised by another species still tend to select a conspecific song tutor? Animal Behaviour, 35(5), 1347–1355. https://doi.org/10.1016/S0003-3472(87)80007-6

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

Franz, M., & Goller, F. (2002). Respiratory units of motor production and song imitation in the zebra finch. Journal of Neurobiology, 51(2), 129–141. https://doi.org/10.1002/neu.10043

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

Halfwerk, W., Varkevisser, J., Simon, R., Mendoza, E., Scharff, C., & Riebel, K. (2019). Toward testing for multimodal perception of mating signals. Frontiers in Ecology and Evolution, 7, 2013–2019. https://doi.org/10.3389/

fevo.2019.00124
Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. Speech Communication, 47(3), 360–378. https://doi.org/10.1016/j.specom.2005.04.007
Higham, J. P., & Hebets, E. A. (2013). An introduction to multimodal communication. Behavioral Ecology and Sociobiology, 67(9), 1381–1388. https://doi.org/10.1007/s00265-013-1590-x
Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. Journal of Speech, Language, and Hearing Research : JSLHR, 53(April), 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243)
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
Kuhl, P. K. (2003). Human speech and birdsong: communication and the social brain. Proceedings of the National Academy of Sciences of the United States of America, 100(17), 9645–9646. https://doi.org/10.1073/pnas.1733998100
Kuhl, P. K. (2007). Is speech learning "gated" by the social brain? Developmental Science, 10(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x
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
Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. Proceedings of the National Academy of Sciences of the United States of America, 100(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100
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
Liu, Y., Massaro, D. W., Chen, T. H., Chan, D., & Perfetti, C. (2007). Using visual speech for training chinese pronunciation: an in-vivo experiment. SLaTE Workshop on Speech and Language Technology in Education. ISCA Tutorial and Research Workshop. The Summit Inn, Farmington, Pennsylvania USA, (SLaTE), 29–32. Retrieved from http://www.isca-speech.org/archive_open/archive_papers/slate_2007/sle7_029.pdf
Mann, N. I., & Slater, P. J. B. (1995). Song tutor choice by zebra finches in aviaries. Animal Behaviour, 49(3), 811–820. https://doi.org/10.1016/0003-3472(95)80212-6
Mann, N. I., Slater, P. J. B., Eales, L. A., & Richards, C. (1991). The influence of visual stimuli on song tutor choice in the zebra finch, *Taeniopygia guttata*. Animal Behaviour, 42(2), 285–293. https://doi.org/10.1016/S0003-3472(05)80560-3
Mello, C. V. (2014). The zebra finch, *Taeniopygia guttata*: An avian model for inves-

tigating the neurobiological basis of vocal learning. Cold Spring Harbor Protocols, 2014(12), 1237–1242. https://doi.org/10.1101/pdb.emo084574

Ménard, L., Dupont, S., Baum, S. R., & Aubin, J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults. The Journal of the Acoustical Society of America, 126(3), 1406–1414. https://doi.org/10.1121/1.3158930

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. Journal of the Acoustical Society of America, 82(6), 2145–2147. https://doi.org/10.1121/1.395659

Ohms, V. R., Snelderwaard, P. C., ten Cate, C., & Beckers, G. J. L. (2010). Vocal tract articulation in zebra finches. PLoS ONE, 5(7). https://doi.org/10.1371/ journal.pone.0011923

Partan, S., & Marler, P. (1999). Communication goes multimodal. Science, 283, 1272–1274. https://doi.org/0.1126/science.283.5406.1272

Podos, J., Southall, J. A., & Rossi-Santos, M. R. (2004). Vocal mechanics in Darwin's finches: Correlation of beak gape and song frequency. Journal of Experimental Biology, 207(4), 607–619. https://doi.org/10.1242/jeb.00770

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. Child Development, 85(3), 956–970. https://doi.org/10.1016/j.surg.2006.10.010.Use

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., & Halfwerk, 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

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26(2), 212–215. https:// doi.org/10.1121/1.1907309

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
Wang, X., Hueber, T., & Badin, P. (2014). On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French. 10th International Seminar on Speech Production, 449–452.
Westneat, M. W., Long, J. H., Hoese, W., & Nowicki, S. (1993). Kinematics of birdsong: functional correlation of cranial movements and acoustic features in sparrows. The Journal of Experimental Biology, 182, 147–171.
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

Chapter 2

Multimodal cues in songbird vocal learning provide perspective on discrepancies between live and audio-only tutoring

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

Abstract
Bird song is a particularly well-characterised example of a socially learned vocal behaviour in non-human animals with striking analogies to human speech acquisition. Bird song learning is highly accessible to experimental manipulation, and audio playback experiments have been instrumental in the study of song learning. However, many songbird species learn less well from song playbacks than from live tutors. It is often assumed that this is because social interaction with a tutor is essential for song learning. This view has been criticised by several authors, stressing the differences between live and tape tutors in non-social dimensions such as contingencies and variability. We here want to raise awareness for the unimodal versus multimodal contrast between tape and live tutors, that constitutes for an additional overlooked dimension in this debate. Birdsong, like many animal signals and human speech, is accompanied by visual components, and thus a multi- rather than unimodal signal. A case in point is the zebra finch, *Taenyiopygia guttata*, the foremost neuroethological model for vocal learning and an often-cited example for the importance of social interactions in song learning. Reviewing zebra finch song learning studies shows that research to date has not systematically differentiated between 'social' and 'multimodal' tutoring, but outcomes are often in line with the hypothesis that vocal learning may be facilitated by multimodal experiences with the signal. We conclude with an appeal and suggestions to systematically test this hypothesis regarding fundamental mechanisms in a cultural transmission process thought to be at the base of the evolution of complex communication systems.

I Introduction

Songbirds are versatile vocal learners. Bird song is a prominent example of a vocally learned signal in non-human animals (Catchpole & Slater, 1995) and often used as a model for human speech acquisition, because of the many similarities in the development of human speech and bird song (Bolhuis, Okanoya, & Scharff, 2010; Doupe & Kuhl, 1999) and the increasing interest in understanding the role of learned communication systems in the evolution of cultural transmission and cumulative culture (Whiten, 2021). Vocal learning in birds has been more extensively studied than in any other animal group given the experimental tractability of the system and early introduction of suitable learning paradigms. Thorpe's pioneering studies in the chaffinch, *Fringilla coelebs*, (e.g. Thorpe, 1954; fully reviewed in Riebel, Lachlan, & Slater, 2015) introduced 'tape tutoring' (playback of pre-recorded song via loudspeakers) in combination with the analyses of sound spectrograms for the systematic study of vocal development. Tape tutoring provides excellent stimulus control and this paradigm started modern bird song research. With increasing use of this

approach it became apparent that many species learned less well from tape than from social tutors (e.g. Kroodsma & Verner, 1978; Thielcke, 1984; Baptista & Petrinovich, 1986; Kroodsma & Pickert, 1984; Waser & Marler, 1977), presumably because the signal lacked social salience as tutees could not socially interact with these tape tutors (Baptista & Gaunt, 1997; Catchpole & Slater, 1995; Slater, Eales & Clayton, 1988; Soma, 2011). However, other differences between live and tape tutors could be decisive and the all-importance of social interaction has remained a debated issue (e.g. Beecher, 2017; Nelson, 1997, 1998). For example, tape tutoring often consists of non-interactive exposure to looped song sequences and is thus more stereotyped and lacking the context, diurnal variability and possible contingencies of real singing (e.g. Baptista & Gaunt, 1997; Beecher, 2017; Houx & ten Cate, 1999a; Nelson, 1997, 1998). A dimension that has seen little systematic attention, is that birdsong, like many mating signals, is often multimodal (Halfwerk et al., 2019; Hebets & Papaj, 2005; Partan & Marler, 1999; Rowe, 1999) whereas a (classic) tape tutor is unimodal, providing audio-only exposure (see Table 1). We suggest a systematic investigation into the presence of multimodal cues during song exposure as an alternative (non-mutually exclusive) explanation for the often improved learning with live compared to tape tutors. Singing is accompanied by visual components, such as beak, throat, head and body movements. From other contexts, it is well documented that multi-component signals (in one or multiple modalities) can increase salience by improving detection and memorisation by receivers compared to single component signals (reviewed in Rowe, 1999). Bats, for example, learn to avoid warning signals of noxious fireflies faster with multi- than with unimodal warning signals (Leavell et al., 2018). Starlings, *Sturnus vulgaris,* perform better in temporal order judgement when auditory stimuli are preceded by visual cues (Feenders, Kato, Borzeszkowski, & Klump, 2017). Auditory filial imprinting in birds is enhanced with visual stimulation (van Kampen & Bolhuis, 1991; van Kampen & Bolhuis, 1993) and nightingales, *Luscinia megarhynchos*, learn songs from audio-only playback less well than songs combined with light flashes (Hultsch, Schleuss, & Todt, 1999).

In 6 month old human infants, fixating more on mouth movements during interactions is associated with higher levels of expressive language at age 2 (Young, Merin, Rogers, & Ozonoff, 2009) and visual speech enhances learning of phoneme contrasts (Teinonen, Aslin, Alku, & Csibra, 2008). In 12-month-olds, hearing an unknown rather than their own language increases how much infants watch a speaker's mouth (Lewkowicz & Hansen-Tift, 2012). Inspired by the human literature, and given the well-established parallels between human speech and avian song acquisition (Bolhuis et al., 2010; Doupe & Kuhl, 1999;

Soha & Peters, 2015), this review investigates whether uni- versus multimodal exposure could have been a systematic confound of comparisons of live versus tape tutors by focussing on the foremost neuroethological model for avian vocal learning, the zebra finch (Griffith & Buchanan, 2010; Mello, 2014).

Table 1. Overview of the sensory stimulation, contingencies and social interactions experienced by tutees exposed to a live tutor versus different types of tape tutoring methods. In this table, 'sensory stimulation' refers to auditory, visual, tactile or olfactory sensory exposure to an adult conspecific tutor, 'contingencies' refer to auditory, visual or tactile actions (by the tutor or the tutee) that can predict exposure to tutor song and 'social interactions' refer to social companionship, auditory social tutor-tutee interaction or visual social tutor-tutee interaction.

	Sensory stimulation				Contingencies*			Social interactions		
Tutoring method								company		
Live	+	+	+	+	+	+	+	+	+	+
Passive tape	+									
Operant tape	+						+			
Tape with vocal interaction	+								+	
Tape with visual stimulation	+					+				

*contingencies can arise if tutor behaviour reliably predicts song or if song is triggered by vocal, visual or physical actions by the tutee.

II Vocal learning: the zebra finch as model

Both sexes in zebra finches have an extensive call repertoire, but only males produce courtship song which consists of a string of motifs, i.e. a stereotyped sequence of individual sound elements called syllables (see Figure 1A). Motifs are learned with varying accuracy from adult conspecifics (see Figure 1B) during a sensitive period lasting roughly from 20-65 days post hatching (dph) (Eales, 1985; Gobes, Jennings, & Maeda, 2017; Immelmann, 1969; Jones, ten Cate, & Slater, 1996; Roper & Zann, 2006). Without suitable models, zebra finches develop impoverished song (reviewed in Slater et al., 1988). The exact mechanisms underlying differences in copying between individuals within and between tutoring settings are an unresolved issue (Derégnaucourt, 2011; Gobes et al., 2017; Houx, Feuth, & ten Cate, 2000; Slater, Eales, & Clayton, 1988). Next to the white crowned sparrow (Baptista & Petrinovich, 1986), zebra finches are almost generically referred to exemplify impoverished learning from tape versus live tutoring (reviewed in Derégnaucourt, 2011; Slater, Eales, & Clayton, 1988). The generally favoured explanation for this is that social interaction with the tutor facilitates song learning (e.g. Chen, Matheson, & Sakata, 2016; Derégnaucourt, Poirier, van der Kant, & van der Linden, 2013; Slater et al.,

1988, but see e.g. Nelson, 1997).

We here systematically revisit the extensive literature on zebra finch song learning (Gobes, Jennings, & Maeda, 2017; Griffith & Buchanan, 2010; Slater, Eales, & Clayton, 1988) to check whether social and multimodal tutoring were always combined or whether studies manipulated these dimensions separately. The results suggest that future studies should more systematically study whether multimodal cues enhance song learning independently or on top of social properties of a live tutor.

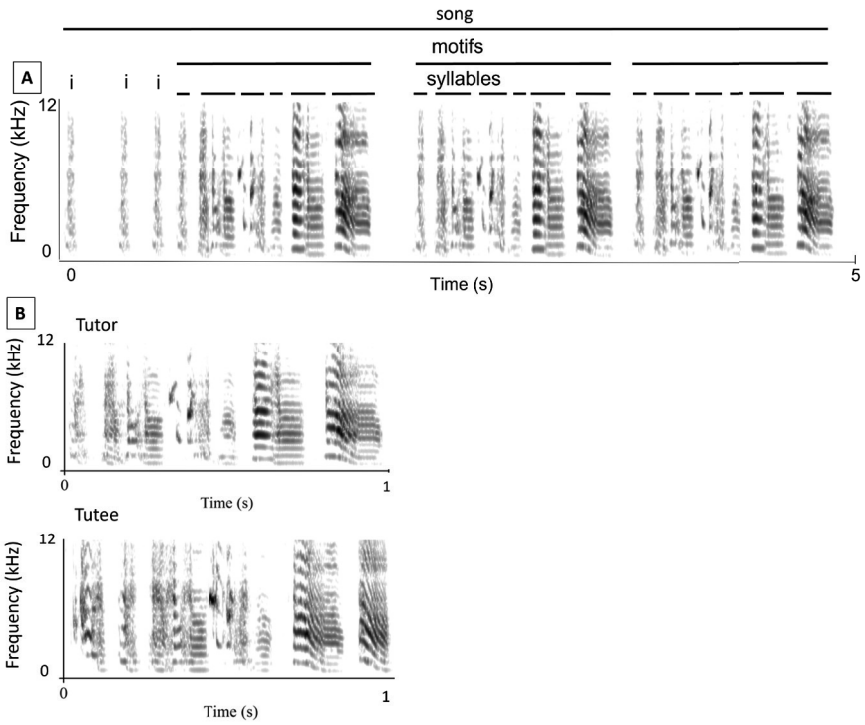

Figure 1. (A) Spectrogram of zebra finch song (also referred to as strophe) with three motif repetitions. Introductory notes are indicated with 'i'. Song units within motifs that are separated by silent intervals are called syllables (Sossinka & Böhner, 1980).The motif of this bird consists of 6 different syllables. (B) Example of the motif of a tutor above and the motif of a tutee that was exposed to this tutor's song during its sensitive period for song learning below. This tutee copied several syllables from the tutor. In most song learning studies, pupils received a particular type of tutoring during the sensitive period for song learning. When the pupil is adult, the amount of its song that matches the tutor song is set equal to the amount of song that the pupil has copied from the tutor. This is usually taken as a measure of song learning success (but see e.g. Geberzahn, Hultsch, and Todt (2002) that this might underestimate learning).

III The visual dimensions of singing

Acoustic signals require rhythmic mechanical movements to set the physi-

cal carrier of the signal (air, water, substrate) in motion, which often leads to visual components obligatorily coupled with specific sounds, e.g. human lip or anuran air sac movements (Bradbury & Vehrencamp, 2011). Birdsong is also accompanied by such obligatory visual components arising mainly from the beak and throat movements of song production (Goller, Mallinckrodt, & Torti, 2002; Ohms, Snelderwaard, ten Cate, & Beckers, 2010; Williams, 2001). There are additional 'free' signal components, like dance and wing movements, that co-occur with but are not inextricably linked to song production (Dalziell et al., 2013; Ota, Gahr, & Soma, 2015; Ullrich, Norton, & Scharff, 2016; Williams, 2001). In zebra finches, the beak and dance movements of tutors and tutees are more similar than those of unrelated males (Williams, 2001), which is consistent with tutees attending to both auditory and visual components during vocal learning. Visual cues also guide tutor choice: zebra finches preferentially choose tutors visually resembling the colour morph of birds that reared them (Mann & Slater, 1995; Mann et al., 1991). Mate recognition is also enhanced by a correct match between a male's morph and song (Brazas & Shimizu, 2002; Campbell & Hauber, 2009). The combination of visual and auditory information can thus influence bird behaviour and it can facilitate song learning (Hultsch et al., 1999; Todt, Hultsch, & Heike, 1979). In view of these observations, we hypothesize that multimodality of a live tutor might facilitate song learning independently of the live tutor's 'social properties'. In the following sections, we revisit the zebra finch song learning literature comparing tutoring paradigms to ask whether studies conclusively show that social interactions - rather than differences in (multimodal) stimulus properties of tutor song - increase the salience of song models.

IV Comparing live and tape tutors across modalities

Auditory modality

Live and tape tutors could in principle provide identical auditory input, e.g. if playback is established via an audio link from a live tutor. However, most playback tutoring repeats pre-recorded song sequences (e.g. Adret, 1993a; Houx & ten Cate, 1999b; Tchernichovski, Mitra, Lints, & Nottebohm, 2001). In stark contrast, live tutors vary pitch, tempo, amplitude, number of syllables, motifs and introductory notes during singing (Glaze & Troyer, 2006; Helekar, Marsh, Viswanath, & Rosenfield, 2000) and differently so to male, female or juvenile audiences (Chen et al., 2016; Hyland Bruno & Tchernichovski, 2019; Jesse & Riebel, 2012; Williams, 2004). Variable song exposure increases stimulus engagement in females (Collins, 1999) and might prevent habituation (Krebs & Kroodsma, 1980), but we are unaware of systematic investigations whether and how variability of the stimulus affects song learning. Studies exposing tutees

to live tutors and using audio links to yoked controls (Chen et al., 2016; Eales, 1989) can provide clues here, as they eliminate differences in song variability among live and tape tutors. In these two studies, young birds learned more from live tutors (Chen et al., 2016; Eales, 1989). An observational study suggests that variability of tutor song predicts how well biological and foster offspring learn (Tchernichovski, Eisenberg-Edidin, & Jarvis, 2021). The next step will be to find out whether variability of tutor song is causal or covarying with tutor (song) properties. Future studies could test this by manipulating stereotypy and variability between and within different tutoring methods.

Visual modality

Visual access to the tutor could affect song learning via several processes. Seeing a tutor provides social stimulation and social relevance of the auditory stimulus, which might benefit the tutee's welfare or motivation to learn. Seeing the visual correlates of song production might increase the salience of the auditory signal and draw the pupil's attention towards the song or, as in humans, birds might experience improved reception and perception: in noisy conditions, seeing orofacial articulatory movements improves speech perception (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007).

In several experiments, visual tutor models, such as pictures (Funabiki & Funabiki, 2008) or plastic models (Tchernichovski et al., 2000), were offered during passive or operant tape tutoring. However, these studies did not test whether adding visual models improved song learning and had no control groups without visuals. Deshpande and colleagues (2014) used a minimalistic single session operant tutoring design (75 s tutoring during a 2 h session using a single tutor across tutees) to capture song template formation. During a single session (at 35 or 45 dph), tutees had operant control over audio or audio-visual song playbacks where the visual stimulus was either preceding (VA), concurrent with or following audio (AV) playback. Only tutees from the simultaneous audio-visual or AV conditions showed significantly more song learning than untutored birds. Song learning was low overall, probably due to the minimal exposure level. However, the results support the hypothesis that visual exposure to a singing tutor can facilitate song learning. Further studies using replicate high-quality video tutors throughout the sensitive period could systematically test whether the addition of a video to a tape tutor leads to increased song learning outside a single session tutoring design. Price (1979) compared song learning in five males raised with two live tutors behind a visual separation to learning in a male that was not visually separated from a live tutor. The visually separated birds copied few song elements, while the male with visual access

produced only elements copied from its tutor. Although it is hard to draw conclusions from such a small sample size, this study does suggest that visual exposure to a tutor facilitates song learning in zebra finches.

Other modalities

Multicomponent signals generally lead to better detection and learning than single component signals (reviewed in Rowe, 1999). Olfactory and gustatory cues enhance the learning of visual warning signals in chicken (Rowe & Guilford 1996). Zebra finches show olfactory guided natal nest and kin recognition (reviewed in Krause et al., 2018). Hence, stimulation in other than auditory and visual modalities might influence song learning. It is feasible that olfactory components of the tutor guide tutor choice, or reinforce proximity, but as yet this has not been studied. Studies have, however, investigated the effect of physical interactions on song learning. Eales (1985) suggested that physical contact with a live tutor is unnecessary for song learning after observing song learning in tutees separated from their tutor by wire mesh. However, Adret (1992) found poor copying from a tutor at 50 cm away and hypothesized that this was due to a lack of physical interaction. We could not find studies directly comparing song learning in tutees that could or could not physically interact with the same tutor. This issue is far from resolved, as physical tutor-tutee interactions are frequent (Adret, 2004; Clayton, 1987; Mann & Slater, 1995; Morris, 1954). Future studies will thus have to investigate a potential influence of olfactory and tactile cues on song learning.

V Contingencies with tutor song exposure: live versus tape tutors

During live tutoring, the tutor's or tutee's behaviour might predict tutor song exposure, possibly facilitating attention to tutor song and song learning. For instance, tutor song might always be preceded by specific tutor behaviour. The absence of such contingencies is a pronounced and systematic difference between tape and live tutoring (Houx & ten Cate, 1999a; ten Cate, 1991).

Vocal contingencies with tutor song

Searching for contingencies between tutor behaviour and singing, Houx and ten Cate (1998) found the only tutor behaviour to predict singing was tutor song itself, as tutors usually produce song in bouts. Calls were not recorded in this study, but might also provide vocal contingencies with tutor song. Tutees that had an audio-link to a live tutor were passively exposed to all of the tutor's vocalizations, but still copied less from the tutor than tutees co-housed with the tutor (Chen et al., 2016; Eales, 1989). Evidence from other species suggests such contingencies might affect song learning – white crowned sparrows learn

better if songs start with a specific acoustic cue (Soha & Marler, 2000).

Visual contingencies with tutor song

Investigating the effect of visual contingencies with tutor song on zebra finch song learning, two studies found that groups receiving visual stimulation contigent on song did not learn more than those that received audio playback only (Bolhuis, van Mil, & Houx, 1999; Houx & ten Cate, 1999a): the visual stimulus was a taxidermic mount of an adult male zebra finch that was revealed right before or after song playbacks (Houx & ten Cate, 1999a) or in another study during and after playbacks (Bolhuis, van Mil, & Houx, 1999). Although a cylinder that was raised and lowered provided some motion, the taxidermic mount itself was stationary (Bolhuis et al., 1999; Houx & ten Cate, 1999a). A temporary coupling between auditory and visual stimulation may be necessary to facilitate song learning. Nightingales (Hultsch et al., 1999), for example, learned better from song playbacks presented with a synchronously flashing stroboscope than from songs presented without stroboscope flashes. A similar temporary coupling of auditory and visual stimulation might occur in the beak or body movements of a singing bird. We thus hypothesize that visual contingencies with song exposure might be one of the factors facilitating song learning from a live tutor.

Contingencies with tutee behaviour

Bird song and language learning have both been suggested to be a form of operant conditioning, where young learners experience reinforcement when motivated to hear vocalizations and actively elicit exposure to speech or song (language: Sturdy & Nicoladis, 2017, birdsong: Adret, 1993a). Operant tasks allow to test these ideas experimentally. The most powerful tests to date used paired designs, where one bird could actively trigger playback by perch hopping or key pecking and a yoked control simultaneously received the same playback via a second loudspeaker. Experimental birds (with operant control over playback) copied more song than their yoked controls in one study (Adret, 1993), but not in other studies (Houx & ten Cate, 1999b; ten Cate, 1991), possibly because otherwise single housed experimental and control birds could vocally interact in one study (Adret 1993). The interaction between the birds in combination with exposure to the operant tape tutor, might have affected song copying success (Houx & ten Cate, 1999b). However, small sample sizes (n = 3 in Adret (1993)), and lack of a live control condition make it difficult to settle these questions. Several subsequent studies have successfully used operant tutoring and reported substantial song imitation compared to untutored birds (Derégnaucourt, Mitra, Fehér, Pytte, & Tchernichovski, 2005; Phan, Pytte, &

Vicario, 2006; Tchernichovski, Mitra, Lints, & Nottebohm, 2001). Unfortunately, none of these studies included a control treatment involving passive audio exposure to the songs used for the operant tutoring, which supposedly leads to higher tutor song similarity than in untutored birds (Chen et al., 2016). Additional differences also hamper direct comparison between these operant and standard tape tutoring paradigms. First, there was potential additional visual stimulation, as the playback loudspeaker was hidden inside a plastic zebra finch model in the tutee's cage. Second, restricted song exposure (maximum 20 reinforced key pecks per day) creates a variable reward schedule which can be more reinforcing than continuous rewarding (Ferster & Skinner, 1957).

We found one study directly comparing live, operant and tape tutors, where the tutees in the operant and passive tape tutoring condition also got restricted song exposure (Derégnaucourt et al., 2013). Birds from these two conditions, once adult, had a lower similarity to the tutor song than birds from the live condition. However, there were several differences between the live and the operant or passive tape tutoring paradigm (for instance, variable versus stereotyped song exposure and raised with a social companion versus in social isolation, see Table 2), making it difficult to discern which factors contributed to this difference. The songs of the operantly trained birds were significantly more similar to the tutor song than the songs from the passively exposed birds. This suggests that active control over song exposure or the partial reward scheme positively affected song learning.

VI Social companionship versus social isolation

An aspect of live tutoring that has to date seen little attention is that the mere presence of a companion could affect song development. Live tutors provide social company while tape-tutored birds are often housed in social isolation, which could affect a bird's hormonal and physiological state and consequently song learning. Adret (2004), for example, observed that tutees co-housed with a female in addition to their tutor learned better than those without. Also, birds reared in song isolation show better song learning when reared with peers than when reared alone (Jones et al., 1996). While this is generally interpreted as using other isolate males' song as model, the presence of social companions could improve song learning by several other mechanisms. In social animals like zebra finches, social isolation might have a negative effect on welfare (e.g. corticosterone levels in the blood of zebra finches that were socially isolated for 10 minutes are significantly higher than baseline levels (Banerjee & Adkins-Regan, 2011)) and on the motivation to practice and learn song. Zebra finches produce more song with a male or female companion, compared to socially

isolated housing (Jesse & Riebel, 2012). Less singing could reduce practicing during motor learning. In young zebra finches that produce immature songs, a female conspecific can elicit songs with more mature properties (Kojima & Doupe, 2011). Companions could thus lead to more practice which is crucial as demonstrated by temporarily pharmacological blocking of vocal motor control during late motor practice (but not other ages) which impairs learning (Pytte & Suthers, 2000). Social companions could not only encourage practice but also guide song development. In cowbirds, *Molothrus ater ater*, non-singing females shape male song production (West & King, 1988) and there is increasing evidence that female zebra finches might affect male song learning (Carouso-Peck & Goldstein, 2019; Jones & Slater, 1993; Williams, 2004). For example, zebra finches were found to learn better if housed with a companion than when housed alone or with a deaf female companion (Williams, 2004). Also, non-vocal feedback (fluff-ups performed by the mother before, during or after tutee song production) was positively correlated with song learning success (Carouso-Peck, Menyhart, DeVoogd, & Goldstein, 2020). This suggests that reactions of companions to the tutee's song play a role in the song learning process.

Rearing in isolation from conspecifics also affects adult auditory discrimination, e.g. birds reared in isolation perform worse in auditory discrimination tasks than socially reared birds (Sturdy, Phillmore, Sartor, & Weisman, 2001). To investigate the effect of social interaction with a tutor on neuronal responsiveness in the auditory cortex, juvenile zebra finches were exposed to playbacks of their tutor's song while in social isolation or paired with their tutors (Yanagihara & Yazaki-Sugiyama, 2019). In the juveniles paired with their tutor, but not in the juveniles in social isolation, neurons exhibited selective auditory responses to the playbacks. Social isolation was only compared with the tutor -present situation, making it impossible to disambiguate whether this effect arose from having a social companion in general, or more specifically from being able to socially interact with the tutor during song exposure.
Albeit as yet not a subject of systematic study, the combined indirect evidence from the studies discussed above suggests that social companionship per se, which is absent in standard tape tutoring settings, may affect song development and learning in zebra finches. Future studies should aim for a comparable social environment in different tutoring treatments.

VII Tutor reaction to tutee song

In live tutoring paradigms, tutors might respond vocally or visually to tutee vocalizations, thereby reinforcing particular song elements or singing behaviour. Tape-tutored tutees do not receive tutor feedback on their vocalizations.

Tutor reacts vocally to tutee song

Among adult zebra finches, vocal interactions can influence singing: males are more likely to alter or end a motif if a female conspecific calls while they are singing (Williams, 2004). While the idea of reinforcement by vocal tutor reactions is appealing theoretically, detailed observations of tutor/tutee interactions showed no vocal contingent tutor reactions, defined as tutor behaviour occurring more often within 15 seconds after tutee song onset than expected by chance (Houx & ten Cate, 1998). However, the magnitude of father's singing responses to their sons singing was positively correlated with sons' song learning accuracy in another study (Carouso-Peck et al., 2020).

Eales (1989) compared song learning in zebra finches in three different treatments: birds that could visually and vocally interact with a tutor in an adjoining cage all learned at least some elements from this tutor. In a group that could interact only vocally with a tutor behind an opaque screen, four birds copied elements from the tutor, while three birds copied elements they heard before 35 dph. None of the birds that could only hear the tutor's song from a loudspeaker copied from it. This suggests that vocal interaction facilitated song learning to some degree. However, as pointed out by Nelson (1997), interpretation of these results is difficult as the birds that could vocally and visually interact with the tutor were housed in a room with many conspecifics, while birds in the other two groups were housed in sound-isolation boxes. Using both a one-way and two-way audio-link, i.e. one link that gave a tutee passive tutor exposure and one that allowed vocal tutor-tutee interaction, Chen et al., (2016) compared song learning between vocally interactive and non-interactively tutored birds, but found no differences in song learning success. Song copying was, however, poor in all groups, as exposure was limited to only one day. These results are therefore not comparable to standard live tutoring situations. Further studies are necessary to find out to what extent vocal exchanges between tutors and tutees affect vocal learning.

Tutor reacts visually to tutee song

An observational study found no tutor behaviour to occur more often than expected after tutee song (Houx & ten Cate, 1998). There is no experimental study investigating this question involving male tutors, but female cowbirds reinforce the production of specific song elements by wing movements (West & King, 1988). Likewise, young zebra finches receiving contingent visual feedback (video playback) of fluff-up behaviour by a female on their immature song production (Carouso-Peck & Goldstein, 2019) were found to copy more tutor

song than birds receiving non-contingent visual feedback. Live tutors could also provide visual feedback to song learning juveniles. This might be one of the factors facilitating song learning from a live tutor.

VIII Is social interaction or multimodal exposure the key difference between live and tape tutors?

As evidenced from the previous paragraphs and Table 2, we have as yet not unambiguously identified which stimulus properties of a live tutor improve learning compared to tape tutoring paradigms. Owing to the logistics of tutoring experiments and research interests, most studies investigated single stimulus dimensions, but the facilitating effect of a live tutor might arise from a combination of factors. The most favoured hypothesis in the literature is that the social interaction between tutor and tutee is decisive for song learning (Baptista & Petrinovich, 1986; Catchpole & Slater, 1995; Slater, Eales, & Clayton, 1988) and also for other forms of channelled learning such as filial and sexual imprinting (ten Cate, 1994). A mechanism here might be that social interactions enhance attention (Chen et al., 2016) or engage birds in more practice (Jesse & Riebel, 2012) and thereby promote learning. This raises the question which stimulus properties of a 'social interaction' are decisive (Nelson, 1997). Table 2 shows that many studies did not systematically control for the fact that live song exposure is multimodal whereas taped song is not. Much of the evidence for 'social interaction' with careful re-evaluation could also be interpreted as evidence for higher stimulus salience arising from multimodality or contingencies. Moreover, more time with a stimulus leads to more interactions – we can thus not establish any causality from observations showing that birds learned more from the male they interacted with most (Eales, 1987; Williams, 1990) or that showed most aggressive behaviour towards them (Clayton, 1987, but see Houx & ten Cate, 1998; Mann & Slater, 1995; Williams, 1990), as this could be a matter of total time in close proximity (Slater & Richards, 1990; Mann & Slater, 1995). Active approaches between tutors and tutees after the tutor started singing (Houx & ten Cate, 1998) can be interpreted as social attraction, but also as attraction to the multidimensional properties of song. Closer and longer proximity and more interaction also mean more opportunity for the tutee to see the tutor singing and these observations thus support either the social or multimodal hypothesis. Similarly, the activation of a mesocortical dopamine circuit by the presence of a singing tutor might be related to social aspects of the tutor but also to multimodal exposure to it (Tanaka, Sun, Li, & Mooney, 2018).

Some studies have tried to address the effect of social interaction with a tu-

tor on zebra finch song learning: testing whether social interaction affects the duration of the sensitive period for song learning, Morrison and Nottebohm (1993) compared tutees visually separated from their tutor by cardboard dividers, social + song isolated tutees and aviary-reared control birds. Unlike the control birds, the other two groups learned from a new unrelated tutor beyond the normal closure of the sensitive phase, which the authors interpreted to show that lack of visually guided social interactions delayed the closure of the sensitive phase. However, the mere visual exposure to the singing tutor could have been crucial. Chen et al. (2016) housed young zebra finches in social isolation during the sensitive period, except for five days when a live tutor was housed next to them. Individuals in one treatment could vocally and visually interact with the tutor, while individuals in a yoked control treatment could only hear the tutor from a loudspeaker. Song in these birds was more similar to the tutor song than song in untutored birds, and more so in the live than in the passively tutored birds. Chen et al. (2016) conclude that higher song copying success in the live tutored group is caused by the social interactions with the tutor in this condition, but again, the results are also in line with the hypothesis that song learning could improve because of multimodal exposure to the tutor. A study investigating song learning in zebra finch tutees housed in an aviary with peers, with a visible and audible adult tutor housed in a separate cage outside of the aviary, found that the adult song of the tutees resembled the song of their peers more than the song of their tutor (Honarmand, Riebel, & Naguib, 2015). This suggests that the ability to socially interact with peers made the tutees more likely to learn from them than from the tutor. We agree that social interaction is probably important for song learning, however, we argue that simpler mechanisms such as visual exposure to a tutor should be investigated as possible additional contributing factors to zebra finch song learning.

IX A proposed framework for disentangling multimodal and social dimensions

Overall, studies found more tutor song copying by live than tape-tutored tutees (see Table 2) with the exception of one study using a plastic tutor bird and restricted song exposure (Phan et al., 2006). From the many studies and different learning outcomes, we could not identify one single key factor systematically associated with the difference in song learning from live or tape tutors. Probably, a combination of factors associated with a live tutor has a facilitating effect on song learning. As we hypothesized, one of these factors might be that a live tutor offers multimodal exposure to song. Revisiting the literature has shown great asymmetries in uni- versus multimodal tutoring regimes for the majority of non-live tutoring approaches (see Table 2). Song tutoring studies have not

systematically controlled for this dimension and the results of many studies are in line with both the interpretations that social interaction or multimodal exposure to song facilitated song learning. Several authors have pointed out the potential importance of both auditory and visual stimulation for zebra finch song learning (Adret, 1992, 1997; Bolhuis et al., 1999; Carouso-Peck & Goldstein, 2019; Houx & ten Cate, 1999a). However, as Table 2 shows, in all studies to date (but Deshpande et al. 2014) visual interaction with a tutor and multimodal exposure to tutor song were confounded: in all conditions with auditory and visual tutor exposure, tutees were also able to socially interact with the tutor. Future studies should thus aim at investigating visual interaction and multimodal exposure separately to test their effects on song learning.

To investigate potential beneficial effects of multimodal tutor exposure on song learning, live tutored birds should be compared with birds that can also see and hear a singing tutor, but without being able to socially interact with it. This can be achieved with one-way audio-visual links, for example using a one-way mirror such that a tutee can see the tutor, while the tutor cannot see the tutee. If this tutee copies more from the tutor than a tutee receiving the same auditory input without visual access to the tutor, this might indicate that multimodal cues facilitate song learning. However, it is then impossible to disentangle the effect of mere visual exposure to a moving and singing adult, which might affect the tutee's general motivation or attention, from the effect of exposure to the specific visual correlates of song production, e.g. the tutor's beak and head movements. To focus on exposure to the visual component of song, one could use artificial tutors, such as videos of singing zebra finches. Creating these videos is relatively easy, although one should be aware that standard video systems are developed for human vision, which makes it important to adjust e.g. colours and frame rate to make them suitable for bird vision (Chouinard-Thuly et al., 2017; Fleishman & Endler, 2000; Oliveira et al., 2000; Tedore & Johnsen, 2017). We are aware of three studies that have tutored zebra finches with videos (Adret, 1997; Deshpande, Pirlepesov, & Lints, 2014; Ljubičić, Hyland Bruno, & Tchernichovski, 2016). Passive and operant exposure to a video tutor led to impoverished adult songs in the pioneering study by Adret (1997), who worried that the sound playback through the TV monitor loudspeakers was distorted and so impaired learning. With very low amounts of exposure to another video tutor (only 75 seconds in total), there was very little learning (Deshpande et al., 2014). Finally, preliminary results suggest that zebra finches can adapt the pitch of already learned syllables according to the song of a video tutor singing toward the tutee (Ljubičić et al., 2016). However, as yet no study has compared birds exposed to the same live and videotaped tutor.

An important caveat regarding videos studies is their limited two-dimensional representation of a bird. The lack of depth might influence the salience of visual cues and thereby their effect on song learning. This could be overcome by using moving 3-dimensional models, such as robotic birds which have already been used successfully in studies on the importance of multimodal song in territory defence (Anderson, DuBois, Piech, Searcy, & Nowicki, 2013, Ręk & Magrath, 2016), suggesting they could be applied in song tutoring studies as well (Simon et al., 2019).

This review mainly focused on production learning, but additional perspectives arise from preference learning. Both male and female zebra finches develop preferences for tutor over unfamiliar songs from live (Riebel, Smallegange, Terpstra, & Bolhuis, 2002) and tape tutors (Holveck & Riebel, 2014; Houx & ten Cate, 1999a, b; Riebel, 2000). In tape-tutored males, song production learning (amount of elements copied) does not predict preference strength for this tutor song (Houx & ten Cate, 1999b), indicating that different mechanisms might be involved in preference and production learning. In males, no direct comparison has been made so far between preference learning from live and tape tutors. In females, preference learning did not differ between live and tape tutors (Holveck & Riebel, 2014). However, the females were housed in pairs and exposed to a series of a live tutor before 35 days and then two tape tutors between 35 and 65 days. It would be premature to conclude that preference and production learning or female and male learning differ in how susceptible they are to uni- versus multimodal – or interactive tutoring (Riebel, Odom, Langmore, & Hall, 2019). Controlled tutoring studies manipulating one modality at a time, as described earlier in this paragraph, and assessing production and preference learning can shed light on whether visual cues have an effect on the song learning process in general, and whether there is a difference in the effect of visual cues on song production and preference learning.

Table 2. Zebra finch tutoring studies comparing song copying across tutoring regimes. '+' means present, '-' means absent. * is a value taken from a graph, ** is calculated based on data in the article.

Study (in order of publication year)	Treatment groups	N	Learning success across groups[1]	Sensory stimulation				Contingencies			Social interaction				Tutoring [days]	Scoring song copying[2]	Mean % ± SE pupil song copied from tutor	Mean %± SE of tutor song copied by pupil[3]
				Auditory	Visual	Tactile	Olfactory	Vocal	Visual	Other	Companionship	Vocal	Visual	General				
Price 1979	1. Non-visible live tutor	5		+	-	-	+	+	-	-	?	+	-	-	110	V		
	2. Live tutor	1		+	+	-	+	+	+	+	+	+	+	+	110	V		
Eales 1989	1. One-way audio-link live tutor	8	3 > 1+2 1=2	+	-	-	-	+	-	-	-	-	-	-	85	V		
	2. Non-visible tutor	7		+	-	-	-	+	-	+	?	+	-	-	85	V		
	3. Live tutor	11		+	+	-	+	+	+	+	+	+	+	+	85	V		
Adret 1993	1. Tape tutor with vocal interaction with peers	3	2,3>1[4]	+	-	-	-	-	-	-	-	-	-	-	30	V	36±9*	39±13
	2. Operant tape tutor	3		+	-	-	-	-	-	+	-	-	-	-	30	V	25±8*	±54**
	3. Operant tape tutor with vocal interaction with peers	3		+	-	-	-	-	-	+	-	-	-	-	30	V	33±7*	76±3
Bolhuis et al. 1999	1. Tape tutor sim. with visual stimulus	7	1=2	+	-	-	-	-	+	-	-	-	-	-	41	V	53±9*	32±11*
	2. Tape tutor followed by visual stimulus	8		+	-	-	-	-	-	-	-	-	-	-	41	V	50±9*	30±9*
Houx & ten Cate 1999a	1. Tape tutor	8	1=2=3	+	-	-	-	-	-	-	-	-	-	-	31	V	43±8	
	2. Tape tutor preceded by visual stimulus	8		+	-	-	-	-	+	-	-	-	-	-	31	V	47±10	
	3. Tape tutor followed by visual stimulus	8		+	-	-	-	-	-	-	-	-	-	-	31	V		
Houx & ten Cate 1999b	1. Tape tutor	8	1=2	+	-	-	-	-	-	-	-	-	-	-	30	V		
	2. Operant tape tutor	8		+	-	-	-	-	-	+	-	-	-	-	30	V		

Adret 2004	1. Live tutor, tutee's eyes occluded	8	2>1	+	-	+	+	+	-	+	+	+	-	+	30	V		42±6
	2. Live tutor, tutee's eyes occluded, female sibling in cage	8		+	-	+	+	+	-	+	+	+	-	+	30	V		63±9
Phan et al. 2006	1. Operant tutor + plastic model	9	1=2	+	-	-	-	-	-	+	-	-	-	-	33-47	S		61±6
	2. Live tutor	11		+	+	+	+	+	+	+	+	+	+	+	33-47	S		71±4
Derégnaucourt et al. 2013	1. Tape tutor	19	3>2>1	+	-	-	-	-	-	-	-	-	-	-	57	S		60±4*
	2. Operant tape tutor	33		+	-	-	-	-	-	+	-	-	-	-	57	S		61±4*
	3. Live tutor	10		+	+	+	+	+	+	+	+	+	+	+	57	S		76±4*
Desphande et al. 2014	1. Untutored isolates	10	4A>1	-	-	-	-	-	-	-	-	-	-	-	0	S		13*[5]
	2A. Operant tape tutor[6]	6		+	-	-	-	-	-	+	-	-	-	-	1	S		16*
	3A. Operant video tutor with audio and video simultaneous	6		+	+	-	-	-	+	+	-	-	-	-	1	S		16*
	4A. Operant video tutor with audio before video	5		+	+	-	-	-	-	+	-	-	-	-	1	S		20*
	5A. Operant video tutor with video before audio	5		+	+	-	-	-	+	+	-	-	-	-	1	S		17*
	2B. Operant tape tutor	5	3B, 4B>1	+	-	-	-	-	-	+	-	-	-	-	1	S		20*
	3B. Operant video tutor with audio and video simultaneous	6		+	+	-	-	-	+	+	-	-	-	-	1	S		21*
	4B. Operant video tutor with audio before video	7		+	+	-	-	-	-	+	-	-	-	-	1	S		21*
	5B. Operant video tutor with video before audio	6		+	+	-	-	-	+	+	-	-	-	-	1	S		19*

Chen et al. 2016	1. One-way audio-link to live tutor	9	2>1 3=4=5	+	-	-	-	+	-	-	-	-	-	-	5	S		34±6*
	2. Live tutor	10		+	+	-	+	+	+	+	+	+	+	+	5	S		59±6*
	3. One-way audio-link to live tutor	6		+	-	-	-	+	-	-	-	-	-	-	<1	S		33±4*
	4. Two-way audio-link to live tutor	5		+	-	-	-	+	-	-	-	+	-	-	<1	S		34±4*
	5. Live tutor	6		+	+	-	+	+	+	+	+	+	+	+	<1	S		43±4*
Carou-so-Peck & Goldstein 2019	1. Live tutor (0-35 dph) and video non-contin-gent to tutee song	9	2>1	+	+	+	+	+	-	+	+	+	+	+	35	S		35±17
	2. Live tutor (0-35 dph) and video contingent to tutee song[6]	9		+	+	+	+	+	+	+	+	+	+	+	35	S		62±12

[1] > indicates a significant difference between groups, while = indicates that groups did not differ significantly from each other.
[2] V = Visual inspection of spectrogram, S = similarity score in Sound Analysis Pro
[3] If original papers gave decimals, these have been rounded off to standardize display
[4] Difference between 2 and 3 was not tested
[5] This study only reports a similarity score where three different SAP measures (similarity, accuracy and sequential match) are combined
[6] 2A, 3A, 4A and 5A were tutored at 35 days post-hatching, while 2B, 3B, 4B and 5B were tutored at 45 days post-hatching

X Conclusion

Re-evaluating previous song tutoring experiments showed that research to date has not systematically controlled for uni- versus multimodality of exposure when testing the importance of social interaction for zebra finch song learning. Investigating multimodal tutoring separately from social components might help to better understand the differences in stimulus properties that arise from live versus tape tutoring. By separately investigating the effect of visual cues and interaction on song learning, the contribution of multimodal cues and social interaction to the song learning process can be disentangled. Additionally, by standardizing the social environment of birds in different tutoring conditions, the effect of a social companion versus a social tutor can be investigated. Identification of relevant stimulus properties should improve our insights into the mechanisms underlying social vocal learning processes that are at the heart of the evolution of cultural transmission and cumulative culture in communication.

Acknowledgements

We would like to thank Carel ten Cate for comments on an earlier version of this manuscript and members of the Behavioural Biology group in Leiden for discussion. Funding for this review was provided by the Human Frontier Science Program (No RGP0046/2016).

References

Adret, P. (1992). Imitation du chant chez les diamants mandarins: voir, entendre et interagir. Annales de La Fondation Fyssen, 7, 73–82.

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

Adret, P. (1997). Discrimination of video images by zebra finches (*Taeniopygia gut tata*): direct evidence from song performance. Journal of Comparative Psychology, 111(2), 115–125. https://doi.org/10.1037/0735-7036.111.2.115

Adret, P. (2004). Vocal imitation in blindfolded zebra finches (*Taeniopygia guttata*) is facilitated in the presence of a non-singing conspecific female. Journal of Ethology, 22(1), 29–35. https://doi.org/10.1007/s10164-003-0094-y

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

Banerjee, S. B., & Adkins-Regan, E. (2011). Effect of isolation and conspecific presence in a novel environment on corticosterone concentrations in a social avian species, the zebra finch (*Taeniopygia guttata*). Hormones and Behavior, 60(3), 233–238. https://doi.org/10.1016/j.yhbeh.2011.05.011

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
Beecher, M. D. (2017). Birdsong learning as a social process. Animal Behaviour, 124, 233–246. https://doi.org/10.1016/j.anbehav.2016.09.001
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
Bradbury, J. W., & Vehrencamp, S. L. (2011). Principles of animal communication. Sinauer Associates, Sunderland.
Campbell, D. L. M., & Hauber, M. E. (2009). The disassociation of visual and acoustic conspecific cues decreases discrimination by female zebra finches (*Taeniopygia guttata*). Journal of Comparative Psychology, 123(3), 310–315. https://doi.org/10.1037/a0015837
Carouso-Peck, S., & Goldstein, M. H. (2019). Female social feedback reveals non-imitative mechanisms of vocal learning in zebra finches. Current Biology, 29, 631–636. https://doi.org/10.1016/j.cub.2018.12.026
Carouso-Peck, S., Menyhart, O., DeVoogd, T. J., & Goldstein, M. H. (2020). Contingent parental responses are naturally associated with zebra finch song learning. Animal Behaviour, 165, 123–132. https://doi.org/10.1016/j.anbehav.2020.04.019
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.
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
Chouinard-Thuly, L., Gierszewski, S., Rosenthal, G. G., Reader, S. M., Rieucau, G., Woo, K. L., & Witte, K. (2017). Technical and conceptual considerations for using animated stimuli in studies of animal behavior. Current Zoology, 63, 5–19. https://doi.org/10.1093/cz/zow104
Clayton, N. S. (1987). Song tutor choice in zebra finches. Animal Behaviour, 35, 714–721. https://doi.org/10.1016/0003-3472(95)80212-6
Collins, S. A. (1999). Is female preference for male repertoires due to sensory bias? Proceedings of the Royal Society B: Biological Sciences, 266(1435), 2309–2314. https://doi.org/10.1098/rspb.1999.0924
Dalziell, A. H., Peters, R. A., Cockburn, A., Dorland, A. D., Maisey, A. C., & Magrath, R. D. (2013). Dance choreography is coordinated with song repertoire in a complex avian display. Current Biology, 23(12), 1132–1135.

https://doi.org/10.1016/j.cub.2013.05.018
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., Mitra, P. P., Fehér, O., Pytte, C., & Tchernichovski, O. (2005). How sleep affects the developmental learning of bird song. Nature, 433(7027), 710–716. https://doi.org/10.1038/nature03275
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
Deshpande, M., Pirlepesov, F., & Lints, T. (2014). Rapid encoding of an internal model for imitative learning. Proceedings of the Royal Society of London. Series B: Biological Sciences, 281(1781), 20132630. https://doi.org/10.1098/rspb.2013.2630
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. (1985). Song learning in zebra finches: some effects of song model availability on what is learnt and when. Animal Behaviour, 33(4), 1293–1300. https://doi.org/10.1016/S0003-3472(85)80189-5
Eales, L. A. (1987). Do zebra finch males that have been raised by another species still tend to select a conspecific song tutor? Animal Behaviour, 35(5), 1347–1355. https://doi.org/10.1016/S0003-3472(87)80007-6
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
Ferster, C. B., & Skinner, B. F. (1957). Schedules of reinforcement. Appleton-Century-Crofts.
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
Franz, M., & Goller, F. (2002). Respiratory units of motor production and song imitation in the zebra finch. Journal of Neurobiology, 51(2), 129–141. https://doi.org/10.1002/neu.10043
Funabiki, Y., & Funabiki, K. (2008). Song retuning with tutor model by adult zebra finches. Developmental Neurobiology, 68(5), 645–655. https://doi.org/10.1002/dneu.20597
Geberzahn, N., Hultsch, H., & Todt, D. (2002). Latent song type memories are accessible through auditory stimulation in a hand-reared songbird. Animal

Behaviour, 64(5), 783–790. https://doi.org/10.1006/anbe.2002.3099
Glaze, C. M., & Troyer, T. W. (2006). Temporal Structure in Zebra Finch Song: Implications for Motor Coding, 26(3), 991–1005. https://doi.org/10.1523/JNEUROSCI.3387-05.2006
Gobes, S. M. H., Jennings, R. B., & Maeda, R. K. (2017). The sensitive period for auditory-vocal learning in the zebra finch: consequences of limited-model availability and multiple-tutor paradigms on song imitation. Behavioural Processes, 163, 5–12. https://doi.org/10.1016/j.beproc.2017.07.007
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
Halfwerk, W., Varkevisser, J., Simon, R., Mendoza, E., Scharff, C., & Riebel, K. (2019). Toward testing for multimodal perception of mating signals. Frontiers in Ecology and Evolution, 7, 2013–2019. https://doi.org/10.3389/fevo.2019.00124
Hebets, E. A., & Papaj, D. R. (2005). Complex signal function: Developing a framework of testable hypotheses. Behavioral Ecology and Sociobiology, 57(3), 197–214. https://doi.org/10.1007/s00265-004-0865-7
Helekar, S. A., Marsh, S., Viswanath, N. S., & Rosenfield, D. B. (2000). Acoustic pattern variations in the female-directed birdsongs of a colony of laboratory-bred zebra finches. Behavioural Processes, 49(2), 99–110. https://doi.org/10.1016/S0376-6357(00)00081-4
Holveck, M. J., & Riebel, K. (2014). Female zebra finches learn to prefer more than one song and from more than one tutor. Animal Behaviour, 88, 125–135. https://doi.org/10.1016/j.anbehav.2013.11.023
Honarmand, M., Riebel, K., & Naguib, M. (2015). Nutrition and peer group composition in early adolescence: impacts on male song and female preference in zebra finches. Animal Behaviour, 107, 147–158. https://doi.org/10.1016/j.anbehav.2015.06.017
Houx, B. B., & ten Cate, C. (1998). Do contingencies with tutor behaviour influence song learning in zebra finches? Behaviour, 135(5), 599–614.
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
Houx, B., Feuth, E., & ten Cate, C. (2000). Variations in zebra finch song copying: an examination of the relationship with tutor song quality and pupil behaviour. Behaviour, 137, 1377–1389. https://doi.org/10.1163/156853900501980
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
Hyland Bruno, J., & Tchernichovski, O. (2019). Regularities in zebra finch song beyond the repeated motif. Behavioural Processes, 163, 53–59. https://doi.

org/10.1016/j.beproc.2017.11.001
Immelmann, K. (1969). Song development in the zebra finch and other estrildid finches. In R. A. Hinde (Ed.), Bird vocalizations. Cambridge, England: Cambridge University Press,.
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
Jones, A. E., & Slater, P. J. B. (1993). Do young male zebra finches prefer to learn songs that are familiar to females with which they are housed. Animal Behav iour, 46, 616–617. https://doi.org/10.1006/anbe.1993.1233
Jones, A. E., ten Cate, C., & Slater, P. J. B. (1996). Early experience and plasticity of song in adult male zebra finches (*Taeniopygia guttata*). Journal of Compara tive Psychology, 110(4), 354–369. https://doi.org/10.1037/0735-7036.110.4.354
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
Krause, E. T., Bischof, H. J., Engel, K., Golüke, S., Maraci, Ö., Mayer, U., … Caspers, B. A. (2018). Olfaction in the zebra finch (*Taeniopygia guttata*): what is known and further perspectives. Advances in the Study of Behavior, 50, 37–85. https://doi.org/10.1016/bs.asb.2017.11.001
Krebs, J. R., & Kroodsma, D. E. (1980). Repertoires and geographical variation in bird song. Advances in the Study of Behavior, 11, 143–177.
Kroodsma, D. E., & Pickert, R. (1984). Sensitive phases for song learning: Effects of social interaction and individual variation. Animal Behaviour, 32(2), 389–394. https://doi.org/10.1016/S0003-3472(84)80274-2
Kroodsma, D. E., & Verner, J. (1978). Complex singing behaviors among Cistothorus wrens. The Auk, 95(4), 703–716.
Leavell, B. C., Rubin, J. J., McClure, C. J. W., Miner, K. A., Branham, M. A., & Barber, J. R. (2018). Fireflies thwart bat attack with multisensory warnings. Science Advances, 4(8). https://doi.org/10.1126/sciadv.aat6601
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
Ljubičić, I., Hyland Bruno, J., & Tchernichovski, O. (2016). Social influences on song learning. Current Opinion in Behavioral Sciences, 7, 101–107. https://doi.org/10.1016/j.cobeha.2015.12.006
Mann, N. I., & Slater, P. J. B. (1995). Song tutor choice by zebra finches in aviaries. Animal Behaviour, 49(3), 811–820. https://doi.org/10.1016/0003-3472(95)80212-6
Mann, N. I., Slater, P. J. B., Eales, L. A., & Richards, C. (1991). The influence of visual stimuli on song tutor choice in the zebra finch, *Taeniopygia guttata*. Animal

Behaviour, 42(2), 285–293. https://doi.org/10.1016/S0003-3472(05)80560-3
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
Mennill, D. J., Newman, A. E. M., Thomas, I. P., Woodworth, B. K., Norris, D. R., Doucet, M., … Thomas, I. P. (2018). Wild birds learn songs from experimental vocal tutors. Current Biology, 28(20), 3273-3278.e4. https://doi.org/10.1016/j.cub.2018.08.011
Morris, D. (1954). The reproductive behaviour of the zebra finch (*Poephila Guttata*), with special reference to pseudofemale behaviour and displacement activities. Behaviour, 6(1), 271–322. https://doi.org/10.1163/156853954X00130
Morrison, R. G., & Nottebohm, F. (1993). Role of a telencephalic nucleus in the delayed song learning of socially isolated zebra finches. Journal of Neurobiology, 24(8), 1045–1064.
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.
Nelson, D. (1998). External validity and experimental design: the sensitive phase for song learning. Animal Behaviour, 56(2), 487–491. https://doi.org/10.1006/anbe.1998.0805
Ohms, V. R., Snelderwaard, P. C., ten Cate, C., & Beckers, G. J. L. (2010). Vocal tract articulation in zebra finches. PLoS ONE, 5(7). https://doi.org/10.1371/ journal.pone.0011923
Oliveira, R. F., Rosenthal, G. G., Schlupp, I., McGregor, P. K., Cuthill, I. C., Endler, J. A., … 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
Ota, N., Gahr, M., & Soma, M. (2015). Tap dancing birds: The multimodal mutual courtship display of males and females in a socially monoga- mous songbird. Scientific Reports, 5(16614). https://doi.org/doi:10.1038/srep16614
Partan, S., & Marler, P. (1999). Communication goes multimodal. Science, 283, 1272–1274. https://doi.org/0.1126/science.283.5406.1272
Phan, M. L., Pytte, C. L., & Vicario, D. S. (2006). Early auditory experience generates long-lasting memories that may subserve vocal learning in songbirds. Proceedings of the National Academy of Sciences of the United States of America, 103(4), 1088–1093. https://doi.org/10.1073/pnas.0510136103
Price, P. H. (1979). Developmental determinants of structure in zebra finch song. Journal of Comparative and Physiological Psychology, 93(2), 260–277. https://doi.org/10.1037/h0077553
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

Riebel, K., Lachlan, R. F., & Slater, P. J. B. (2015). Learning and cultural transmission in chaffinch song. Advances in the Study of Behavior, 47, 181–227. https://doi.org/10.1016/bs.asb.2015.01.001

Riebel, K., Smallegange, I. M., Terpstra, N. J., & Bolhuis, J. J. (2002). Sexual equality in zebra finch song preference: evidence for a dissociation between song recognition and production learning. Proceedings of the Royal Society of London. Series B: Biological Sciences., 269(1492), 729–733. https://doi.org/10.1098/rspb.2001.1930

Riebel, K. (2000). Early exposure leads to repeatable preferences for male song in female zebra finches. Proceedings of the Royal Society of London. Series B: Biological Sciences., 267(1461), 2553–2558. https://doi.org/10.1098/rspb.2000.1320

Riebel, K, Odom, K. J., Langmore, N. E., & Hall, M. L. (2019). New insights from female bird song: towards an integrated approach to studying male and female communication roles. Biology Letters, 15(4), 1–7. https://doi.org/10.1098/rsbl.2019.0059

Roper, A., & Zann, R. (2006). The onset of song learning and song tutor selection in fledgling zebra finches. Ethology, 112(5), 458–470. https://doi.org/10.1111/j.1439-0310.2005.01169.x

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., & Halfwerk, 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 Be haviour, 18, 1–34. https://doi.org/10.1016/S0065-3454(08)60308-3

Slater, P. J. B., & Richards, C. (1990). Renesting and song learning in the zebra finch, *Taeniopygia guttata*. Animal Behaviour, 40(6), 1191–1192. https://doi.org/10.1016/S0003-3472(05)80190-3

Soha, J. A., & Marler, P. (2000). A species-specific acoustic cue for selective song learning in the white-crowned sparrow. Animal Behaviour, 60(3), 297–306. https://doi.org/10.1006/anbe.2000.1499

Soha, J. A., & Peters, S. (2015). Vocal Learning in Songbirds and Humans: A Retrospective in Honor of Peter Marler. Ethology, 121(10), 933–945. https://doi.org/10.1111/eth.12415

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

Sturdy, C. B., & Nicoladis, E. (2017). How much of language acquisition does operant conditioning explain? Frontiers in Psychology, 8(OCT), 1–5. https://doi.org/10.3389/fpsyg.2017.01918

Sturdy, C. B., Phillmore, L. S., Sartor, J. J., & Weisman, R. G. (2001). Reduced social contact causes auditory perceptual deficits in zebra finches, *Taeniopygia guttata*. Animal Behaviour, 62(6), 1207–1218. https://doi.org/10.1006/anbe.2001.1864

Tanaka, M., Sun, F., Li, Y., & Mooney, R. (2018). A mesocortical dopamine circuit enables the cultural transmission of vocal behaviour. Nature, 563(7729), 117–120. https://doi.org/10.1038/s41586-018-0636-7

Tchernichovski, O., Eisenberg-Edidin, S., & Jarvis, E. (2021). Balanced imitation sustains song culture in zebra finches. Nature Communications, 1–21. https://doi.org/10.1038/s41467-021-22852-3

Tchernichovski, O, Mitra, P. P., Lints, T., & Nottebohm, F. (2001). Dynamics of the vocal imitation process: how a zebra finch learns its song. Sci ence, 291(5513), 2564–2569. https://doi.org/10.1126/science.1058522

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

Tedore, C., & Johnsen, S. (2017). Using RGB displays to portray color realistic imagery to animal eyes. Current Zoology, 63, 27–34. https://doi.org/10.1093/cz/zow076

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

ten Cate, C. (1991). Behaviour-contingent exposure to taped song and zebra finch song learning. Animal Behaviour, 42(5), 857–859. https://doi.org/10.1016/S0003-3472(05)80131-9

ten Cate, C. (1994). Perceptual mechanisms in imprinting and song learning. In J. A. H. & J. J. Bolhuis (Ed.), Causal mechanisms of behavioural development. (pp. 116–146). Cambridge University Press.

Thielcke, G. (1984). Gesangslernen beim Gartenbaumläufer (*Certhia brachydactyla*). Die Vogelwarte, 32, 282–297.

Thorpe, W. H. (1954). The process of song learning in the chaffinch as studied by means of the sound spectrograph. Nature, 173, 465–469.

Todt, D., Hultsch, H., & Heike, D. (1979). Conditions affecting song acquisition in nightingales (*Luscinia megarhynchos L.*). Zeitschrift Für Tierpsychologie, 51(1), 23–35. https://doi.org/10.1111/j.1439-0310.1979.tb00668.x

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

Waser, M. S., & Marler, P. (1977). Song learning in canaries. Journal of Comparative and Physiological Psychology, 91, 1–7.

Whiten, A. (2021). The burgeoning reach of animal culture. Science, 372(6537). https://doi.org/10.1126/SCIENCE.ABE6514

Williams, H. (1990). Models for song learning in the zebra finch: fathers or others? Animal Behaviour, 39(4), 745–757. https://doi.org/10.1016/S0003-3472(05)80386-0

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

Williams, H. (2004). Birdsong and singing behavior. Annals of the New York Academy of Sciences, 1016, 1–30. https://doi.org/10.1196/annals.1298.029

Yanagihara, S., & Yazaki-Sugiyama, Y. (2019). Social interaction with a tutor modulates responsiveness of specific auditory neurons in juvenile zebra finches. Behavioural Processes, 163, 32–36. https://doi.org/10.1016/j.beproc.2018.04.003

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

Chapter 3

Multimodality during live tutoring is relevant for vocal learning in zebra finches

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

This chapter is published in Animal Behaviour:
https://doi.org/10.1016/j.anbehav.2022.03.013

Abstract
In many songbird species, young birds learn their song from adult conspecifics. Like much animal communication, birdsong is multimodal: singing is accompanied by beak and body movements. We hypothesized that these visual cues could enhance vocal learning thus partly explaining the reduced learning from unimodal audio playbacks compared to multimodal live social tutoring observed in many birdsong studies. To test this, juvenile zebra finches, *Taeniopygia guttata*, were tutored in a yoked design where replicate tutoring groups of three male–female dyads were exposed to the same live tutor simultaneously in three different ways. (1) Tutees were housed with the tutor in a central compartment; hence they could hear, see and interact with their tutor ('live'). (2) Tutees placed in one of two adjacent compartments could hear but not see the same tutor from behind a black loudspeaker cloth ('audio-only'). (3) Tutees could likewise hear the tutor through loudspeaker cloth but could also see the tutor through a one-way mirror ('audiovisual'). Comparisons of subadult and adult song showed more changes in the audio-only than in the audiovisual or live tutored tutees, suggesting the audio-only group's song development was delayed. According to (blinded) human observer similarity scoring, the audio-only tutees' singing was least similar and the live tutees' singing most similar to their tutor's singing, while the audiovisual tutees showed an intermediate level of similarity, but the between-treatment differences in similarity were not significant. Conversely, the audio-only group showed the highest similarity values with their father's song, which they only heard before the experimental tutoring. Given that the quantity and quality of the tutor song input were the same across treatments within tutoring groups, the results support the hypothesis that visual in addition to auditory exposure to a tutor can affect the timing and possibly also the amount of vocal learning.

Introduction
Songbirds are well-known vocal learners (Catchpole & Slater, 2003; Doupe & Kuhl, 1999). For the majority of species studied, learning from conspecific social tutors is crucial to develop fully functional species-specific song (Catchpole & Slater, 2003). In some species, hearing adult song from playback provides birds with sufficient input to develop their song, but in a considerable number of the species studied, playing tutor song back via loudspeakers (so-called tape tutoring) resulted in lower tutor song copying accuracy than from a live conspecific as tutor (reviewed in Baptista & Gaunt, 1997; Soma, 2011).

The zebra finch, *Taeniopygia guttata*, one of the commonest animal models for studies on vocal learning (Griffith & Buchanan, 2010; Mello, 2014), is a species

that learns better from a live conspecific than from passive tutor song exposure (reviewed in Derégnaucourt, 2011; Slater, Eales & Clayton, 1988). It is usually concluded that this difference is due to a lack of social interaction with the tutor in the tape–tutor condition (Chen et al., 2016; Derégnaucourt et al., 2013; Eales, 1989), but there are more dimensions that differ between live and tape tutoring (for discussion see Nelson, 1997). For example, live tutoring offers multimodal tutor exposure, as tutees can hear and see their tutor, while tape tutoring only offers auditory, unimodal song exposure. The majority of studies comparing social versus nonsocial tutoring used live and tape tutors and were thus also comparing multi- versus unimodal tutoring, meaning that these issues have been confounded in previous studies (Varkevisser et al., in preparation).

Several lines of evidence suggest that multi- rather than unimodal presentation of song might increase the salience of the stimulus regardless of a social component. Zebra finch song, like birdsong in general and many signals in animal communication (Halfwerk et al., 2019; Higham & Hebets, 2013; Partan & Marler, 1999), is a multimodal signal, as auditory song production is accompanied by visual cues such as beak and body movements (Goller, Mallinckrodt, & Torti, 2004; Ullrich, Norton, & Scharff, 2016; Williams, 2001). Visual stimulation together with an auditory stimulus can facilitate learning of an auditory stimulus as has been demonstrated in domestic chickens, *Gallus g. domesticus,* in the context of filial imprinting (van Kampen & Bolhuis, 1991; van Kampen & Bolhuis, 1993) and in young nightingales, *Luscinia megarhynchos*, that learned songs from audio playbacks combined with light flashes better than those presented as audio-only playbacks (Hultsch et al., 1999). Zebra finches also seem to attend to visual information during song learning. First, tutees show beak and dance movements that are highly similar to the individual-specific movements produced by their tutors, while they differ from those of unfamiliar males (Williams, 2001). Second, the visual appearance of a bird plays a role in tutor choice (Clayton, 1988; Mann et al., 1991; Mann & Slater, 1995). Moreover, improved song learning was found when zebra finch tutees received visual stimulation contingent on their song production (Carouso-Peck & Goldstein, 2019). There are also a number of studies that used static, nonmoving taxidermic mounts as a visual stimulus, but found no improved learning in birds briefly seeing this visual stimulus either right before, during or after song presentation (Bolhuis et al., 1999; Houx & ten Cate, 1999). Overall, it thus seems worth investigating more systematically whether the visual cues associated with song production facilitate song learning.

Our aim in the present study was to create an experimental situation that would allow us to compare tutees receiving either multi- or unimodal exposure to the same live tutor to see whether the additional visual cues, independent of visual interaction, would facilitate song learning in zebra finches. Tutees in our experiment could all hear and vocally interact with the same tutor, since in each set-up a tutor and three male tutees were kept in the same cage which was separated into three compartments (Fig. 1). In the middle compartment, a tutee was housed together with a tutor so that they could interact visually, acoustically and physically. This represents a situation in which normally high levels of song copying occur (Derégnaucourt, 2011). In one of the adjacent compartments, a tutee was housed and separated from the tutor in the middle compartment with acoustically transparent cloth, but the tutee could see the tutor through a small one-way mirror. This provided tutees in this treatment group with multimodal tutor exposure, but it prevented visual tutor–tutee interaction, as the tutees could see the tutor, but the tutor could not see them. In the other side compartment, a tutee was also separated from the tutor in the central compartment with loudspeaker cloth, but without a one-way mirror. This group could hear, but not see the tutor, thus receiving only unimodal audio exposure to the tutor. We provided all male tutees with a juvenile female as a social companion. This was to prevent social isolation, which was a confound in previous studies comparing live tutoring and audio-only song exposure. In addition, the experiment was run simultaneously in two locations with birds from two different breeding colonies using the same paradigm, but with slightly different technical realization on location.

Birds were tutored experimentally between 35 and 65 days posthatching (DPH). The peak of the sensitive phase for song learning in zebra finches is 35–65 DPH, but sensory learning starts as early as 20–25 days DPH (reviewed in Gobes et al., 2017). Zebra finches normally primarily learn the motif of the tutor they can socially interact with between 35 and 65 DPH, but if learning conditions are suboptimal at that time, they might incorporate syllables heard before that age in their song (Böhner, 1986; Eales, 1989; Jones, ten Cate, & Slater, 1996; reviewed in Gobes et al., 2017). As our tutoring methods had not been tried before, we assessed the similarity of tutee's songs with both their father (the first encountered model) and the tutors they encountered during the experimental tutoring phase (35–65 DPH). Song development entails not only learning to sing specific syllables but also the ordering and timing of syllables, as well as the stereotypy of song delivery, all aspects of song that can differ substantially between individual male zebra finches (Helekar et al., 2000; Holveck et al., 2008; Hyland Bruno & Tchernichovski, 2019; Scharff & Nottebohm,

1991). We thus assessed a number of these parameters in addition to song similarity scores. If visual cues play a role in song learning, then song learning in the birds with multimodal tutor exposure should be better than in the birds with unimodal tutor exposure. However, if visual cues do not play a role, and live tutors facilitate song learning mainly because of their 'sociality', then equal song copying is expected in the birds with audiovisual and audio-only tutor exposure.

Methods

Subjects and housing

Subjects for this study were 13 adult male tutors and 44 male and 90 female juvenile domesticated wild-type zebra finches from two different breeding colonies. One colony was located at Leiden University (contribution to experiment: N = 9 adult male tutors and 27 male and 54 female juveniles) and the other at the Free University Berlin (contribution to experiment: N = 4 adult male tutors and 18 male and 36 female juveniles). In Leiden, subjects were bred in several rounds: 12 male and 24 female juveniles (four tutor groups, see below) hatched in March 2017, three male and six female juveniles (one tutor group) hatched in August 2017 and 12 male and 24 female juveniles (four tutor groups) hatched in November 2017. In Berlin, breeding in the colony was continuous and juveniles for the first tutor group hatched in January 2017 and for the last tutor group in November 2018. All young birds were the offspring of established breeding pairs and were housed in breeding cages (Leiden: 80 x 40 cm and 41 cm high; Berlin: 180 x 42 cm and 33 cm high) until chicks were 35 DPH (age of chicks was determined as the median hatching day of all chicks within the nest). In Berlin, the father remained in the same cage but was separated from the juveniles at 23 DPH by a wire mesh covered in paper allowing vocal communication but not visual or physical contact. In Leiden, the father remained in the breeding cage with the tutees until 35 DPH, at which age (mean ± SD = 35.3 ± 1.2 days) young males and females were assigned to tutor groups that were exposed to the song of the same unrelated (coefficient of relation < 0.125) adult male (the 'tutor'). The adult tutors had been housed in same-sex aviaries prior to the experiment (Leiden, age range at the start of the experiment 120–806 days, mean ± SD: 509 ± 300; Berlin, range 1919–2945 days, mean ± SD: 2482 ± 424). Next to one adult tutor, each tutor group consisted of three male and three female tutees for the tutoring period and an additional three female companions that were cohoused with the tutees after the tutoring phase. For each tutoring group, whenever possible, we chose three males from the same nest. This was possible in two tutor groups in Leiden and all six tutor groups in Berlin (in the other groups in Leiden, two siblings were

spread over the different treatment conditions across tutor groups). Each tutor group was then complemented with three female tutees always chosen from a different brood as females develop a preference for the song they hear early in life, and might guide male song development to a certain degree based on this preference (Jones & Slater, 1993). Therefore, whenever possible, these three females originated from the same brood and had heard the same tutor song early in life (N = 8 tutor groups). In the other groups, two siblings were spread over the different treatment conditions across tutoring groups. Tutors, together with one male and female tutee (live condition), were placed in the middle compartment of an experimental cage consisting of three compartments (Fig. 1; Leiden: 150 x 40 cm and 50 cm high, located in a larger bird room where other birds were audible, but not visible; Berlin: 90 x 33 cm and 42 cm high, located in a sound-attenuated chamber 91 x 150 cm and 78 cm high).

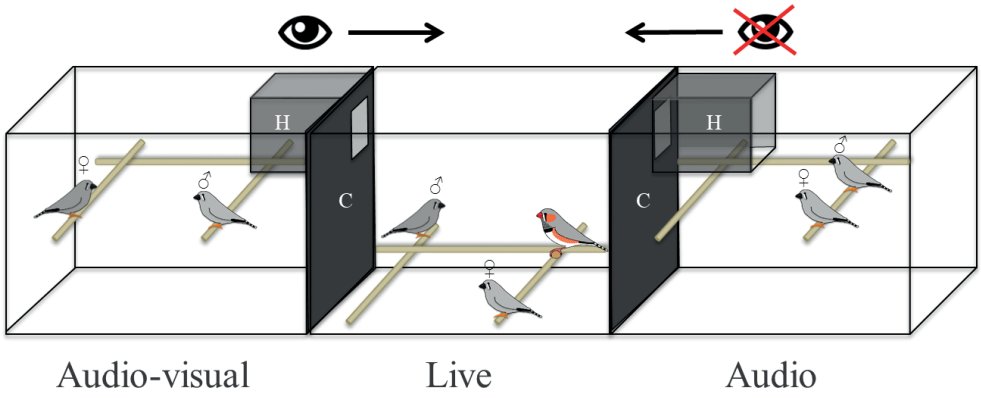

Figure 1. Schematic front view of the experimental set-up in which the song tutoring took place. C = separation made from loudspeaker cloth, H = observation hut. Eye and crossed-out eye symbol show through which one-way mirror the tutees could and could not see the tutor, respectively.

A male/female tutee dyad was also placed into each of the compartments to the left and right. The compartments were separated from each other by opaque (black), but acoustically transparent loudspeaker cloth. One of the two side compartments was assigned to the audio-only condition: tutees could only hear (though the loudspeaker cloth) but not see the tutor in the central compartment. The other compartment was designated to the audiovisual treatment: as in the audio-only treatment the tutor could be heard via the cloth, but in addition, the tutee could watch the tutor via a small one-way mirror (5x8 cm), when perched on the upper central perch. The audio-only compartment had an identical mirror, but it was rendered opaque (by gluing together two one-way mirrors with a piece of white paper in between). The assignment of audio-only or audiovisual compartment to the left or right compartment was balanced

across set-ups. As the one-way mirrors only function if there is a difference in light intensity between the two sides of the mirror, we built 'observation huts', consisting of black painted wood and opaque loudspeaker cloth (Leiden) or black plastic (Berlin), around the one-way mirrors, such that a bird perching in front of the mirror would find itself in a dimly illuminated space. Perches were arranged in such a way that birds in the side compartments could easily reach the one-way mirrors, while the perches in the middle cage were positioned lower so that it was more difficult to reach the mirrors. In Leiden, we noticed that some of the tutors were nevertheless flying up to the one-way mirrors and clinging to the cloth next to it. To avoid the birds in the middle compartments coming close to the one-way mirrors and seeing their own appearance, we glued transparent plastic hemispheres (10 cm diameter) around the one-way mirrors. We characterized sound propagation through the cages and the loudspeaker cloth as follows. A loudspeaker playing pure tones between 200 Hz and 18 kHz in 200 Hz steps was placed at different positions in the central compartment. We had a microphone in the same compartment and one microphone at different positions in the neighbouring side compartment. We found that at the positions where we made the recordings in the side compartment, the frequency response was similar to that in the central compartment but was attenuated (probably due to spreading loss and atmospheric attenuation) between 3dB for lower frequencies and up to 10 dB at higher frequencies (see Fig. A1). A measurement with a microphone installed in the 'observation hut' showed that the one-way mirror, the hut and the position of the hut probably influenced the frequency response. There was almost no effect for frequencies below 4 kHz, but between 6 kHz and 10 kHz a higher attenuation compared to the other location was measured, indicating that the song of a tutor gets somewhat filtered when tutees listen in the observation hut. Note, however, that the audio-only conditions were the same in the two side cages and that the audio-only and audiovisual tutees thus had the same audio conditions.

When tutees reached 65 DPH, tutors were removed from the experimental set-up. Female tutees were also moved to form all-female groups in aviaries as their development was followed separately. The females were moved to prevent them from learning from the male tutees that start singing adult-like song around this time (Immelmann, 1969). All male tutees that had remained in the set-up received a new female companion each; these were of the same age and from the same breeding round. These females had been housed in sets of three with an adult tutor pair between 35 and 65 DPH to be then moved into the experimental set-up to replace the female tutees. For every tutor group, the three new female companions had been housed with the same adult tutor pair

before 65 DPH. In addition to this, in Leiden, all the mirrors were now covered with cardboard. Male tutees remained in the experimental cages (in Berlin and Leiden) except for a brief recording session at 65 days and then until their song was recorded after 100 DPH. Then, males were moved into large all-male group aviaries. See Fig. 2 for a timeline of the experimental procedure.

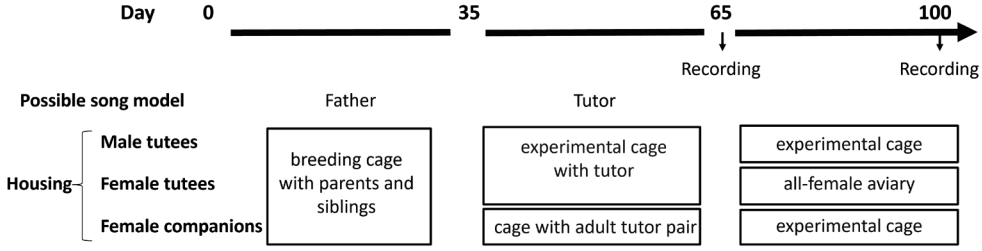

Figure 2. Time line of the experimental procedure. Note that in Berlin, the father was removed from the breeding cage at around 23 days posthatching.

Throughout, birds were housed on a 13.5/10.5 h (Leiden) or 12/12 h (Berlin) light/dark cycle, at 20–25 ºC and 45–65% humidity. Within tutor groups, the tutees from the different treatments were housed in the same bird room (Leiden) or in the same soundproof box (Berlin); thus, within groups tutees always experienced the same temperature and humidity conditions. Birds had ad libitum access to a commercial tropical seed mixture (Leiden: Beyers, Belgium; Berlin: Teurlings, Germany), cuttlebone, grit and drinking water. This diet was supplemented twice a week with hardboiled eggs, germinated tropical seeds, vegetables and fruit.

In total, 15 tutor groups were raised in the experimental set-ups: nine in Leiden and six in Berlin. One tutor group (Berlin) consisted only of three male tutees and one tutor, because there were no same-age females available at the start of song tutoring.

Usage of observation huts by tutees

To investigate whether the tutees in the side compartments (audiovisual and audio-only treatments) were using the observation huts, for the first four tutor groups raised in Leiden, we filmed the tutees in the two treatment groups (GoPro Hero 3+ camera, San Mateo, CA, U.S.A.) two mornings a week throughout the tutoring period. For each group, we analysed 2–4 h of video recorded during one or two mornings (recordings started between 0900 and 1000 hours) in the second week of tutoring, because we expected the birds to be familiar

with the experimental set-up by then. For these videos, every 30 s the positions (inside or outside the hut) of the male and female tutee were scored separately. The proportion of observations during which the male and female tutee of the audiovisual and audio-only treatments were inside the observation huts was then calculated. Zebra finches have a wide visual field (each eye around 170° in the horizontal plane; Bischof, 1988) and can look through the window while their body or head is not directed towards it. From our video recordings, we could therefore only assess whether the birds were in the hut, but not when the tutees were looking through the one-way mirror. However, the proportion of observations where the tutees were inside the huts does give an indication of the total time for which the tutees could have watched the tutor.

Song recording

Both in Berlin and in Leiden, for all song recordings (fathers, tutors and male tutees), birds were moved the afternoon before a recording day to acclimate in a sound-attenuated chamber (Leiden: 125 x 300 cm and 240 cm high; Berlin: 60 x 60 cm and 73 cm high) that contained a small recording cage (Leiden: 76 x 45 cm and 45 cm high; Berlin: 46 x 29 cm and 48 cm high), and then recorded continuously the next morning with a microphone suspended from above the cage (Leiden: Sennheiser MKH40 microphone, Wedemark, Germany connected to a TASCAM DR-100MKiii recorder (TEAC Corp., Los Angeles, CA, U.S.A.), sampling at 96 kHz, 16 bits; Berlin: Earthworks SRO microphone (Milford, NH, U.S.A.) connected to a PC using the recording function of the Sound Analysis Pro software (SAP; Tchernichovski, Nottebohm, Ho, Pesaran, & Mitra, 2000), sampling at 44.1 kHz, 16 bits). The tutees' fathers were recorded after successful breeding and after their offspring (the tutees) had been moved into the experimental set-up. All song tutors had been recorded prior to moving them into the experimental set-ups. All tutees were recorded twice: once at 65 DPH (mean ± SE: 66 ± 1.4 days) and once as young adults after 100 DPH (mean ± SE: 130 ± 11.3 days). The recording at 65 DPH took place while tutee song was still developing, but when most syllables of the final song are usually present (Slater, Eales, & Clayton, 1988). At 65 DPH, in Leiden, male and female tutees were placed together in the recording cage (as in Leiden, many birds had to be recorded around the same time, and males are more likely to sing the first day if housed with a female). In Berlin, young male tutees were recorded without their female companions. At >100 DPH, in both Leiden and Berlin male tutees were recorded while they were temporarily housed individually in the recording cage. After tutoring had started, it turned out that one of the audiovisual groups in Berlin had two females instead of a male and female tutee due to misidentification. From the remaining 44 male tutees, 33 birds could be re-

corded at 65 DPH. At >100 DPH, 41 birds produced more than 20 songs. Only song of these birds was used in the song analysis. The father from one of the tutor groups could not be recorded, so the song of the tutees from this group was only compared to the tutor song.

Song analysis

Comparison of the relative success of tutoring methods has been hampered by the many different analysis methods used in zebra finch song research. Studies up until 1999, including many relevant for this study, mainly used visual inspection of spectrograms by human observers (e.g. Bolhuis et al., 1999; Eales, 1989; Houx & ten Cate, 1999) to assess the similarity between the tutees' song and possible model songs. Zebra finch song studies since 2000 have regularly used automated digital measurement methods, such as SAP (Tchernichovski, Nottebohm, Ho, Pesaran, & Mitra, 2000) and Luscinia (Lachlan et al., 2010). To make our results comparable to these previous studies and to our recent study employing a video tutoring method (Varkevisser et al., 2021), we used all three of these methods (human observers, SAP and Luscinia) in this study. To aid cross-study comparison and interpretation, we also assessed the correlation between the three methods for our complete data set and for the subset of live tutored tutees separately, as previous method comparisons have only taken into account similarity assessment for the song of tutors and their live tutored tutees (Lachlan et al., 2010; Tchernichovski et al., 2000), a condition known to lead to high tutor song similarity (Derégnaucourt, 2011). Next to tutor–tutee similarity, an additional set of structural dimensions (see Table 1) were analysed, again using the same parameters as Varkevisser et al. (2021).

Song and motif selection

Following Sossinka and Boehner (1980), we defined a song syllable as a unit of sound separated from another sound by a silent interval of at least 5 ms and a motif as an individual-specific sequence of syllables. The term 'song' refers to a series of motifs separated from other sounds by more than 2 s of silence or a series of motifs preceded by multiple introductory notes (Sossinka & Böhner, 1980). Selection of songs and sound editing were conducted by visual inspection of combined spectrograms and amplitude waveform displays with Praat software (v. 6.0.19, Boersma & Weenink, 2008; spectrogram settings: fast Fourier transformations with 1000 time and 250 frequency steps, 0.005 s window length, dynamic range 55 dB, Gaussian window). The spectrograms of all audio-recording sessions were visually screened and digitally parsed into songs to be saved as separate audio files using one folder per recording session (65 DPH and >100 DPH) of each male. From each folder, 20 songs were picked at ran-

dom (with custom-written software by Niklas J. Tralles) and from these songs, the motif encountered most often was selected and termed 'the typical motif'. In addition, we selected the motif with the highest number of different syllables (the 'full motif') from the adult (>100 DPH) recordings. We also selected a random subset of 10 motifs by first randomly selecting 10 of the 20 songs and then selecting one motif from each of the 10 songs with a random number generator (http://www.random.org). These motifs were digitally cut from the recordings, band stop filtered from 0 to 420 Hz, and amplitude normalized with the 'scale peak' function (all with Praat Software, v. 6.0.19). Introductory notes that were part of each motif occurrence were kept, but all additional introductory notes were cut off before further analysing these 10 motifs with the SAP and Luscinia software (see below).

Song structure and performance

For the typical and full motifs, one of the authors (J.V.) visually inspected the spectrograms and labelled all different syllables with different letters (see Fig. 3, using the Praat software and settings as described above). For each tutee, we counted the syllables in the typical motif as well as the number of unique syllables in the full motif. We then calculated sequence linearity and consistency (Scharff & Nottebohm, 1991), by assessing the different transitions between the syllables for the sample of 20 randomly selected undirected crystallized songs of each male. Sequence linearity is the total number of different syllables divided by the number of different transitions between syllables and higher scores indicate a more stereotyped syllable order of the different motifs within a song. We determined sequence consistency by first noting all transitions and then determining the most frequent ('typical') transition for each syllable in the 20 songs. We divided the total number of occurrences of typical transitions by the total number of transitions encountered in the 20 randomly selected songs. As with sequence linearity, higher scores indicate more stereotyped songs. As an additional song stereotypy measure, we conducted within-subject comparisons by comparing each of the 10 randomly selected motifs with each other in SAP and Luscinia, using the same settings as for the similarity scores (see below). We continued analyses with the median SAP similarity score and the median 1-d Luscinia distance value, so that for both scores higher scores indicate a higher stereotypy. These values are referred to as the 'SAP stereotypy score' and 'Luscinia stereotypy score'.

Similarity to tutors' and fathers' song

Human observer similarity scoring

We followed the procedures from Houx and ten Cate (1999a) for the human

observer similarity scoring, but with the difference that we chose syllables (see above and Fig. 3) instead of elements as units. We opted for syllables, because based on the literature we expected poor tutor song copying and isolate-like song in the experimental groups (Eales, 1989; Price, 1979) which can make determining element boundaries problematic because of the variance in the frequency patterns being higher than in normal song (Price, 1979). Identifying syllable boundaries is less of a problem, as syllables can be recognized by the short silent intervals that delineate them. Three observers (Ph.D. candidates from the Leiden lab), blinded with respect to birds' IDs and treatments (and with some but varying experience with spectrogram analyses), independently assessed syllable similarity between the song models and the tutees. Each observer scored the complete set of spectrograms (see Fig. 3 and Fig. A2) while working through a PowerPoint presentation on a personal computer. Each new slide presented a new set of spectrograms: one of a tutee's full motif (labelled 'tutee') on top and directly underneath a second spectrogram labelled 'model' (which unknown to the observer was from either the tutor or the father of the tutee). The observers had to compare each tutee with two models: the tutor and the father. They were asked to indicate for each tutee syllable the most similar syllable of the model by paying attention to a syllable's frequency pattern, duration, overall shape and sequential position and to score the degree of similarity on a scale from 0 to 3 (0 = 'no similarity at all', 1 = 'slight similarity', 2 = 'moderate similarity' and 3 = 'very strong similarity'). Interobserver reliability was calculated after normalizing individual observer scores by subtracting the mean of the observer's scores from each score and then dividing it by the standard deviation. Using these normalized similarity values as the response variable and tutee ID as a factor, we then conducted a one-way ANOVA to calculate repeatability (Lessells & Boag, 1987) which was high for all comparisons (Tutor–Tutee: $F_{2,38}$= 12.92, P < 0.01, r ± SE = 0.80 ± 0.05; Tutee–Tutor: $F_{2,38}$ = 10.18, P < 0.01, r ± SE = 0.75 ± 0.06; Father–Tutee: $F_{2,38}$= 7.07, P < 0.01, r ± SE = 0.67 ± 0.07; Tutee–Father: $F_{2,38}$ = 5.17, P < 0.01, r ± SE = 0.58 ± 0.08). While this indicates relatively high agreement, it also shows that observers differed; this was mainly because observers varied in how strict they were regarding the syllables with lower similarity to the tutor syllables. To capture this best, for further analysis, we decided to combine the individual scores of all observers by first summing them and then dividing them by the maximum score a bird could have received from three observers. This resulted in one similarity score for a particular model–tutee comparison, while correcting for between-individual differences in the number of syllables in the motif, thus providing a measure that combines the proportion of syllables copied with a weighting of their similarity.

Tutees can differ in the proportion of copied versus improvised syllables (e.g. Tchernichovski et al., 2021; Williams, 1990), which means that the direction of comparison can affect the scores for syllable sharing. For example, if a tutee has accurately copied the syllables ABC from a tutor with the motif ABCDE, this tutee would score higher on the tutee–model comparison (ABC = 100% of the tutee's syllables are shared with the tutor) than on the model–tutee comparison (ABC = only 60% of the tutee's syllables were copied from ABCDE). Conversely, another tutee singing motif ABCDEFG, where ABCDE are copied and F and G improvised, would score higher on the model–tutee comparison (all tutor syllables, i.e. 100%, copied) than on the tutee–model comparison (only ABCDE, but not F and G are shared, thus this yields only 71%). As the types and direction of effects (i.e. poor copying or improvisation) for this new type of tutoring could not be predicted from the literature, we assessed song similarities between tutors and tutees to capture both overlap and level of improvisation by looking at (1) the proportion and similarity of the model's syllables that the tutee has copied ('similarity score model–tutee') and (2) the proportion and similarity of the tutee's syllables that are shared with the model ('similarity score tutee–model'). For the similarity score model–tutee, for each model syllable, we noted the ID and similarity score of the tutee syllable that received the highest score and then summed these scores. If two or more tutee syllables received the same score, we noted this score once, but for the similarity score tutee–model, the scores for all tutee syllables were included (see Table 1 for full formula).

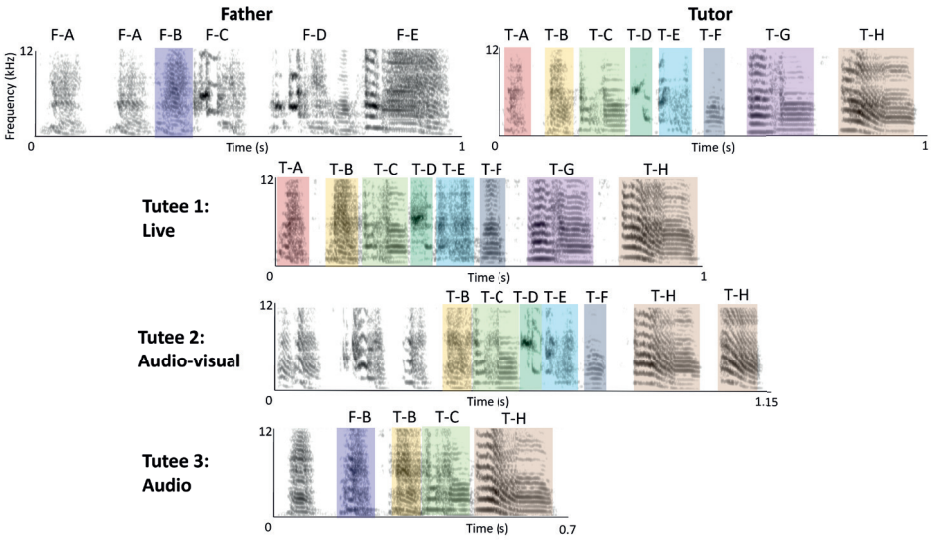

Figure 3. Spectrograms of the songs of the father, tutor and three male tutees from one tutor

group (full motif). Syllables are labelled by two letters indicating the song model (F, T) combined with a second letter indicating the syllable identity. Human observers scored the similarity between two syllables on a scale from 0 to 3. Syllables shaded in the same colour were judged as the most similar syllable by at least two observers. Note, however, that this binary categorization of shared/nonshared syllables is for illustration purposes; it does not reflect the continuous scoring of similarity in the analyses which combined the scores of all three observers and corrected the score for total motif length (see parameters 'human observer similarity score model-tutee' and 'human observer similarity score tutee-model' in Table 1).

Automated similarity scoring (SAP and Luscinia)

For the automatic, quantitative song comparisons, we used Luscinia (version 2.16.10.29.01) and SAP (MxN comparison, default settings tuned for zebra finch, per tutor–tutee pair amplitude thresholds were adjusted for correct syllable segmentation, version 2011.104) to compare each of the 10 randomly selected tutee motifs to each of the 10 randomly selected father's and tutor's motifs. For each possible comparison, we assessed the asymmetric time courses SAP similarity score for the model to tutee and tutee to model comparisons (SAP similarity score model–tutee and tutee–model). Both values indicate the percentage of sounds of one song (tutee or model) observed in the other. As the quantitative measure of similarity, for each individual we used the median value of the scores resulting from the comparisons between the 10 randomly selected motifs per individual. We used the median, because our sample size was too small to create a good-fitting model for all similarity scores and because the SAP scores did not follow a normal distribution and were bound between 0 and 100. For the acoustic distance calculations between model–tutee pairs in Luscinia, which uses a dynamic time warping (DTW) algorithm, we selected the acoustic features 'mean frequency', 'fundamental frequency' and 'fundamental frequency change' (following Lachlan, van Heijningen, ter Haar, & ten Cate, 2016) and we added the feature 'time', which allows for flexible comparison of motifs of different duration. The DTW analysis results in one distance measure (d) between 0 and 1 for each possible motif pair. In contrast to the human observer and SAP similarity scores, this measure is symmetric, meaning that it is the same for the model to tutee and the tutee to model comparisons. In Luscinia, a smaller distance value means a higher similarity, but because the other two methods express higher similarities with higher values, we simply calculated the inverse of the median distance score (1 - d, henceforth 'Luscinia similarity score'), to aid comparison across all three methods.

Structural changes in the typical motif between 65 and 100 DPH

For each tutee, the syllables of the typical motif produced at 65 DPH were compared with those at the second recording at > 100 DPH by visually inspect-

ing the spectrograms to assess the number of changes (i.e. syllable deletions, repetitions or insertions). For this analysis, spectrograms were saved under a code number and then inspected by one of the authors (J.V.) without knowing a tutee's treatment group or whether syllables were improvised or copied from model song.

Table 1. All song analysis parameters and the formulas and sample sizes for calculation.

Parameter	Definition	Sample per bird
Typical motif	Most frequently produced motif	20 random songs
Full motif	Motif with highest # different syllables in bird's repertoire	20 random songs
Total number of syllables	# syllables in a tutee's typical motif	Typical motif
Number of unique syllables	# unique syllables in a tutee's full motif	Full motif
Linearity	(# different syllables/song)/(#transition types/song)	20 random songs
Consistency	(total # typical transitions)/(total # of transitions)	20 random songs
Human observer similarity score model-tutee	(Σ similarity scores (all observers) for all model syllables)/(# model syllables*3 (max score)*# observers)	Full motif
Human observer similarity score tutee-model	(Σ similarity scores (all observers) for all tutee syllables)/(# tutee syllables*3 (max score)*# observers)	Full motif
SAP similarity score model-tutee	Median SAP similarity scores comparing tutor's/father's to tutee's motifs	10 random motifs
SAP similarity score tutee-model	Median SAP similarity scores comparing tutee's to tutor's/father's motifs	10 random motifs
Luscinia similarity score	Median 1 – Luscinia distance score for tutor/father motifs compared to tutee motifs	10 random motifs
SAP stereotypy score	Median SAP similarity scores within-tutee motif comparisons	10 random motifs

Luscinia stereotypy score	Median 1 – Luscinia distance scores within-tutee motif comparisons	10 random motifs
Changes 65 to > 100 dph	# changes in motif produced at 65 and >100 dph	Typical motif (65 and 100 dph)

dph: days posthatching. All samples analysed were from the 100 dph recordings, except the sample used to calculate Changes 65 to > 100 dph. For that parameter, the typical motifs recorded at 65 and >100 dph were analysed. The parameters and definitions listed here are identical to those used by Varkevisser et al. (2021).

Statistical analysis

We used RStudio (R version 3.5.1, http://www.rstudio.com/) to build linear mixed-effects models (LMMs) to compare whether treatment groups differed in number of unique syllables, the sequence linearity and sequence consistency scores, and the human observer, SAP and Luscinia scores. Human observer, SAP and Luscinia scores are bounded distributions and were therefore arcsine square-root transformed prior to analyses to meet model assumptions. Generalized linear mixed-effect models (GLMMs) with a Poisson distribution and log-link function were used to assess whether tutees from different treatments differed in the total number of syllables and the number of changes from 65 to > 100 DPH (package lme4: Bates, Mächler, Bolker, & Walker, 2014). For every parameter, we first ran a null model including 'Tutor group' (Number of the tutor group, 15 tutor groups in total) as a random intercept and 'Location' (Leiden or Berlin) as a fixed factor. We always included 'Location' as the locations differed in the technical realization of the experiment (see above). We used ANOVAs to compare this null model to a model that included 'Treatment' (Live, Audiovisual or Audio-only) as a fixed factor. We used a Shapiro–Wilk test to assess whether the model's residuals followed a normal distribution. Post hoc tests with Tukey adjustment for multiple comparisons were conducted for between-treatment comparisons if the model with 'treatment' was significantly better than the model without 'treatment' as a fixed factor (package emmeans: Lenth, Singmann, Love, Buerkner, Herve, 2018). The three similarity scores (human observers, SAP and Luscinia) for all tutees and the live tutored tutees only were compared with Pearson correlation coefficients after the human observer scores were square-root transformed to meet normality assumptions.

Ethical note

We adhered to the ASAB/ABS Guidelines for the Use of Animals in Research and the European and Dutch legislation on animal experimentation. At all stages in the experiment birds had ad libitum access to food and water and were cohoused with at least one other bird (apart from the short song record-

ing sessions). The manipulation of the size and composition of social groups (in accordance with general housing procedures) is not considered a procedure in the Experiments on Animals Act (Wet op de Dierproeven, 2014) which is the applicable legislation in the Netherlands in accordance with the European guidelines (EU directive no. 2010/63/EU) regarding the protection of animals used for scientific purposes. We also had no indication that the described procedures induced distress or impaired welfare. At all times, all birds were housed and cared for in accordance with these regulations and internal guidelines concerning care of the animals and licensing and skill of personnel, including review and monitoring by the Leiden University Animal Welfare Body and following their advice to ensure the wellbeing of all animals at the facility (with or without a licence).

Results

Usage of observation huts by tutees

To assess whether tutees came near the one-way mirrors, tutee position (inside or outside the observation hut) was scored for the audio-only and audiovisual tutees of four tutor groups (N = 16: four male and four female audio-only and audiovisual tutees). All tutees used the perches in the observation hut more often than expected by chance. Although only 14.9% of the total perch area was inside the observation hut, the average percentage of observations (mean ± SD) inside the observation huts was almost double of what was expected: in the audiovisual group males spent 28 ± 19% (N = 4 birds, 15 observation-hours) and females 26 ± 18% (N = 4 birds, 28.5 observation-hours) of observations in the huts. Interestingly, for the audio-only tutees, the percentage of observations where the tutees were inside the hut was similar to that for the audiovisual groups for both males (26 ± 18%, N = 4 birds, 32 observation-hours) and females (25 ± 17%, N = 4 birds, 32 observation-hours) suggesting that the huts were a preferred area for all birds, regardless of whether it allowed them to see the tutor.

Song structure and performance

The parameters used to assess song structure and performance (total number of syllables, number of unique syllables, linearity and consistency) did not vary significantly between tutoring treatments (models including 'treatment' as fixed factor were not significantly better than the models without 'treatment', see Table 2 and the details of the models with treatment in Table 3). To test whether the tutees from the different treatments differed in between-motif stereotypy, we compared the 10 randomly selected tutee motifs to each other in SAP and Luscinia. There was no significant effect of treatment on the SAP or Luscinia

stereotypy scores (adding 'treatment' as fixed factor did not significantly improve the null model: SAP stereotypy score: N = 41, χ_2 = 2.18, P = 0.34; Fig. 4a, Table 4; Luscinia stereotypy score: N = 41, χ_2 = 0.15, P = 0.93; Fig. 4b, Table 4).

Table 2. Mean values for the song structure and performance parameters per treatment group and details on ANOVA comparing the null model with a model including 'treatment' as a fixed effect (both models included 'tutor group' as random factor and 'location' as fixed factor). We did not include the tutor data in the models.

	Tutor (not in models)	**Live**	**Audio-visual**	**Audio**	**ANOVA null model and model with 'treatment'**		
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	N	χ^2	p
Total nr syllables	*7.1 ± 2.0*	7.2 ± 1.5	7.3 ± 3.4	6.4 ± 2.3	41	0.88	0.64
Nr unique syllables	*6.2 ± 1.1*	5.5 ± 1.7	6.1 ± 1.7	5.9 ± 1.7	41	0.87	0.65
Linearity	*0.55 ± 0.12*	0.43 ± 0.09	0.47 ± 0.12	0.47 ± 0.08	41	1.19	0.55
Consistency	*0.93 ± 0.06*	0.92 ± 0.05	0.92 ± 0.06	0.91 ± 0.06	41	0.17	0.92

Table 3. Details of models with treatment as fixed factor for the song structure and performance parameters

Response variable	***Model term***	***Level***	***Estimate***	***SE***	***z/t***
Total no. of syllables[1]	Intercept		1.92	0.12	15.55
	Treatment				
		Audio-visual	0.12	0.15	0.81
		Live	0.12	0.14	0.82
	Location				
		Leiden	-0.11	0.12	-0.96
No. of unique syllables[1]	Intercept		6.62	0.53	12.39
	Treatment				
		Audio-visual	0.15	0.61	0.24
		Live	-0.36	0.57	-0.63
	Location				
		Leiden	-1.21	0.55	-2.21

Linearity[2]	Intercept		0.46	0.03	14.32
	Treatment				
		Audio-visual	0.004	0.04	0.12
		Live	-0.03	0.04	-0.86
	Location				
		Leiden	0.003	0.03	0.09
Consistency[2]	Intercept		0.89	0.02	50.95
	Treatment				
		Audio-visual	0.01	0.02	0.36
		Live	0.01	0.02	0.28
	Location				
		Leiden	0.04	0.02	2.46

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

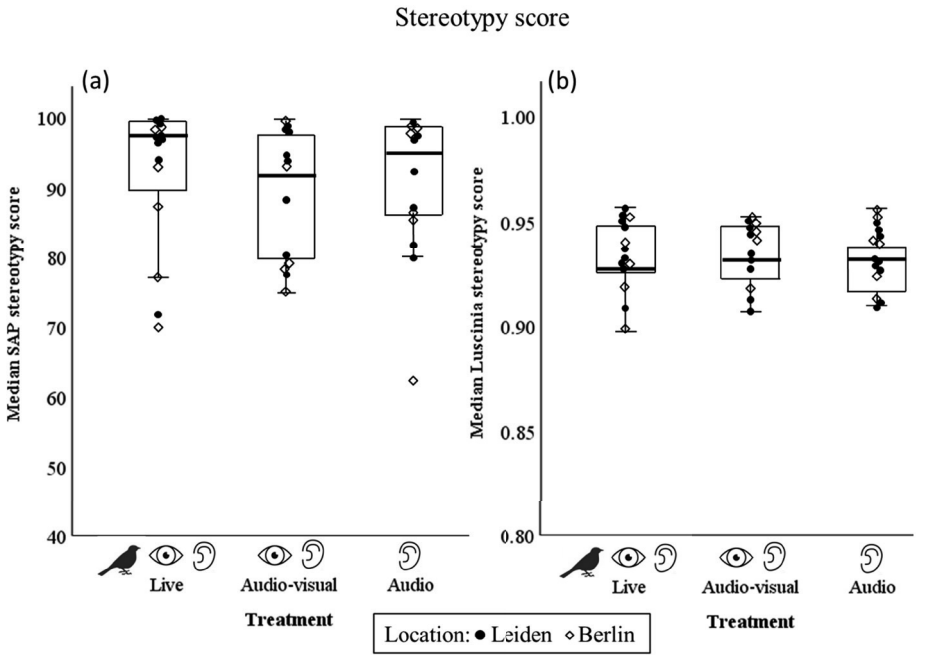

Figure 4. (a) SAP and (b) Luscinia stereotypy scores for the 10 randomly selected tutee motifs recorded at 100 days posthatching in the three treatments. Box plots indicate the median (midline), interquartile range (box) and 1.5 times the interquartile range (whiskers).

Table 4. Details of models with 'treatment' as fixed factor for the (arcsine square-root transformed) SAP and Luscinia stereotypy scores

Response variable[1]	Model term	Level	Estimate	SE	t
SAP stereotypy score	Intercept		1.24	0.06	21.87
	Treatment				
		Audio-visual	-0.05	0.07	-0.77
		Live	0.04	0.06	0.66
	Location				
		Leiden	0.09	0.06	1.61
Luscinia stereotypy score	Intercept		0.93	0.006	165.83
	Treatment				
		Audio-visual	0.002	0.005	0.33
		Live	0.002	0.005	0.33
	Location				
		Leiden	-0.002	0.006	-0.32

[1] LMM with 'Tutor group' as random factor.

Similarity to tutors' and fathers' song

Comparison of different similarity assessment methods

For the similarity scores of all tutees, we found a significant correlation between the human observer and the Luscinia similarity scores and between the human observer and the SAP similarity scores for the father–tutee comparison. There was no correlation between the Luscinia and SAP similarity scores (see Table 5). When looking at the similarity scores of the live tutored tutees only (to enable comparison with earlier studies comparing Luscinia or SAP with human observer similarity scores for live tutored tutees), we only found a significant correlation between the Luscinia and the human observer similarity scores (see Table 5). To find out whether the low correlation between the SAP and the human observer similarity scores was related to the different song samples used to calculate these (one typical motif for the human observer scores and 10 randomly selected motifs per tutee for the SAP scores), we repeated the SAP similarity score calculations with the same sample that was used for the human observer similarity scores (one typical motif for each tutee compared to one typical motif of each tutor, the same motifs that were used for the human observer similarity scores). This led to a significant correlation between the

SAP and human observer scores for the tutor–tutee comparison, but not for the father–tutee comparison (see Table 5). None of the correlation coefficients were very high (all below 0.73), suggesting that the three methods measured different aspects of song similarity. Below, we present the data from all three similarity assessment methods.

Table 5. Pearson correlation coefficients for the human observer similarity scores (square-root transformed), the median SAP and the median Luscinia similarity scores

	Tutor-Tutee		Father-Tutee	
Comparison[1]	r	p	r	p
All tutees				
Human – SAP	0.15	0.37	**0.32**	**0.04**
Human – Luscinia	**0.73**	**<0.01**	**0.47**	**<0.01**
SAP - Luscinia	0.10	0.53	0.04	0.82
Live tutored tutees only				
Human – SAP	0.41	0.15	0.21	0.47
Human – Luscinia	**0.69**	**<0.01**	**0.59**	**0.03**
SAP - Luscinia	0.04	0.89	-0.11	0.72
All tutees: one motif of each tutee				
Human – SAP	**0.40**	**<0.01**	0.26	0.12

Sample sizes: all tutees: tutor–tutee: N = 41; father–tutee: N = 39; live-tutored tutees only: tutor–tutee: N = 14; father–tutee: N = 13. Significant values are given in bold, for all tutees, for the live tutees only and for the SAP and human observer similarity scores calculated for the same sample of one typical motif of each tutee compared to one typical motif of the tutor.

[1] Human = human observers similarity score, SAP = SAP similarity score, Luscinia = Luscinia similarity score.

Similarity between tutees' and their tutors' songs

For the human observer similarity scores calculated by comparing the tutor's syllables to the tutee's syllables (tutor–tutee comparison), adding 'treatment' as fixed factor to the null model did not lead to a significant improvement (N = 41, $\chi_2 = 2.78$, P = 0.25). Similarity was highest for the tutees in the live treatment group (model estimates LMM: mean ± SE: 0.68 ± 0.04; Table 6, Fig. 5a), followed by the audiovisual (mean ± SE: 0.59 ± 0.04) and the audio-only tutees (mean ± SE: 0.58 ± 0.04). Likewise, for the tutee–tutor comparison, adding

'treatment' as fixed factor to the null model did not lead to a significant improvement (N = 41, χ_2 = 1.08, P = 0.58; Table 6). Again, similarity was highest in the live group (model estimates LMM: mean ± SE: 0.84 ± 0.03; Table 6, Fig. 5b), intermediate in the audiovisual group (mean ± SE: 0.80 ± 0.04) and lowest in the audio-only group (mean ± SE: 0.73 ± 0.03).

For the comparison of the tutor's and tutee's songs in SAP, there was no significant effect of tutoring treatment in the tutor–tutee or tutee–tutor comparison; the model including 'treatment' as fixed factor was not significantly better than the null model for the SAP tutor–tutee similarity scores (N = 41, χ_2 = 0.44, P = 0.80; Table 6, Fig. 5c) and the SAP tutee–tutor similarity scores (N = 41, χ_2 = 2.49, P = 0.29; Table 6, Fig. 5d).

Treatment had a significant effect on Luscinia similarity scores for the comparison between tutees and their tutors' songs (adding 'treatment' as fixed factor significantly improved the null model: N = 41, χ_2 = 8.72, P = 0.01; Fig. 5e, Table 6): this score was higher in the live than in the audiovisual tutees and there was a nonsignificant trend for the live tutored tutees also having higher scores than the audio-only tutees (for post hoc test results see Table 6).

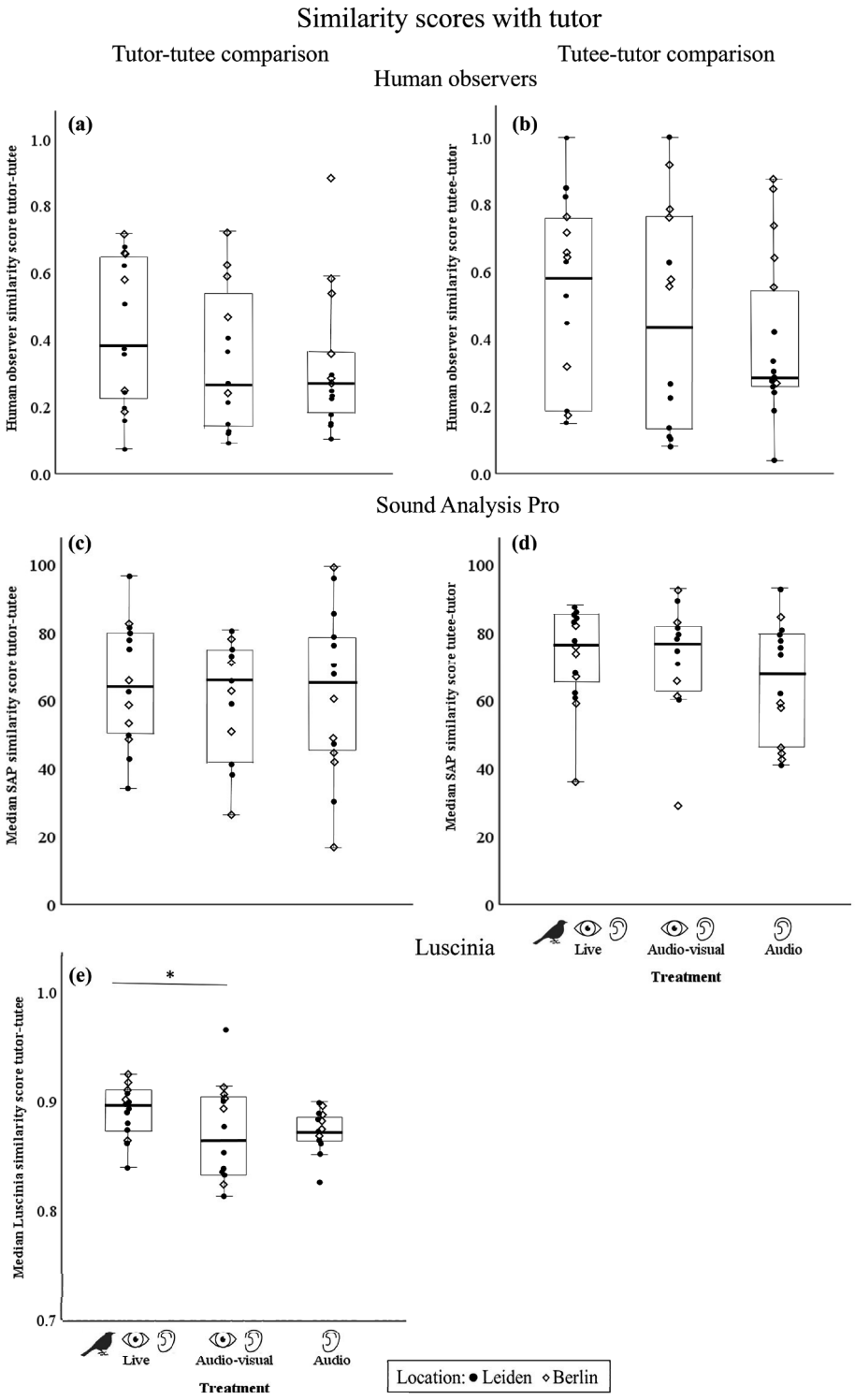
Similarity scores with tutor
Tutor-tutee comparison
Tutee-tutor comparison
Human observers
(a)
(b)
Human observer similarity score tutor-tutee
Human observer similarity score tutee-tutor
Sound Analysis Pro
(c)
(d)
Median SAP similarity score tutor-tutee
Median SAP similarity score tutee-tutor
Luscinia
(e)
*
Median Luscinia similarity score tutor-tutee
Live
Audio-visual
Audio
Treatment
Location: Leiden Berlin

Table 6 Details of models with 'Treatment' as fixed factor for the arcsine square-root transformed human observer, SAP and Luscinia 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***
Human observers sim. scores[1]	Intercept		0.74	0.07	11.19	0.87	0.10	8.47
	Treatment							
		Audiovisual	0.01	0.07	0.19	0.07	0.12	0.59
		Live	0.10	0.07	1.52	0.12	0.12	0.99
	Location							
		Leiden	-0.28	0.07	-4.06	-0.24	0.10	-2.40
SAP sim. scores[1]	Intercept		0.87	0.08	10.35	0.87	0.06	14.85
	Treatment							
		Audiovisual	-0.03	0.07	-0.36	0.08	0.06	1.36
		Live	0.02	0.07	0.28	0.08	0.06	1.38
	Location							
		Leiden	0.09	0.10	0.93	0.15	0.06	2.30
Lusc. sim. scores[2]	Intercept		0.09	0.0005	203.8			
	Treatment							
		Audiovisual	-0.0004	0.0005	-0.74			
		Live	0.001	0.0005	2.28			
	Location							
		Leiden	-0.001	0.0005	-2.73			

[1] LMM with 'Tutor group' as random factor.

[2] LMM with 'Tutor group' as random factor. Post hoc comparisons: audiovisual versus live: estimate = -0.001, SE = 0.0005, t = -2.92, P = 0.02; audio-only versus live: estimate = -0.001, SE = 0.0005, t = -2.28, P = 0.08; audio-only versus audiovisual: estimate = 0.0004, SE = 0.0005, t = 0.737, P = 0.74.

Similarity between tutees' and their fathers' songs

We also checked whether birds had learned from the father with which they had been housed before the experimental tutoring (Böhner, 1990). For the human observer similarity scores in the comparison of the father's syllables to the tutee's syllables (father–tutee comparison), adding 'treatment' as fixed factor did not significantly improve the null model (N = 39, χ_2 = 3.38, P = 0.18), but

as above we kept the experimental 'treatment' as fixed factor in the final model (Table 7). The similarity scores for the father–tutee comparison were highest in the group that learned the least during the experimental phase, namely the audio-only group (model estimates LMM: mean ± SE: 0.72 ± 0.03; Table 7, Fig. 6a), followed by the live (mean ± SE: 0.64 ± 0.02) and the audiovisual group (mean ± SE: 0.58 ± 0.03). For the human observer similarity scores in the tutee–father comparison, adding 'treatment' as fixed factor also did not significantly improve the null model (N = 39, χ_2 = 0.20, P = 0.91). The human observer similarity scores for this comparison were highest in the audio-only group (mean ± SE: 0.60 ± 0.09; Table 7, Fig. 6b) compared to the audiovisual (mean ± SE: 0.57 ± 0.10) and live groups (mean ± SE: 0.56 ± 0.09).

For the comparison of the father's song and tutee's song in SAP, there was no significant effect of tutoring treatment in the father–tutee or tutee–father comparison (model including 'treatment' as fixed factor was not significantly better than the null model for the SAP father–tutee comparison (N = 39, χ_2 = 2.07, P = 0.35; Table 7, Fig. 6c) and the SAP tutee–father comparison (N = 39, χ_2 = 0.23, P = 0.89; Table 7, Fig. 6d).

Treatment did not significantly affect Luscinia similarity scores for the comparison between tutees and their fathers' songs (model with 'treatment' as fixed factor was not significantly better than the null model: N = 39, χ_2 = 3.31, P = 0.19; Table 7, Fig. 6e).

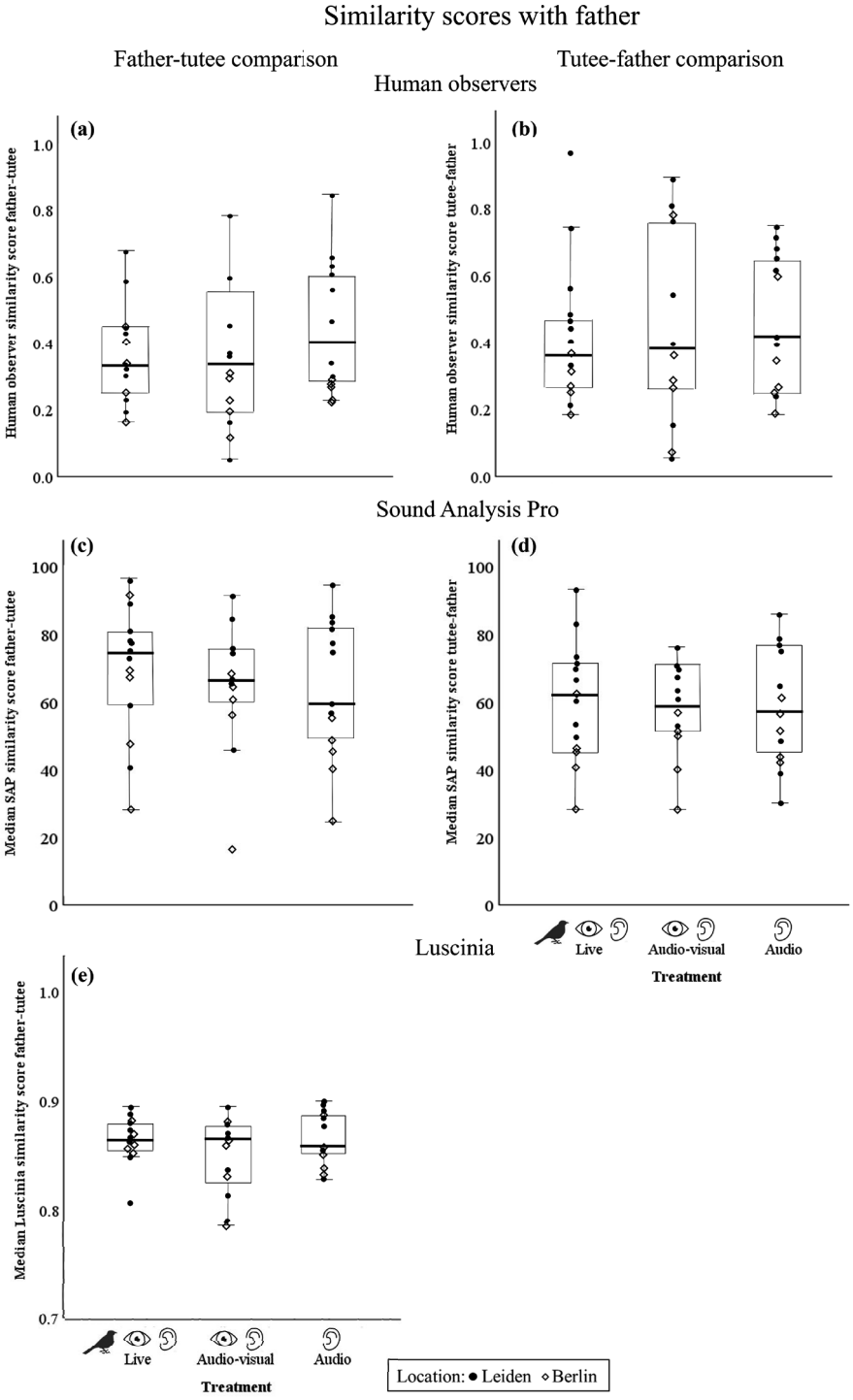
Similarity scores with father
Father-tutee comparison
Tutee-father comparison
Human observers
(a)
(b)
Human observer similarity score father-tutee
Human observer similarity score tutee-father
1.0
0.8
0.6
0.4
0.2
0.0
Sound Analysis Pro
(c)
(d)
Median SAP similarity score father-tutee
Median SAP similarity score tutee-father
100
80
60
40
20
0
Live
Audio-visual
Audio
Treatment
Luscinia
(e)
Median Luscinia similarity score father-tutee
1.0
0.9
0.8
0.7
Live
Audio-visual
Audio
Treatment
Location: ● Leiden ◇ Berlin

Table 7 Details of models with 'Treatment' as fixed factor for the arcsine square-root transformed human observer, SAP and Luscinia similarity scores for the comparison of father and tutee song

			Father-tutee			*Tutee-father*		
Response variable	***Model term***	***Level***	***Estim.***	***SE***	***t***	***Estim.***	***SE***	***t***
Human observers sim. scores[1]	Intercept		0.61	0.06	9.76	0.60	0.09	6.84
	Treatment							
		Audio-visual	-0.13	0.07	-1.74	-0.03	0.10	-0.28
		Live	-0.08	0.07	-1.18	-0.04	0.09	-0.43
	Location							
		Leiden	0.18	0.06	3.00	0.23	0.09	2.69
SAP sim. scores[1]	Intercept		0.78	0.08	10.04	0.74	0.05	14.50
	Treatment							
		Audio-visual	0.003	0.06	0.05	-0.006	0.06	-0.10
		Live	0.07	0.06	1.28	0.02	0.06	0.33
	Location							
		Leiden	0.23	0.09	2.60	0.20	0.05	4.11
Lusc. sim. scores[1]	Intercept		0.09	0.0005	174.6			
	Treatment							
		Audio-visual	-0.0009	0.0006	-1.59			
		Live	-0.0001	0.0006	-0.12			
	Location							
		Leiden	0.0006	0.0005	1.13			

[1] LMM with 'Tutor group' as random factor.

Changes between 65 and > 100 dph

The amount of changes in the typical motif between 65 and >100 DPH differed per treatment group: there were more changes in the motif of the audio-only tutored birds than the live tutored birds and there was a nonsignificant trend for the audio-only birds showing more changes in their motif than the audio-visual birds (model with 'treatment' significantly better than the null model: N = 33, $\chi_2 = 9.29$, $P < 0.01$; Fig. 7, for post hoc test results see Table 8). Of the 12 birds in the audio-only group that we could record at both 65 and >100 DPH, three did not change anything, five had added one or more syllables to their typical motif and four birds had deleted one or more syllables from their typi-

cal motif between 65 and >100 DPH.

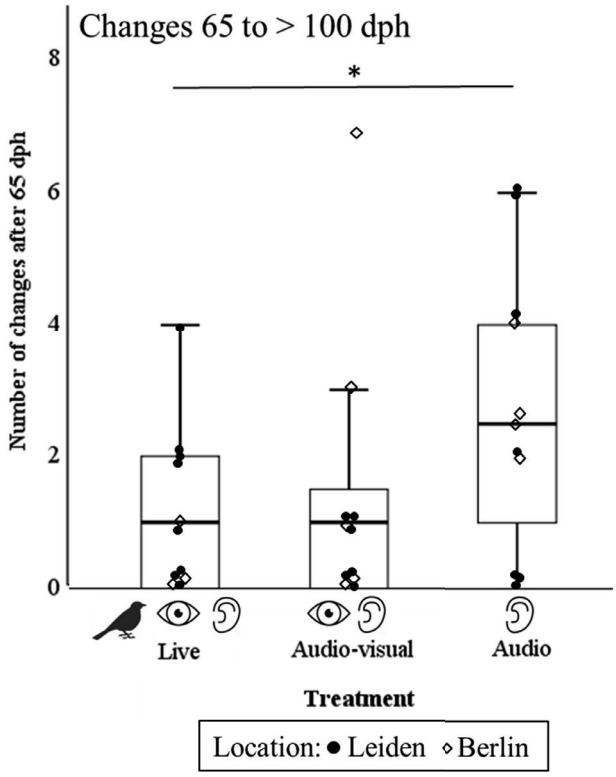

Figure 7. Number of changes in the typical motif of the tutees in the three treatment groups between 65 and >100 days posthatching. Box plots indicate the median (mid-line), interquartile range (box) and 1.5 times the interquartile range (whiskers). *P < 0.05, GLMM see Table 8.

Table 8. Details of best model (GLMM) for the number of changes between 65 and > 100 days posthatching (response variable)

Model term	***Level***	***Estimate***	***SE***	***z***	***p***
Intercept		1.01	0.38	2.66	**<0.01**
Treatment					
	Audio-visual	-0.77	0.33	-2.33	**0.02**
	Live	-0.91	0.36	-2.53	**0.01**
Location					
	Leiden	-0.25	0.45	-0.57	0.57

GLMM with Poisson distribution and 'Tutor group' as random factor. Post hoc comparisons: audio-only versus live: estimate = 0.91, SE = 0.36, z = 2.53, P = 0.03; audio-only versus audio-visual: estimate = 0.77, SE = 0.33, z = 2.32, P = 0.052; audiovisual versus live: estimate = 0.20, SE = 0.40, z = 0.49, P = 0.88. Significant P values are given in bold.

Discussion

The aim of this study was to test whether audiovisual exposure to a live tutor would facilitate song learning in zebra finches in comparison to auditory exposure only. To test this hypothesis, birds within each tutor group (with 15 replicates) were simultaneously tutored by an adult male in one of three conditions: cohousing with the tutor ('live' tutoring group), audio-only exposure to the same tutor via an acoustically transparent loudspeaker cloth ('audio') or both auditory (loudspeaker cloth) and visual (through a one-way mirror) exposure to the same tutor ('audiovisual'). Overall, the findings in this study suggest improved song learning in the tutees with audiovisual tutor exposure ('live' and 'audiovisual' group) compared to the group with audio-only tutor exposure ('audio'). First, the tutees with audiovisual exposure showed a different developmental trajectory: they made fewer changes between the subadult and adult recordings than the tutees with audio-only exposure. Second, their songs tended to have a higher similarity to the tutor's song and a lower similarity to the father's song (to which they were exposed before the peak of the sensitive period for song learning) than the tutees with audio-only exposure, but these between-treatment differences in similarity were not significant.

Birds from the audio-only condition differed from the birds in the other treatments in how their song developed. During song development, tutees start producing highly variable subsong, which becomes more stereotyped in structure and sequence over time. In socially reared zebra finches, around 60 DPH almost all syllables of the final song are produced and often in the same sequence as in adulthood (Arnold, 1975; Slater, Eales, & Clayton, 1988). A higher number of changes after 65 DPH might thus indicate a delay in song development, compared to birds with fewer changes after 65 DPH. More song plasticity after the peak of the sensitive period for song learning has been found in zebra finches housed in social isolation between 35 and 120 DPH compared to zebra finches housed in peer groups during this period (Jones et al. 1996). In the study described here, there were more changes in the typical motif of the audio-only birds than in that of the live birds after 65 DPH, which is in line with earlier findings showing that zebra finch tutees that were only auditorily exposed to adult conspecifics during the sensitive period for song learning change their song up until a later age than control birds reared in aviaries together with adult conspecifics (Morrison & Nottebohm, 1993). This earlier study concluded that the closing of the sensitive period depends on whether a bird was able to have visual social interactions with a tutor. Here, however, we did not find a difference in the number of changes between the live group and the audiovisual group, while tutees in this latter group could not have visual

social interactions with the tutor. This suggests that the timing of song development might be influenced not only by visual social interaction, but also by mere visual exposure to the tutor. Our results thus provide direct support for the hypothesis that seeing as well as hearing a tutor has a facilitating effect on the timing of song learning, independent of visual tutor–tutee interaction.

The zebra finch literature generally reports less learning from audio-only than from live tutors (e.g. Derégnaucourt et al., 2013; Eales, 1989; reviewed in Derégnaucourt, 2011). In contrast, in the current study, audio-only and live tutored birds did not differ significantly in how much they learned from the tutor. In our experiment, other than in earlier tape versus live tutor comparisons, audio-tutored birds could vocally interact with the live tutor and also had a nonsinging female as a social companion to avoid the potential confound of social isolation on song development in the audio-only and audiovisual groups, which could explain why audio-only and live tutored tutees showed comparable song learning in our study. It is also possible that being housed with a female companion resulted in fewer syllables copied from the tutor in all conditions, including the live condition, because tutees incorporated calls from the female in their song (see e.g. Price, 1979) or were reinforced by their female companions' behaviour to retain specific syllables that resembled the song of the female's father, which was different from that of the male tutee's father or tutor (Carouso-Peck et al., 2020; Carouso-Peck & Goldstein, 2019; Jones & Slater, 1993). We can compare the absolute scores directly with those from Phan et al. (2006), who computed SAP similarity scores in a similar fashion to us for live tutored tutees, but their tutees only had their father as a tutor and were housed continuously with their parents and siblings. These tutees had a higher average similarity to their tutor's song (71±4) than the tutees in the current study (62±3). Derégnaucourt et al. (2013) computed SAP similarity scores in a similar fashion for an experiment also involving a live and audio-only condition, but where all tutees' fathers had been removed at 25 days and where live and audio-only tutees were housed without a female companion. Derégnaucourt et al. (2013) also found a higher average SAP similarity score for the live tutored tutees (76±4) than we found in the current study (62±3), but a comparable average similarity score for the audio-only tutored tutees (60±4; this study: 61±3). This suggests that in the current study the female companion, the prolonged exposure to the father's song in most tutees or the vocal interaction with the other tutees behind the loudspeaker cloths (Honarmand et al., 2015) could have contributed to the lack of a significant difference between the live and audio-only tutored tutees.

The highest tutor song copying rates in the live group and the lowest tutor song copying rates in the audio-only group are in line with earlier studies that have shown that reduced quality of tutor access during the peak of the sensitive period for song learning (35–65 DPH) can lead to increased copying of song heard before this period (reviewed in Gobes et al., 2017), which in our experiment was the father's song. The audio-only group indeed showed the highest and the live tutored group the lowest father's song copying rate, although this was not significantly different. This suggests that causes other than social isolation or a lack of vocal isolation might play a role in the lower song copying by audio-only tutees as well. The lowest overall song copying in this group is in line with a lack of visual cues being one possible cause of poorer song copying in audio-only tutees.

The study included tutees raised in Leiden and Berlin. Compared to the Leiden Song similarity with tutor song was highest in the birds from the live group. Multiple factors may have contributed here: unlike the other tutees, live tutees could visually and physically interact with and physically approach the tutor (Liu et al., 2021), which might have contributed to song learning success. Besides, due to our experimental set-up, extrapolating from the acoustic transmission properties, unless the tutor was sitting close to the loudspeaker cloth, the tutor song was louder (and likely to be so on average) in the central compartment than in the other two compartments, which, if amplitude affects learning, might have contributed to the higher similarity scores of the live tutees. Our results are in line with other observations of improved learning from live tutors, but the question of which factors contribute to the improved song learning from live versus different types of audio-only tutors (ranging from stereotyped playbacks to visually occluded tutors, e.g. Baptista & Gaunt, 1997; Beecher, 2017; Houx & ten Cate, 1999; Nelson, 1997, 1998) is an ongoing discussion.

Earlier studies addressed the question whether additional visual stimulation improved song learning but have not found an effect. Presentations of a stationary taxidermic mount of a zebra finch male as a visual stimulus right before, during or after tutor song presentation or presentation of a video of a singing tutor synchronized with tutor song did not lead to more song copying than tape tutoring only (Bolhuis et al., 1999; Houx & ten Cate, 1999; Varkevisser et al., 2021). In the study presented here, the tutees with visual tutor exposure (the live and audiovisual groups) tended to have a higher similarity to the tutor's song and a lower similarity to the father's song than the birds without visual tutor exposure (the audio-only group), but this difference was not significant.

The study included tutees raised either in Leiden or Berlin. Compared to the Leiden tutees, the Berlin tutees produced songs more similar to their tutor's song and less similar to their father's song. It is difficult to pinpoint the reason for this at this stage, as the differences could be stochastic or arise from a number of differences in the technical realization of the experiment at the two locations. While all tutees in Leiden and Berlin were moved to the tutoring set-up at day 35, tutees in Berlin were in a different compartment of the cage than their father between day 23 and day 35, so that they could hear the father but not see him or interact with him physically. In Leiden the father remained in the same space with the juveniles until day 35. The tutees in Leiden may therefore have picked up more from their father than the tutees in Berlin. This will, however, need systematic study as there were also other differences between Berlin and Leiden. For example, in Berlin, the three compartment cages were positioned in a soundproof box, while in Leiden they were in a room with other birds present. This probably made the tutor in Berlin more audible to the tutees in the side compartments than in Leiden. Conversely, the tutees in the compartments adjacent to the tutor might also have been more audible to the tutor which could have led to more vocal interactions. Individual rather than population differences might also have contributed: there were only 15 tutors in total. If in our colonies some tutors are copied more and more readily by (related and unrelated) young birds as reported for other colonies (Tchernichovski et al., 2021) and some of these preferentially and better copied tutors were better represented in one location, this could be mistaken for a location or population effect.

The three similarity assessment methods used in this study (human observers, Luscinia and SAP) differed in whether they picked up a significant treatment effect. In previous studies, Luscinia and SAP were both found to be highly correlated with human observer similarity scores (Luscinia: $r = 0.96$, $N = 18$, Lachlan et al. 2010; SAP: $r = 0.91$, $N = 10$, Tchernichovski et al., 2000). In the current study, in all comparisons the human observers' scores were significantly correlated with the Luscinia but not the SAP scores. Our study design had several features that differed from the previous comparison between SAP and human observer similarity scoring (Tchernichovski et al., 2000) that might have led to this lower correlation. For instance, the previous study used only live tutoring and one tutor, and therefore tutees in the previous comparison probably copied more from the tutor than the tutees in the current study. As the correlation between human observers and SAP was weaker in the larger sample including all groups than in the smaller sample only involving the live tutored birds, SAP or the human observers might have more difficulty assessing similarity be-

tween model song and poorly copied tutee song, than between model song and well-copied song. Human visual scoring was used to validate the two automated methods and is considered a suitable method for assessing song similarity if multiple independent observers that are blind to the expected outcome of the comparisons are used (Jones, ten Cate, & Bijleveld, 2001), which was the case in our study, and which is why our conclusions are mainly based on the results from the human observer similarity scoring.

Overall, our findings suggest that birds with multimodal exposure developed their adult song faster and tended to produce songs that were more similar to the tutor's song than birds with unimodal tutor exposure, which is in line with our hypothesis that visual exposure to a singing tutor has a facilitating effect on zebra finch song learning. There are various ways in which visual exposure to the tutor might have facilitated song learning in this experiment. First, as we hypothesized, the beak and throat movements associated with song production might have made song easier to detect and remember, which would be in line with a study showing that visual stimulation matched in rhythm to auditory song presentation can facilitate song learning (Hultsch et al., 1999) and studies showing that stimuli with multiple components (in one or multiple modalities) are easier to detect and remember than unicomponent stimuli (reviewed in Hebets & Papaj, 2005; Rowe, 1999). On the other hand, the facilitating effect might not necessarily have to do with the coupling between visual and auditory song exposure. It might also be that the tutor provided visual feedback in reaction to songs produced by the tutees. Young zebra finches that received contingent visual feedback (a video of a female conspecific) on their immature song production copied more tutor song than birds that received noncontingent visual feedback (Carouso-Peck, & Goldstein, 2019) and an observational study showed that the number of fluff-ups performed by the mother before, during or after tutee song production was positively correlated with tutee song learning success (Carouso-Peck et al., 2020). In this observational study, no visible behaviour of the father was studied as possible feedback to juvenile song production (Carouso-Peck et al., 2020). From our tutoring experiment, we cannot rule out the possibility that the tutor provided visual feedback to the tutees, which might have facilitated song learning in the birds with visual access to the tutors. Follow-up studies, for example where no vocal interaction is possible between the tutor and tutees, could help find out which mechanism underlies the effect that visual exposure to a tutor has on zebra finch song learning.

In this study, we disentangled the effect of multimodal exposure to a tutor from that of social visual interaction with a tutor on the song learning process in

zebra finches. Our results suggest that multimodal exposure to a tutor affects zebra finch song development and might be one of the factors involved in the difference in song learning success from live and tape tutors. Follow-up studies are necessary to get more insight into the mechanism through which multimodal exposure to a tutor facilitates song learning. This can give more insight into the factors involved in the vocal learning process.

Acknowledgements
Funding for this research was provided by a Human Frontier Science Program Grant (No RGP0046/2016). We would like to thank Carel ten Cate for comments on an earlier version of this manuscript and members of the Behavioural Biology group in Leiden for discussion. We would like to thank Jing Wei, Quanxiao Liu and Zhiyuan Ning for the visual comparison of the spectrograms.

References

Arnold, A. P. (1975). The effects of castration on song development in zebra finches (*Poephila guttata*). Journal of Experimental Zoology, 191(2), 261–278. https://doi.org/10.1002/jez.1401910212

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.

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

Bischof, H. J. (1988). The visual field and visually guided behavior in the zebra finch (*Taeniopygia guttata*). Journal of Comparative Physiology A, 163, 329–337. https://doi.org/10.1007/BF00604008

Böhner, J. (1986). Der zeitliche Verlauf des Gesangerwerbs beim Zebrafinken. Ver handlungen Der Deutschen Zoologischen Gesellschaft, 79.

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

Carouso-Peck, S., & Goldstein, M. H. (2019). Female social feedback reveals non-imitative mechanisms of vocal learning in zebra finches. Current Biology, 29, 631–636. https://doi.org/10.1016/j.cub.2018.12.026

Carouso-Peck, S., Menyhart, O., DeVoogd, T. J., & Goldstein, M. H. (2020). Contingent parental responses are naturally associated with zebra finch song learning. Animal Behaviour, 165, 123–132. https://doi.org/10.1016/j.anbehav.2020.04.019

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

Clayton, N. S. (1988). Song tutor choice in zebra finches and bengalese finches: the relative importance of visual and vocal cues. Behaviour, 104, 281–299.

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 gut tata)* 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

Gobes, S. M. H., Jennings, R. B., & Maeda, R. K. (2017). The sensitive period for auditory-vocal learning in the zebra finch: consequences of limited-model availability and multiple-tutor paradigms on song imitation. Behavioural Processes, 163, 5–12. https://doi.org/10.1016/j.beproc.2017.07.007

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/MUv110n3ED

Halfwerk, W., Varkevisser, J., Simon, R., Mendoza, E., Scharff, C., & Riebel, K. (2019). Toward testing for multimodal perception of mating signals. Frontiers in Ecology and Evolution, 7, 2013–2019. https://doi.org/10.3389/fevo.2019.00124

Hebets, E. A., & Papaj, D. R. (2005). Complex signal function: Developing a framework of testable hypotheses. Behavioral Ecology and Sociobiology, 57(3), 197–214. https://doi.org/10.1007/s00265-004-0865-7

Helekar, S. A., Marsh, S., Viswanath, N. S., & Rosenfield, D. B. (2000). Acoustic pattern variations in the female-directed birdsongs of a colony of laboratory-bred zebra finches. Behavioural Processes, 49(2), 99–110. https://doi.org/10.1016/S0376-6357(00)00081-4

Higham, J. P., & Hebets, E. A. (2013). An introduction to multimodal communication. Behavioral Ecology and Sociobiology, 67(9), 1381–1388. https://doi.org/10.1007/s00265-013-1590-x

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. (1999). 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
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
Hyland Bruno, J., & Tchernichovski, O. (2019). Regularities in zebra finch song beyond the repeated motif. Behavioural Processes, 163, 53–59. https://doi.org/10.1016/j.beproc.2017.11.001
Immelmann, K. (1969). Song development in the zebra finch and other estrildid finches. In R. A. Hinde (Ed.), Bird vocalizations. Cambridge, England: Cambridge University Press,.
Jones, A. E., & Slater, P. J. B. (1993). Do young male zebra finches prefer to learn songs that are familiar to females with which they are housed. Animal Behaviour, 46, 616–617. https://doi.org/10.1006/anbe.1993.1233
Jones, A. E., ten Cate, C., & Bijleveld, C. C. J. H. (2001). The interobserver reliability of scoring sonagrams by eye: A study on methods, illustrated on zebra finch songs. Animal Behaviour, 62(4), 791–801. https://doi.org/10.1006/anbe.2001.1810
Jones, A. E., ten Cate, C., & Slater, P. J. B. (1996). Early experience and plasticity of song in adult male zebra finches (*Taeniopygia guttata*). Journal of Comparative Psychology, 110(4), 354–369. https://doi.org/10.1037/0735-7036.110.4.354
Lachlan, R. F., Verhagen, L., Peters, S., & ten Cate, C. (2010). Are there species-uni versal 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
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: estimated marginal means, aka least-squares means.
Liu, W. chun, Landstrom, M., Schutt, G., Inserra, M., & Fernandez, F. (2021). A memory-driven auditory program ensures selective and precise vocal imitation in zebra finches. Communications Biology, 4(1). https://doi.org/10.1038/s42003-021-02601-4
Mann, N. I., & Slater, P. J. B. (1995). Song tutor choice by zebra finches in aviaries. Animal Behaviour, 49(3), 811–820. https://doi.org/10.1016/0003-3472(95)80212-6
Mann, N. I., Slater, P. J. B., Eales, L. A., & Richards, C. (1991). The influence of visual stimuli on song tutor choice in the zebra finch, *Taeniopygia guttata*. Animal Behaviour, 42(2), 285–293. https://doi.org/10.1016/S0003-3472(05)80560-3
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

Morrison, R. G., & Nottebohm, F. (1993). Role of a telencephalic nucleus in the delayed song learning of socially isolated zebra finches. Journal of Neurobiology, 24(8), 1045–1064.

Partan, S., & Marler, P. (1999). Communication goes multimodal. Science, 283, 1272–1274. https://doi.org/0.1126/science.283.5406.1272

Price, P. H. (1979). Developmental determinants of structure in zebra finch song. Journal of Comparative and Physiological Psychology, 93(2), 260–277. https://doi.org/10.1037/h0077553

Rowe, C. (1999). Receiver psychology and evolution of multicomponent signals. Animal Behaviour, 58, 921–931. https://doi.org/10.1006/anbe.1999.1242

Scharff, C., & Nottebohm, F. (1991). A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: Implica tions for vocal learning. Journal of Neuroscience, 11(9), 2896–2913. https://doi.org/10.1523/JNEUROSCI.11-09-02896.1991

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 re search. 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

Tchernichovski, O., Eisenberg-Edidin, S., & Jarvis, E. (2021). Balanced imitation sustains song culture in zebra finches. Nature Communications, 1–21. https://doi.org/10.1038/s41467-021-22852-3

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

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-real istic 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. (1990). Models for song learning in the zebra finch: fathers or others?

Animal Behaviour, 39(4), 745–757. https://doi.org/10.1016/S0003-3472(05)80386-0

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

Appendix

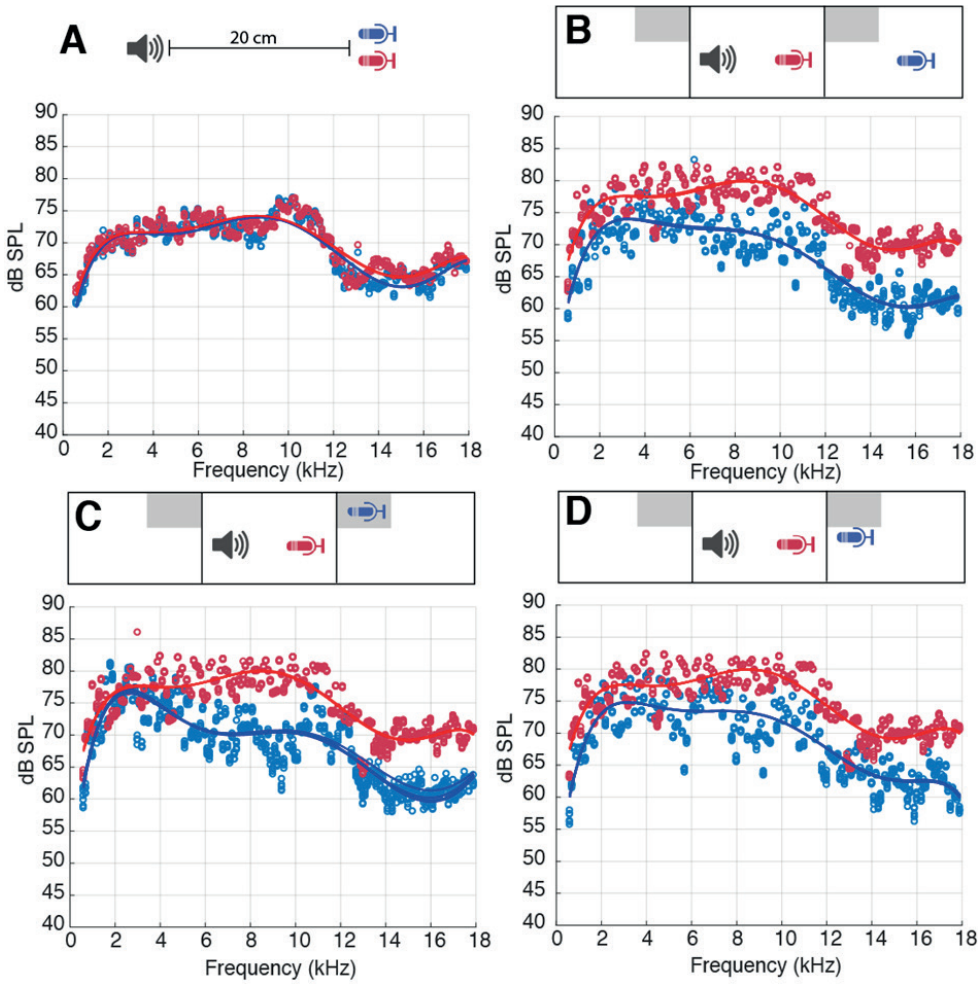

Figure A1. Acoustic transmission properties of the cages measured with two free-field microphones (40BF, preamplifier 26AB, power module 12AA; G.R.A.S. Sound & Vibration) and one speaker (Vifa, Viborg, Denmark) which played tones in frequency steps on 200 Hz. (A) Frequency response for a reference measurement where both microphones were installed in 20cm distance to the speaker. (B) Frequency response where one microphone (red) was in the central compartment and the other one (blue) was in the neighbouring compartment at the same height. (C) Frequency response where one microphone (red) was in the central compartment and the other one (blue) was in the neighbouring compartment in the 'observation hut' behind the one-way mirror. (D) Frequency response where one microphone (red) was in the central compartment and the other one (blue) was in the neighbouring compartment next to the observation hut not covered by the one-way mirror.

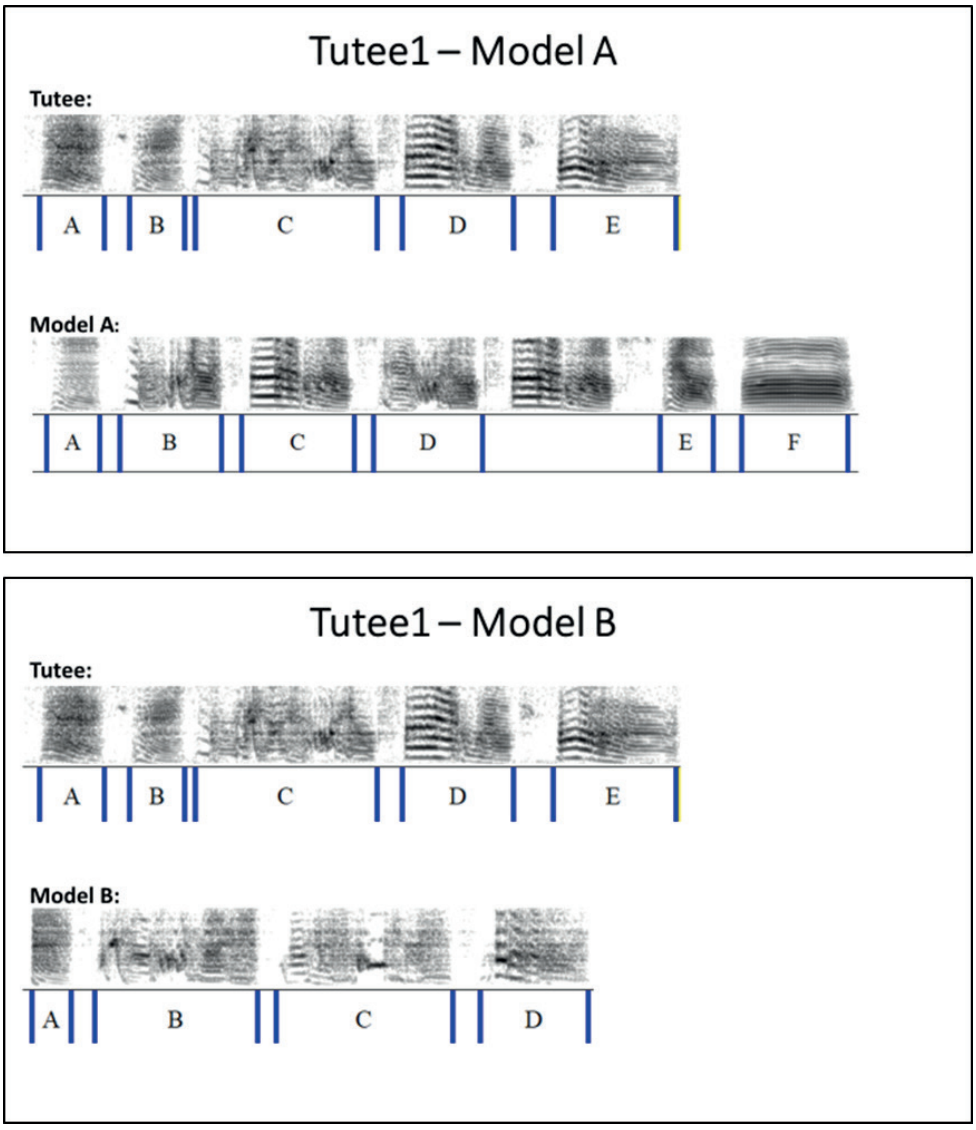

Figure A2. Example of slides used for human observer similarity scoring.

Chapter 4

Adding colour realistic video images to audio playbacks increases stimulus engagement but does not enhance vocal learning in zebra finches

Judith Varkevisser, Ralph Simon, Ezequiel Mendoza, Martin How, Idse van Hijlkema, Rozanda Jing, Qiaoyi Liang, Constance Scharff, Wouter Halfwerk & Katharina Riebel

This chapter is published in Animal Cognition:
https://doi.org/10.1007/s10071-021-01547-8

Abstract
Bird song and human speech are learned early in life and for both cases engagement with live social tutors generally leads to better learning outcomes than passive audio-only exposure. Real-world tutor-tutee relations are normally not uni- but multimodal and observations suggest that visual cues related to sound production might enhance vocal learning. We tested this hypothesis by pairing appropriate, colour-realistic, high frame-rate videos of a singing adult male zebra finch tutor with song playbacks and presenting these stimuli to juvenile zebra finches (*Taeniopygia guttata*). Juveniles exposed to song playbacks combined with video presentation of a singing bird approached the stimulus more often and spent more time close to it than juveniles exposed to audio playback only or audio playback combined with pixelated and time-reversed videos. However, higher engagement with most realistic audio-visual stimuli was not predictive of better song learning. Thus, although multimodality increased stimulus engagement and biologically relevant video content was more salient than colour and movement equivalent videos, the higher engagement with the realistic audio-visual stimuli did not lead to enhanced vocal learning. Whether the lack of three-dimensionality of a video tutor and/or the lack of meaningful social interaction make them less suitable for facilitating song learning than audio-visual exposure to a live tutor remains to be tested.

Introduction
Bird song is one of the best-studied animal examples of vocally learned signalling (Catchpole and Slater 1995) and it is often used as a model system for human speech acquisition, because of the many similarities between human speech and bird song (Doupe and Kuhl 1999; Bolhuis et al. 2010). One of the open research questions in the study of both speech and bird song development is whether, and to what extent, exposure to the visual cues accompanying the production of vocalizations, such as lip movements in speech and beak movements in bird song, plays a role in vocal development (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). Given the well-established experimental tutoring paradigms, bird song offers a system in which the effect of visual cues on the vocal learning process can be studied experimentally (Doupe and Kuhl 1999; Brainard and Doupe 2002; Goldstein et al. 2003).

In the study of bird song learning, experimental tape tutoring has been crucial. Instead of learning from a bird that is physically present, young birds are

tutored by playing back pre-recorded conspecific song via loudspeakers, either under operant control of the juvenile bird or passively (Derégnaucourt 2011). These methods allow researchers control over the quantity, quality and timing of song exposure. This high level of experimental and stimulus control has greatly contributed to understanding vocal learning processes (Catchpole and Slater 1995; Derégnaucourt 2011). Not all songbird species, however, learn as well from a tape tutor as from a live conspecific (reviewed in Baptista & Gaunt 1997; Soma 2011). Many researchers have argued that this is because social interaction with a tutor is important for song learning (e.g. see Baptista and Petrinovich 1986; Slater et al. 1988; Catchpole and Slater 1995; Carouso-Peck et al. 2020). However, tape and live tutors differ in more aspects than sociality. For example, bird song, like much animal communication, is multimodal, offering simultaneous information from several modalities (Partan and Marler 1999; Higham and Hebets 2013; Halfwerk et al. 2019). Bird song production is accompanied by visual components, such as beak, head, throat and body movements. Multimodal signals are often easier detected and remembered by receivers than unimodal signals (reviewed in Rowe, 1999) and might thus be beneficial to learning. In line with this, improved learning of paired auditory -visual stimuli has been demonstrated in several bird species and contexts, for example in the context of filial imprinting (van Kampen & Bolhuis, 1991; van Kampen & Bolhuis, 1993) or song learning (e.g. in nightingales, *Luscinia megarhynchos*, Hultsch, Schleuss, & Todt, 1999). However, the difference between multi- and unimodal tutoring has rarely been considered in the discussion on why several bird species learn better from live than from tape-tutors (Nelson, 1997, Baptista & Gaunt, 1997; Soma, 2011).

One of the songbird species often cited for learning poorly from audio playbacks is the zebra finch (*Taeniopygia guttata*), an important animal model to study vocal learning (Griffith and Buchanan 2010; Mello 2014). Zebra finches learn better from a live tutor than when passively exposed to audio-only presentation of tutor song (Eales 1989; Derégnaucourt et al. 2013; Chen et al. 2016). The most favoured hypothesis regarding these differences is that social interactions with a tutor increase the salience of the tutor song (Chen et al., 2016; Derégnaucourt et al., 2013; Slater, Eales, & Clayton, 1988). However, social and tape tutors also differ in non-social aspects: tape tutoring is often more stereotyped than a live tutor, shows no circadian activity patterns, is less or not interactive and is often non-contingent on tutee behaviour (for discussion see Nelson 1997). The effect of contingencies on song learning has seen some experimental testing in zebra finches, but with mixed results regarding whether they facilitate song learning from playback and whether similar

learning outcomes can be attained with behaviour contingent playback as with live tutoring (ten Cate 1991; Adret 1993; Houx and ten Cate 1999a; Phan et al. 2006; Derégnaucourt et al. 2013). There is, however, yet an additional systematic difference that studies investigating social versus non-social tutoring have not controlled for, namely the multi- versus unimodal presentation of song in live compared to classic tape tutoring paradigms. In this study, we aim to specifically test whether multimodal exposure (rather than social interaction) to a tutor might improve learning and could thus (partly) explain the differences in learning from tape and live tutors. To do so, a method is required that allows investigating whether song learning from passive song playback is improved by simultaneous visual exposure to the singing tutor when, akin to tape tutoring, tutees cannot also socially interact with the song tutor.

This study follows up on earlier pioneering experiments that added visual stimuli right before, during or after the presentation of tutor song and found no improvement of learning with the added visual stimuli (Bolhuis, van Mil, & Houx, 1999; Houx & ten Cate, 1999). These studies used non-moving taxidermic mounts of male zebra finches as visual stimuli, which might have been suboptimal because they were stationary (Bolhuis et al. 1999). Interestingly, painted plaster images of female conspecifics were sufficient to stimulate adult males to sing more than when alone (Bischof et al. 1981), suggesting that the degree of naturalistic visual stimulation necessary for song learning in juveniles and song production in adults might differ.

Videos provide moving images, but when using videos in animal research, it should be taken into consideration that standard video systems are designed for human visual perception. This aspect was until recently rarely controlled and adjusted for during video stimulus preparation and presentation to animals that often have different colour and movement perception (Chouinard-Thuly et al. 2017). Birds have a higher flicker-fusion frequency and different colour, brightness and depth perception than humans (Cuthill et al. 2000; Fleishman and Endler 2000; Oliveira et al. 2000). It is unclear, however, how much deviation from naturalist colour and movement fluidity is still acceptable to birds. Human-vision adapted videos can trigger natural behaviour in zebra finches, such as copying food choices from demonstrators via live streaming video's (Guillette and Healy 2016) or courtship singing by males towards females on video screens (Ikebuchi and Okanoya 1999; Galoch and Bischof 2007; James et al. 2019) and presenting a video of a female conspecific contingent with immature song production by juvenile male zebra finches improves song learning (Carouso-Peck and Goldstein 2019). Importantly, zebra finches do react

differently to a video than a live presentation of particular stimuli (Ikebuchi and Okanoya 1999; Swaddle et al. 2006; Guillette and Healy 2019; James et al. 2019). Zebra finches tutored with a passive or operant video tutor copied song poorly (Adret 1997). Adret (1997) speculated that the poor sound quality of the TV monitor loudspeakers used for playbacks might have been responsible for the poor learning and other authors later wondered whether the low flicker frequency of the monitor in this experiment was suboptimal (Derégnaucourt 2011). Neither factor has been systematically tested so far in the context of song learning. High fidelity audio-video playbacks could open a window into investigating the potential role of multimodal cues in song learning, a potential confound of 'social' tutoring not controlled in classic audio-only playback studies. Deshpande, Pirlepesov, and Lints (2014) conducted a study in which juvenile zebra finches had operant control over either just audio or audio-visual (simultaneous or staggered audio and video) playback of song. In this study, only the groups tutored with simultaneous audio-visual playback or with staggered playback where audio preceded video showed significant song learning compared to birds without tutoring. Song learning in all birds was poor, possibly because the video was suboptimally adjusted to avian vision (e.g. no colour adjustments) or because of the very limited amount of song exposure that birds received (only 75 seconds in total over the sensitive period for song learning). In addition, only one tutor video was used, so any unintended cue or flaw in this particular video may have unduly influenced the results. Technical advancement and increased insights into avian vision allow addressing several of the potential issues with stimulus quality discussed above and formulated in a recent consensus on the usage of video stimuli in animal research (Chouinard-Thuly et al. 2017). A recurrent, neglected issue in this context is how the frame rate of the presented video relates to the study species' speed of vision. Neglecting such aspects can affect animals' responses, as has been demonstrated for social responses of pigeons, *Columbia livia*, towards video stimuli (Ware et al. 2015). In the present study, we therefore made use of recent technical, empirical and theoretical advancements to produce videos of multiple tutors. We recorded them with high frame rates (120 fps) to accommodate the higher temporal resolution of zebra finch vision and videos were displayed on gaming monitors with high refresh rates (120 Hz), which, in combination with the high frame rates of the video itself, should make the movements in the videos look smooth to the birds. We also adjusted the colours of our videos following the 'colour realistic imagery' technique (Tedore and Johnsen 2017), to mimic as closely as possible the animals' colour perceptual experience of a real conspecific. Combining these videos with high quality sound recordings, enabled us to present auditory and visual information linked in real-time (or experimen-

tally dissociated) to zebra finch tutees, thus controlling for currently known potential sources of artefacts (Chouinard-Thuly et al. 2017).

In the current study, tutees were exposed to either audio playback only or to song playbacks accompanied by colour realistic videos of the singing tutor or in a control condition by the same colour realistic, but now pixelated and reversed versions of the video stimuli. If accurate rhythmic correspondence between the beak, head, and throat movements and the song facilitates song learning, the birds receiving video presentations of the tutor together with audio playback should show improved song learning. It is also possible that any moving visual stimulus presented together with the song would facilitate song learning. For instance, the detectability of a signal can be positively affected if it is presented together with an additional stimulus in another sensory modality, possibly by drawing the receiver's attention to the signal (Feenders, Kato, Borzeszkowski, & Klump 2017; reviewed in Rowe 1999). We therefore also included a group of tutees exposed to videos created by pixelating the frames of the original videos before playing them back in reversed order. This created videos of comparable complexity in colours and movements without presenting a video image of a bird and without direct rhythmic correspondence between the song and the video. To prevent possible effects that social isolation might have on song learning, which in tape versus live tutoring is a rarely addressed confound (Varkevisser et al., in prep.), we decided to not house the tutees solitarily, as was usually the case in previous zebra finch tape tutoring studies (e.g. Bolhuis et al. 1999; Derégnaucourt et al. 2013; Houx & ten Cate 1999), but together with an age-matched female companion. Being housed with a companion will likely be beneficial for welfare and can potentially motivate a bird to sing (Jesse and Riebel 2012), thereby creating a better comparison with a situation where a live tutor is present. As all female companions, like the male tutees, came from families where the father had been removed before the onset of the sensitive phase for song learning, females might reinforce singing in males (as in the natural nest), but any influence they might have will be unspecific with regard to song content.

By thus keeping the social environment the same, but varying whether song presentation was accompanied by visual stimulation (song unspecific versus song specific), we created an experimental situation to test the hypothesis that visual stimulation in addition to auditory song exposure facilitates song learning. If this were the case, then all video tutored birds should learn better compared to birds experiencing only unimodal auditory song exposure. In addition, the video groups might differ from each other in learning outcomes

if visual exposure to the specific movements accompanying song production, e.g. song related beak and body movements, had greater salience in this context than equally colourful and equally animated, but unspecific visual exposure. This expectation was based on the human literature where such sound-specific motor gestures attract the attention of infants more than unspecific gestures (Kuhl and Meltzoff 1982; Patterson and Werker 1999), but also on increased insights from the animal literature showing effects of correctly synchronised visual and acoustic information on perceptual salience (e.g. Taylor et al. 2011; Rek 2018). We thus expected the tutor videos with the synchronous auditory-visual information to lead to better song learning than the pixelated and reversed videos.

Methods

Subjects and housing

We used 44 juvenile males and 44 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, calculated as days from the median hatching day within a 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 the birds 2.40 m across on the opposite side of the aisle could also be seen. When subjects reached 35 dph, they were moved in dyads consisting of a young male and an unrelated young female into sound-attenuated chambers (125 x 300 x 240 cm) for song tutoring (details below) until they reached 65 dph, when they were moved to a recording cage (see below). After recording at 65 dph, the dyads were housed 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 13.5/10.5 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 and Böhner 1980). Motifs were defined as the repeated syllable sequence in a song, and syllables as sounds separated from other sounds by at least 5 milliseconds of silence.

A male-female tutee dyad was exposed to one of three different tutoring treatments (Figure 1): 1) song only playback ("Audio"), 2) song playback combined with a time-aligned video of the tutor singing ("Audio-video") or 3) song playback combined with a pixelated version of the same video and with the individual frames of the video played back in reversed order ("Audio-pixel").

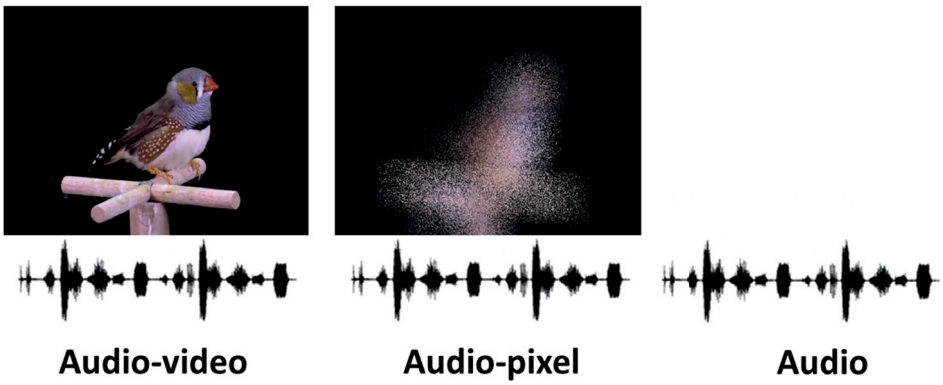

Figure 1. Overview of the different tutoring treatments in this study. The Audio-video treatment consists of a synchronous sound and video exposure (120 fps video, sound and beak movements aligned, for an example see Online Resource 1); the Audio-pixel treatment consists of the same song and the same video, but the video is pixelated and played back in reversed order (for an example see Online Resource 2) and in the Audio treatment only the audio channel of the song is played back.

We used song from twelve different tutors. The same tutor song was presented to three tutee dyads, each in a different tutoring treatment (Audio, Audio-video and Audio-pixel). Tutees exposed to the same tutor song were tutored simultaneously and will be referred to as one 'tutor group'. We raised 12 tutor groups with these three treatments. Due to a technical delay in another experiment, additional young birds could be tutored and post-hoc, we raised four additional tutor groups. In these four groups, we only included the Audio-video and Audio-pixel treatment to increase the statistical power for the pairwise comparisons in the subquestion as to whether the quality of the video material affected learning. For these groups, we used four tutors that had previously been used as tutors for other groups. Within one tutor group, wherever possible, all males and all females originated from the same nest (all 3 male

siblings: 8/12 tutor groups; 2 siblings and 1 additional male: 3/12 tutor groups; 3 unrelated males: 1/12 tutor groups; all 3 female siblings: 11/12 tutor groups; 2 siblings and 1 additional female: 1/12 tutor groups). Tutoring took place between 35 and 65 days post-hatching. Tutor songs were presented in daily tutoring sessions following one of three different tutoring schedules (see Table 1 for details). For each tutor, per treatment, three different stimuli were made which were played back in random order throughout the day. It is currently unclear how often a tutee should hear a tutor song to optimally learn it. Some studies suggested that a high amount of song exposure might negatively affect zebra finch song learning (Tchernichovski et al. 1999; Tchernichovski and Mitra 2002; Chen et al. 2016). However, previous passive play-back studies have found a low degree of tutor song copying using exposure frequencies ranging from 20 (Derégnaucourt et al. 2013) to approximately 250 songs per day (Bolhuis et al. 1999; Houx and ten Cate 1999b). Even less is known about how much a tutee should be exposed to a video tutor, but given the limitations of producing sufficient high-quality videos and a potential effect of overexposure, we decided to first offer limited song exposure to the first three tutor groups. These groups (i.e. 3 x 3 male tutees, in the Audio-video, Audio-Pixel and Audio condition) received three tutoring sessions daily with 10 songs played during each session (schedule 1). We made daily observations of how tutees responded to the stimulus presentation (through the one-way mirrors in the doors of the sound-attenuated chambers). At the end of the song tutoring period, tutees in these groups still responded to the stimulus presentation by approaching the loudspeakers and thus did not seem to lose interest in the stimuli over time. We also observed that it took a while before the birds reached the best position to see the videos, which they left again during the inter-song intervals. This sometimes meant they only saw part of the video. We thus decided to increase the number of tutoring sessions and the amount of song presented per session and to shorten the inter-song intervals. The next 9 tutor groups thus received four tutoring sessions daily with 12 songs per session (schedule 2). As the tutees still seemed to remain interested in the stimuli throughout the experiment, we decided to increase exposure even further during the third schedule. Given the exploratory nature of the study, using several exposure frequencies seemed safest to detect potential effects of exposure frequency that could then inspire future studies and also safest to avoid both floor and ceiling effects from exposure frequency. The last four tutor groups therefore received eight tutoring sessions daily with 24 songs per session (schedule 3), reaching daily song exposures of 192 songs and an average of 768 motifs, which falls into the range of daily song output observed in adult males housed socially (range between 0 and 1262 motifs, average ± SD: 395 ± 362 motifs; Jesse and Riebel 2012, range

between 0 and 891 motifs, average ± SD: 237 ± 208 motifs; Böhner 1983). In all schedules and for all treatments, the first session began at 08:15, half an hour after the lights went on in the room and every tutoring session started with the audio-only presentation of three introductory notes of the tutor followed by one second of silence. After this, one of three different videos and/or songs of the same tutor was presented. After the stimulus presentations, the screens went back to black.

Table 1. Description of the different tutoring schedules used in this study.

Schedule	# daily tutoring sessions	Daily tutoring times	#songs/ session	# songs/ day	Inter-song interval	N groups
1	3	8:15, 12:15, 16:15	10	30	fixed, 1 min.	3
2	4	8:15, 10:15, 12:15, 16:15	12	48	variable, range 2-6s[3]	9
3[1]	8	8:15, 8:45, 9:15, 10:15, 12:15, 13:30, 14:45, 16:15	24	192	variable, range 2-6s	4[2]

[1] With this schedule, no birds were tutored in the Audio condition.
[2] All tutor groups had a different tutor song, but these four groups received the songs of 4 of the tutors used in schedule 2.
[3] The playback program used random inter-song intervals in the given range.

Stimulus preparation

Audio and video recordings

Stimuli consisted of audio and video recordings of undirected song of 12 adult male zebra finches from the colony (3 songs per bird, 36 songs in total). All songs were recorded in an identical manner and using the same equipment: 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, during the time of highest singing activity after lights on, the male was recorded between 08:00 and 11:00, or until we had recorded three songs. 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. A single cross-shaped perch was placed in the middle of the cage so that the bird would always be in focus of the camera. The back side of the cage was covered with a black cloth so that the videos had a black background, because this gave the best contrast between the background and the stimulus bird. LED video lights (DV-216VC, FalconEyes, Hong Kong) were projected on the perch from the rear above and the left and right front sides. Audio recordings were made with a Sennheiser MKH40 microphone (Wedemark, Germany),

hanging 50 cm above the perch in the recording cage, 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 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 420 Hz using Praat (version 6.0.19, Boersma & Weenink, 2008). Audio and video files were synchronized with Vegas Pro (version 14.0, Magix, Berlin, Germany).

For each male, three songs with introductory 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 motif repetitions ± SD = 3.9 ± 0.8).

Colour adjustments of the videos

Commercially available RGB displays are made for human vision, and their three phosphors (Red, Green, Blue) match the sensitivity of human cones (560 nm, 530 nm and 420 nm, Solomon and Lennie 2007). Zebra finches, like other birds, are tetrachromatic with four cone types with wavelength sensitivities of 567 nm, 502 nm, 429 nm, and 360-380 nm. Birds thus have a wider visual spectrum (approximately 320–700 nm, incl. UV) than humans (approximately 400–700nm). This means images or videos displayed on standard LCD screens that emulate human perception of colour rather than the true light reflectance of objects, video playbacks on RGB screens will not provide the true colours to the birds. There is however a method known as *colour-realistic imagery* which allows to colour-correct images displayed on RGB screens (Tedore and Johnsen (2017) to match the colour perception system of a non-human observer as closely as possible. To calculate the correction factors we needed as input: the colour spectra of the plumage of zebra finches; the sensitivity of their photoreceptors (measured previously by Bowmaker, Heath, Wilkie, and Hunt (1997)); and the output of the phosphors of the experimental RGB displays. As it is not possible to display UV light with monitors we neglected the UV component and only corrected the red, green and blue channel.

Measurements of zebra finch plumage radiance and video screen irradiance
Most zebra finch colour patches are either black, white or grey and they do not need colour correction (or colour correction would only lead to minimal changes), therefore we focused on the three main coloured patches: the red beak, the orange/red cheeks and the brownish lateral patterns beneath the wings. We measured these patches for 6 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)). We then measured the absolute radiance of the gaming monitors (VG248QE, ASUS, Taipei, Taiwan) to be used to display our stimuli. We used a calibrated light source (HL-3P-CAL) and a 400 um Premium Fiber (QP400-2-VIS-BX), both from Ocean Insight (Orlando, FL, USA) to calibrate the spectrometer. To ensure that the fibre did not move between measurements of the different phosphors we clamped the bare fibre firmly in front of the screens. We displayed red, green or blue phosphors by setting the measured phosphor value to a middle magnitude 128 and all other phosphors to zero. Measured radiance values were converted to quantal units, see Appendix, Figure 8 for the results.

Generation of colour adjusted video stimuli
With the zebra finch plumage colour spectra, the birds' photoreceptor sensitivities and the output of the phosphors of the screens, we could calculate correction factors using a Matlab script (R2019a, Mathworks, Natick, Massachusetts, USA - script provided by Tedore and Johnsen (2017)). We then colour corrected the single frames of the videos in Photoshop CC (Adobe Inc., Mountain View, California, USA) using the 'Replace Color' function (Image>-Adjustments>Replace Color) for the different colour patches. For an example of a colour-corrected frame see Appendix, Figure 9. We selected the patch with the eyedropper tool, adjusted the selection threshold in a way the whole patch was chosen and not many other parts of the bird were selected and then adjusted using the correction factor values for the respective patch. We used Photoshop droplets to batch process all colour patches and frames. We also created pixelated videos using the Photoshop displacement filter (Filter>Distort>Displace) and used random pixels as displacement map (see Appendix, Figure 10). The colour corrected frames were then imported in Vegas Pro software to create a video with 119.88 fps. The frames were placed in chronological order for the Audio-video condition and to avoid any rhythmical visual information, in reversed order for the Audio-pixel condition. The audio file was

then added to the video in Vegas Pro. All generated stimuli were exported as mp4 files (Audio: 448 Kbps, 96 kHz, 16 Bit, AAC, Video: 640 x 480 Progressive, YUV, 50 Mbps). After creating these stimuli, we played them back through the loudspeaker above the experimental arena (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 (see Appendix, Figure 11 for an example).

Experimental arena

The experimental arena consisted of a cage (70 x 60 x 45 cm, see Figure 2) with four sides of wire mesh in the audio-only condition and three sides of wire mesh and one side of black plastic in the other two conditions. A window (20 x 15 cm) was cut out of the plastic and the experimental monitor (VG248QE, ASUS, Taipei, Taiwan) placed directly behind it. To ensure reproducible luminance and colour representation for all screens, we calibrated the screens before every tutoring round. For calibration we used a X-Rite i1 Display Studio (Danaher Corp., Grand Rapids, USA) and the program iProfiler with the following settings: White Point CIE Illuminant D65, Luminace 120cd/m2, Tone Response Curve: sRGB. The screen was connected to an Intel NUC computer (NUC7i3BNK, Intel Corporation, California, USA) which controlled stimulus presentation by a custom-made (by one of us - RS) LabView program with a VLC player plugin. Sound was played-back at 74 dB (Fast, A, re 20 µPa, Voltcraft SL-451, Conrad, Hirschau, Germany) at 30 cm from a loudspeaker (Blaupunkt, CB4500, Hildesheim, Germany) suspended from the ceiling at 50 cm above the cage (directly above the video monitor, see Figure 2). We had decided on this position, because positioning the loudspeaker behind the monitor would have negatively affected the sound quality. Visual stimulation can attract the perceived location of spatially discordant but temporally synchronous auditory stimulation (Chen and Vroomen 2013). This phenomenon, known as spatial ventriloquism, has been demonstrated in species as diverse as humans, frogs, spiders and birds (Narins et al. 2005; Lombardo et al. 2008; Chen and Vroomen 2013; Kozak and Uetz 2016). Little is known about crossmodal integration in zebra finches, but in another bird species, spatial ventriloquism was found to take place over a distance of one meter between the auditory and visual stimulus (Lombardo et al. 2008). The loudspeaker above the cage of the audio-only condition was connected to the computer of the audio-pixel condition. Each cage was placed on a table in a sound attenuated room (125 x 300 x 240 cm). A webcam (Renkforce RF-4805778, Conrad, Hirschau, Germany) was

installed next to the cage to record the tutees' behaviour in the cage.

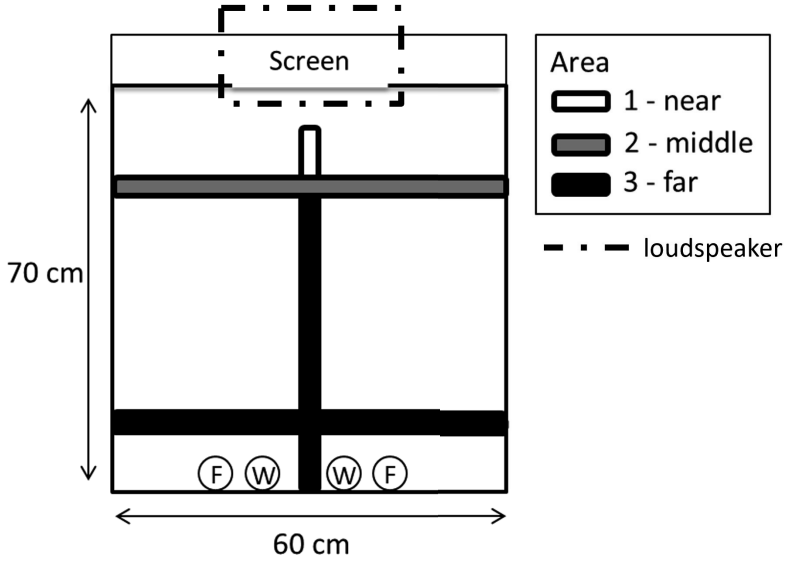

Figure 2. Schematic top view of the experimental set-up. In the set-up for the Audio group, there was no screen next to the cage. For the behaviour observations, we divided the cage into three areas, with 1 being the perch area nearest to the screen (8cm of perch), 2 being an intermediate area (60 cm of perch) and 3 the perch area furthest from the screen (104 cm of perch). The dotted rectangle indicates the location of the loudspeaker (hanging 50 cm above the cage). F = food, W = water. Food and water bottles were placed on the floor of the cage.

Song recordings tutees

All tutees were recorded once as juveniles at 65 dph (X ± SE: 64.6 ± 0.9) and once as young adults after 100 dph (X ± SE: 116 ± 12). For the first recording at 65 days post-hatching, male and female tutees were jointly moved into a cage (76 x 45 x 45 cm) in a sound-attenuated recording room (125 x 300 x 240 cm) between 12:00 and 13:00. A Sennheiser MKH40 microphone (Wedemark, Germany), connected to a TASCAM DR-100MKiii recorder, was hanging at 50 cm above the perch in the recording cage. Recordings were made with a 96 kHz sampling frequency. Recordings were made continuously during the next morning, after which birds were moved back to the experimental set-up. After 100 days post-hatching, male tutees were recorded again using the same recording set-up and the same procedure, but now males were housed singly in the recording room. There were 42 birds that produced more than 20 songs during this recording session. Only song of these birds was used in the song analysis (one tutee from the Audio-video and one tutee from the Audio-pixel treatment did not sing enough).

Song analysis

An overview of all song analysis measures can be found in Table 2. In almost all 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.

Song and motif selection

For all sound analyses 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). First, all songs were cut out of the recording sessions' audio files saving all songs per male into one folder to then randomly select twenty songs from this folder (with custom-written software by Niklas J. Tralles). As mentioned above, 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. This sample was used to calculate linearity and consistency, and to identify a tutee's 'typical' and 'full' motif (a motif was defined as the repeated syllable sequence in a song). The typical motif was defined as the motif encountered most often in the 20 randomly selected songs and the full motif as the motif with the highest number of different syllables. The full motifs were used for the human observer similarity scoring and to determine the total number of syllables in the tutee's repertoire (see below). For each tutee, we labelled different syllables with different letters (see Figure 3). From the 20 songs, we selected a new smaller subsample consisting of 10 out of the 20 randomly selected songs (again using the custom written software making a random selection from each folder). A random number generator (http://www.random.org) was then used to randomly select one motif from each of these ten songs. Using Praat-software, these ten motifs were cut out of the recordings, filtered with a band stop filter from 0 to 420 Hz, and the amplitude was normalized using the 'scale peak' function. Introductory notes that did not occur with every repetition of the motif were not considered to be part of the motif and cut off before proceeding further with the analyses. These ten motifs were used for the SAP and Luscinia similarity and stereotypy scores (see below).

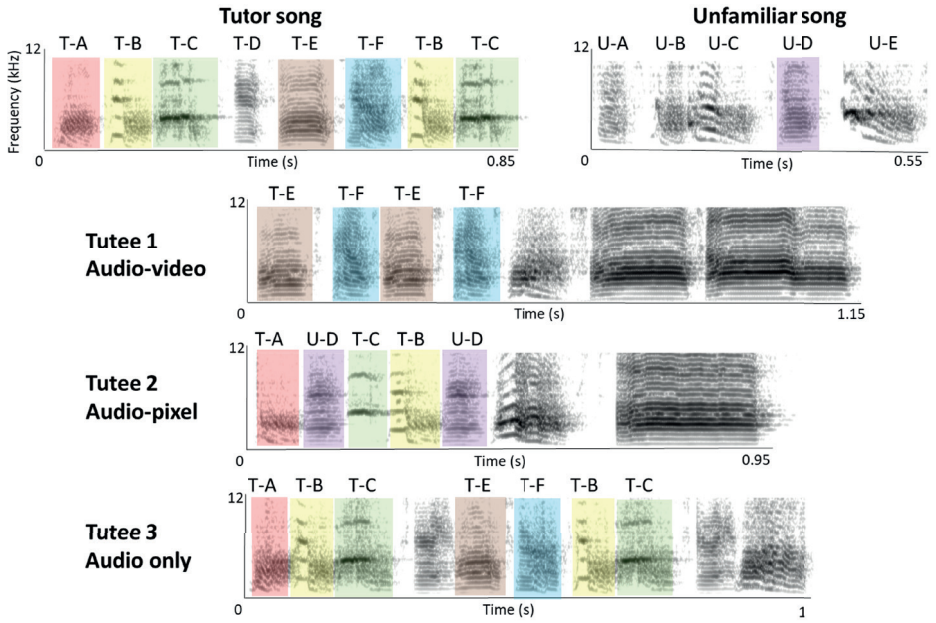

Figure 3. Spectrograms of the full motif of the tutor, the unfamiliar full motif of another adult male and three tutees from one tutor group. Letters above tutor and unfamiliar song spectrograms indicate how syllables were labelled with letters for further analyses. Human observer similarity between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

Song structure and performance

For each tutee, we determined the total number of syllables in the typical motif and the number of unique syllables in the full motif by visually inspecting the spectrograms in Praat (settings as described above). We calculated sequence linearity and sequence consistency (Scharff and Nottebohm 1991) for the twenty randomly selected songs. Sequence linearity was calculated by dividing the number of different syllables (e.g. A, B, C …) by the number of different transitions between syllables (e.g. AB, AC, BC …) in a song. This measure 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.

Similarity between tutee and tutor song

For zebra finch song, the literature up until 1999, including the studies most relevant to this study (Bolhuis et al. 1999; Houx and ten Cate 1999b), mostly used visual inspection of spectrograms by human observers to assess song similarity between tutors and tutees. This is why we also decided to assess song similarity using human observers. Since 2000, automated digital measurement methods, such as Sound Analysis Pro (SAP, Tchernichovski, Nottebohm, Ho, Pesaran, & Mitra 2000, specifically developed to assess zebra finch song learning) and Luscinia (Lachlan, Verhagen, Peters, & ten Cate 2010) have regularly been used. An often-mentioned advantage of automated song similarity assessment is that it compares song objectively. However, human observer similarity scoring is also objective when using observers that are blinded to the origin and/or expected outcome of the spectrogram comparisons, which was the case in this study. Moreover, both of the aforementioned automated comparison methods were validated against comparisons done by human observers (Luscinia: Lachlan et al. 2010; SAP: Tchernichovski et al. 2000), which the developer of SAP considers preferred over automated methods (Tchernichovski 2011). In this study, we will therefore primarily use similarity scoring by human observers to assess song learning success in the birds from the different treatments. However, to allow cross-study comparisons, we also assessed song similarity using Luscinia and Sound Analysis Pro (for details see below). We calculated the correlation between the similarity scores obtained with the three different methods to find out whether they provide a similar outcome.

For the human ratings of similarity, we followed the methods used by Houx and ten Cate (1999a), but compared motifs at the syllable level (continuous sounds separated by at least 5 ms of silence), while Houx and ten Cate (1999a) compared motifs at the element level (sounds separated from other sounds by either an observable gap of silence on the spectrogram or by an abrupt change in frequency or structure, meaning that one syllable can consist of several elements). Based on previous studies, we expected poor song copying in the Audio tutees (Price 1979; Eales 1989) and depending on whether videos would or would not sufficiently substitute for live tutors potentially in the other treatment groups too. In poorly copied and isolate-like song, determining element boundaries can be difficult, for instance due to a higher variance in frequency patterns than in normal song (Price 1979) while determining syllable boundaries is more straightforward. For this reason, we decided to assess similarity on the syllable level. For visual scoring, a PowerPoint presentation was created where each slide contained two spectrograms: on top the full motif of the tutee (labelled 'tutee') and below a second spectrogram labelled 'model'. The model

song was either from the tutor or from the tutor of another tutor group (unfamiliar to the tutee). Each tutee was thus compared with two models: the actual tutor and an unfamiliar control model (the tutor of another group). We included the unfamiliar song to analyse the level of syllable sharing between two birds from the same colony that occurs by chance. Syllables were labelled with different letters by one of us (JV) and these letters were placed below each of the two spectrograms on each slide. Three independent observers (PhD-candidates at the Leiden lab not involved in this study and with varying experience in working with spectrograms of zebra finch song) received the PowerPoint presentation. For each syllable in the tutee's repertoire, the observers were asked to indicate which syllable of the model it resembled most by paying attention to frequency pattern, duration, shape and position with respect to neighbouring syllables, 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'). Observers were given no information on tutees' treatment groups and whether a model song was from the tutor or from another male. To assess inter-observer reliability, we first normalized the scores per observer (for each score we subtracted the mean of all scores of this observer and then divided it by the standard deviation of the scores of the observer). We calculated repeatability using a one-way ANOVA (following Lessells & Boag 1987) with the similarity score as the dependent variable and tutee ID as factor. The repeatability estimate r of the normalized scores was moderate (Tutor-Tutee: r ± SE = 0.54 ± 0.09, $F_{2,39} = 4.45$, $p < 0.01$; Tutee-Tutor: r ± SE = 0.50 ± 0.09 $F_{2,39} = 4.03$, $p < 0.01$). The difference between observers mainly had to do with how strict observers were regarding poorly copied syllables. To capture this best and to have one value for further analyses that would integrate all observer values, we decided to work with the total sums of similarity scores (of all three observers) for a tutee divided by the potential maximum score a bird could receive from three observers (the sum of the similarity scores of all three observers for all pairwise syllable comparisons of a particular model-tutee comparison). This score thus corrected for between individual differences in syllable numbers, thereby providing a measure combining the proportion of syllables copied as well as a weighing of their similarity.

Syllable sharing and similarity values are affected by the direction of such a comparison if model and tutee differ in total number of syllables and therefore can be assessed 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"). The tutee-model comparison was included as tutees can

differ in how many syllables they improvise in addition to song copied from a tutor (Williams, 1990). To clarify, a tutee that has accurately copied the syllables ABC from a tutor with the song ABCDE would get a higher score for the tutee-model comparison than for the model-tutee comparison. A tutee that sings ABCDEFG (with ABCDE accurately copied from the tutor and F and G improvised) would get a higher score for the model-tutee comparison than for the tutee-model comparison. 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. If two or more tutee syllables received the same similarity score, we noted this score once, but the scores for all tutee syllables were included in the tutee-model comparison. For each motif, the scores of all three observers were then summed up and divided by the maximum possible score (see Table 2 for full formula).

For the automatic, quantitative song comparisons, we compared each of 10 randomly selected motifs of a tutee to each of 10 randomly selected motifs of its tutor using both Luscinia (version 2.16.10.29.01) 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). A difference between the two methods is that SAP uses a linear time warping algorithm to align two signals for comparison, while Luscinia uses dynamic time warping (DTW) which searches for the optimal alignment of two time series irrespective of how warped they have been in time (Lachlan et al., 2010). Similarity assessment in Sound Analysis Pro is based on five acoustic features: pitch, frequency modulation, amplitude modulation, goodness of pitch and Wiener entropy. Like with the human observer similarity scores, SAP similarity scores are influenced by the direction of the comparison. 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 that are observed in the tutor's song. We used the median value of these scores as the quantitative measure of similarity (henceforth 'SAP similarity score'), as our sample size of birds was too small to create a good-fitting model for the similarity scores of all comparisons and as the SAP scores were not normally distributed and bound between 0 and 100. Luscinia also calculates global similarity but works with a dynamic time-warping algorithm to calculate acoustic distance scores between tutee-model pairs. We chose the acoustic features 'mean frequency', 'fundamental frequency' and 'fundamental frequency change' for the acoustic distance

calculations (following Lachlan, van Heijningen, ter Haar, & ten Cate 2016). We also included 'time' in the analysis, which allows for flexible comparison of signals that vary in length. The output of the DTW analysis is a distance measure between 0 and 1 for all possible pairs of motifs. Unlike 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 and to get an indication of how similar the 10 randomly selected tutee motifs were to each other, we also compared the 10 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 10 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 10 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'.

Table 2. Overview of song analysis parameters used in this study and the sample that was used to calculate them.

Parameter	**Definition**	**Sample per bird used to calculate the parameter**
Typical motif	most frequently produced motif	20 random songs
Full motif	motif with highest # different syllables in bird's repertoire	20 random songs
Total number of syllables	# syllables in a tutee's typical motif	Typical motif
Number of unique syllables	# unique syllables in a tutee's full motif	Full motif
Linearity	(# different syllables/song)/(#transition types/song)	20 random songs
Consistency	(total # typical transitions)/(total # of transitions)	20 random songs

Human observer similarity score model-tutee	(Σ similarity scores for all model syllables)/(# model syllables*3 (max score)*# observers)	Full motif
Human observer similarity score tutee-model	(Σ similarity scores for all tutee syllables)/(# tutee syllables*3 (max score)*# observers)	Full motif
SAP similarity score tutor-tutee	SAP similarity scores comparing tutors' to tutees' motifs	10 random motifs
SAP similarity score tutee-tutor	SAP similarity scores comparing tutees' to tutors' motifs	10 random motifs
Luscinia similarity score	1 – Luscinia distance score for comparison of tutor and tutee motifs	10 random motifs
SAP stereotypy score	SAP similarity scores for the comparison between tutee motifs	10 random motifs
Luscinia stereotypy score	1 – Luscinia distance scores for the comparison between tutee motifs	10 random motifs

Behaviour recording and analysis

For the 30 days of tutoring, daily web-cam (Renkforce RF-4805778, Conrad, Hirschau, Germany) recordings were made of the tutoring sessions at 8:15, 12:15 and 16:15. For 6 tutor groups (18 male-female tutee dyads) that were tutored with tutoring schedule 2 (see Table 1), videos from every 5th day were coded using BORIS software (version 7.5.1). Coding was done by two of us (IvH and RJ) that first scored the same video's independently until they reached an inter-observer reliability value of $K > 0.9$ (Cohen's Kappa calculated by BORIS). After this, they each coded different videos (N.B. for these videos observer blinding was not possible, as filming and scoring the approach towards the stimuli showed the stimuli. However, observer biases are playing out strongest with ambiguous or continuous categories, but less so for discrete units such as these spatially separated perches). The observers scored the position of the tutees in the different areas of the cage during stimulus presentation (see Figure 3). This was used to calculate the proportion of the observed time that tutees spent in the different areas corrected for perch length in each area ((time

spent$_{area\ x}$/length perch$_{area\ x}$)/(total time/total cm perch length)). In addition, we also scored the amount of times the birds left the perches to fly directly up and against the screen. For the Audio condition, the amount of times the tutees flew up and against the location of the screen was scored, even though the Audio birds did not have a screen next to their cage.

Statistical analysis

RStudio (R: version 3.5.1) was used for all statistical analyses. To assess tutee engagement with the stimuli, the proportion of time spent in different cage areas (corrected for perch length in that area) was arcsine square root transformed before analyses to meet model assumptions. We then created linear mixed models (LMMs, package *lme4*: Bates, Mächler, Bolker, & Walker 2014) and started with a null model that only included 'TutorGroup' (Number of the tutor group) as a random factor. We then added fixed effects in the following order: 'area' (1, 2 or 3), 'treatment' (Audio-video, Audio-pixel or Audio), the interaction between 'area' and 'treatment', and 'sex' (sex of the tutee: male or female). We used ANOVA's to check whether each of these fixed effects led to a significant improvement of the model. For the number of screen approaches, we created negative binomial generalized linear mixed models (GLMMs). We started with a null model with only 'TutorGroup' as random factor, then added fixed effects in the following order: 'Treatment', 'Sex' and 'Tutoring day' (number of days since the tutee was moved to the experimental set-up) and used an ANOVA to test whether these factors significantly improved the model.
For the stereotypy and human observer, SAP and Luscinia similarity scores, we built linear mixed-effects models (LMMs). Human observer, SAP and Luscinia scores were arcsine square root transformed before analyses to meet model assumptions. To calculate the correlation between the three different similarity scores (human observers, SAP and Luscinia), we calculated the Pearson correlation coefficient after a square root transformation of the human observer scores to meet assumptions of normality. Generalized linear mixed-effect models (GLMMs) with a Poisson distribution and log-link function were created for the total number of (unique) syllables. For the analysis of all song parameters, we started with a null model with only 'TutorGroup' (ID of the tutor group) as a random factor. We then added 'Schedule' (the 3 different tutoring schedules) and 'Treatment' as fixed effects. We used ANOVA's to test whether adding each of these model terms led to a significant improvement compared to the simpler model. As the human observer similarity scores were our main parameter of interest for assessing song learning success and we were interested in the similarity scores attained by the tutees from the different tutoring treatment groups, we still ran a model with 'Treatment' as fixed factor for the human observer

similarity scores even if this did not significantly improve the model. 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 'model' (tutor or unfamiliar)' as fixed factor significantly improved the null models (with 'TutorGroup' 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

Tutee behaviour

During the tutoring sessions, birds did not use all areas in the cage equally often (Figure 5). Birds in all groups showed a bias towards area 1 which was closest to where the stimuli could be seen and heard. To test whether this engagement with the stimuli differed across treatments, we analysed the proportion of time during the tutoring sessions that the tutees spent in the different areas of the cage corrected for the perch length in that area. The proportion of time spent was affected by area, treatment and the interaction between area and treatment: tutees spent a significantly higher proportion of time in area 1 (near) in the Audio-video group than in the Audio-pixel and Audio group. Besides, in the Audio-video and Audio-pixel group, more time was spent in area 1 (near) than in area 2 (middle), while this difference was not found in the Audio group (best model included 'treatment', 'area' and the interaction between 'treatment' and 'area', see Table 3 and Figure 4).

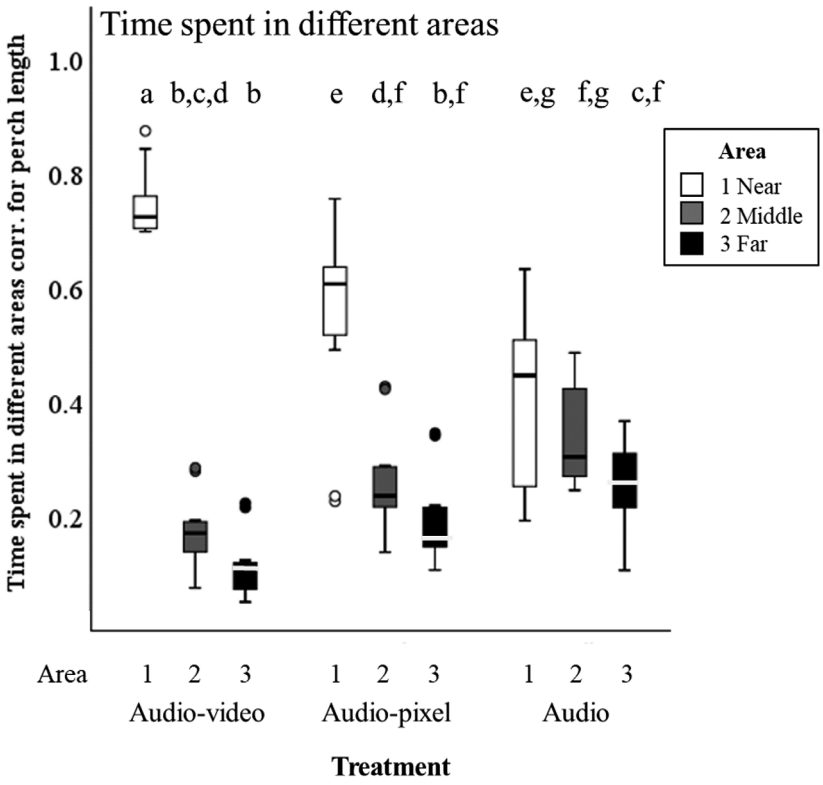

Figure 4. Proportion of time spent in the different cage areas, corrected for the total perch length in that area. Box plots indicate the median (mid-line), interquartile range (box), and 1.5 times the interquartile range (whiskers). Data points beyond this range are plotted as individual points. Different letters above boxes indicate a significant difference of $p < 0.05$ according to post-hoc tests (see Appendix, Table 11), LMM see Table 3.

Table 3. Details of best model (LMM) for the proportion of time spent in different areas of the cage, corrected for the perch length in that area.

Response variable[1]	***Model term***	***Level***	***Estimate***	***SE***	***t***
Prop. of time spent corrected for perch length	Intercept		0.69	0.03	20.49
	Treatment				
		Audio-video	0.32	0.05	6.72
		Audio-pixel	0.14	0.05	2.95
	Location				
		Area 2 (middle)	-0.07	0.05	-1.56
		Area 3 (far)	-0.17	0.05	-3.61

Location x Treatment				
	Area 2 x Audio-video	-0.51	0.07	-7.61
	Area 3 x Audio-video	-0.50	0.07	-7.48
	Area 2 x Audio-pixel	-0.23	0.07	-3.34
	Area 3 x Audio-pixel	-0.21	0.07	-3.16

[1]LMM with random factor 'Tutor group'. For post-hoc comparisons see Appendix, Table 11.

The amount of times that the tutees flew up to the screen (or the location of the screen in the Audio group) differed between the treatment groups: there were more direct screen approaches in the Audio-video condition than in the Audio-pixel and Audio condition, and more screen approaches in the Audio-pixel than in the Audio condition (model including 'treatment' significantly better than model without treatment, N = 36, $\chi^2 = 40.62$, $p < 0.01$, see Table 4 (also for post-hoc test results) and Figure 5). The number of direct screen approaches did not differ between the male and female tutee (adding 'sex' did not significantly improve the model, N = 36, $\chi^2 = 0.73$, $p = 0.39$) and did not change over time (adding 'Tutoring day' also did not significantly improve the model, N = 36, $\chi^2 = 0.12$, $p = 0.73$).

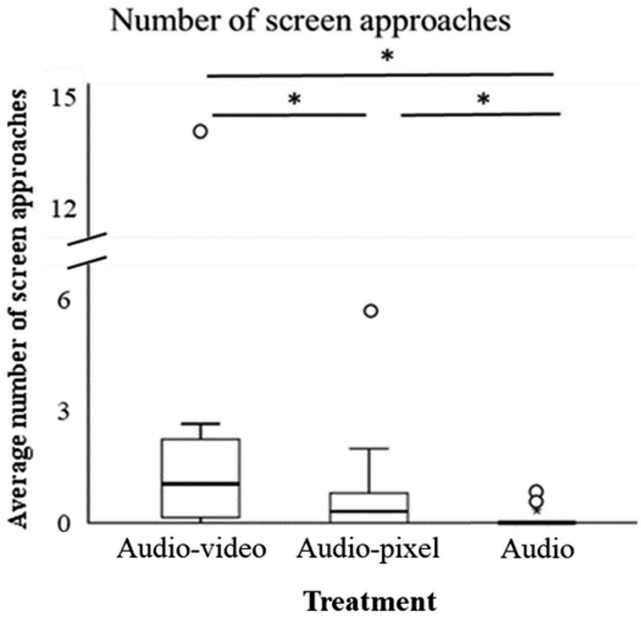

Figure 5. The average number of direct screen approaches during the stimulus presentations (values are the average per tutee for the three scored presentations per recording day (every fifth day of the tutoring period three (out of four) tutoring sessions were recorded and scored)). * indicates $p < 0.05$, GLMM see Table 4.

Table 4. Details of best model (GLMM) for the amount of screen approaches. Significant p-values are given in bold.

Response variable[1]	***Model term***	***Level***	***Estimate***	***SE***	***z***	***p***
Number of screen approaches	Intercept		-4.15	0.70	-5.89	**<0.01**
	Treatment					
		Audio-video	3.46	0.66	5.21	**<0.01**
		Audio-pixel	2.45	0.70	3.51	**<0.01**

[1] Negative binomial GLMM with random factor 'Tutor group'. Significant post-hoc comparisons: Audio vs. Audio-video: estimate: -3.46, SE: 0.66, z: -5.21, $p < 0.01$, Audio vs. Audio-pixel: estimate: -2.45, SE: 0.70, z: -3.51, $p < 0.01$, Audio-video vs. Audio-pixel: estimate: 1.02, SE: 0.38, z: 2.67, $p < 0.05$.

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 5). Presentation schedule affected none of the parameters but linearity, which differed between the 3 tutoring schedules and was higher in schedule 1 (fewer presentations) than in the other schedules (see Table 6C, model including 'schedule' significantly better than null model, N = 42, $\chi^2 = 8.80$, $p = 0.01$, best models for each parameter in Table 6).

Table 5. Mean values of song structure and performance parameters and details on ANOVA for comparison between null model and model including 'treatment' as a fixed effect. 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).

	Tutor (not in models)	**Audio-video**	**Audio-pixel**	**Audio**	**ANOVA null model and model with 'treatment'**		
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	N	$\chi 2$	p
Total nr syllables	*6.33 ± 1.44*	5.08 ± 1.38	6.46 ± 1.76	5.25 ± 2.34	42	2.56	0.28
Nr unique syllables	*5.25 ± 1.60*	4.60 ± 1.30	4.93 ± 1.44	4.42 ± 0.51	42	0.40	0.82
Linearity	*0.46 ± 0.12*	0.41 ± 0.11	0.40 ± 0.10	0.44 ± 0.09	42	0.85	0.66
Consistency	*0.94 ± 0.04*	0.89 ± 0.08	0.90 ± 0.07	0.92 ± 0.08	42	0.77	0.68

Table 6. Details of best models for the song structure and performance parameters.

Response variable	*Model term*	*Level*	*Estimate*	*SE*	*z*
A. Total number of syllables[1]	Intercept		1.72	0.07	25.16
B. Number of unique syllables[1]	Intercept		1.54	0.07	21.57
C. Linearity[2]	Intercept		0.51	0.04	14.49
	Schedule				
		Schedule 2	-0.12	0.04	-3.04
		Schedule 3	-0.12	0.05	-2.47
D. Consistency[2]	Intercept		0.90	0.02	49.34

[1] GLMM with a Poisson distribution and random factor 'Tutor group'.
[2] LMM with random factor 'Tutor group'.

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 human observer and the SAP similarity score or the SAP and the Luscinia similarity score (see Table 7), suggesting that these measures pick up on different dimensions of song similarity. It is important to note, however, that the correlation between the human observer similarity scores on the one hand, and the SAP and Luscinia scores on the other hand is influenced by the different samples that were used to calculate these scores (1 typical motif for the human observer scores and 10 randomly selected motifs per tutee for the SAP and Luscinia scores). In subsequent paragraphs, we will present the results of all three methods, although, as mentioned before, we will primarily focus on the results from the human observer similarity scoring to determine whether song learning success was affected by the different tutoring treatments.

Table 7. 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 values are given in bold.

Comparison	N	r	p
human observer sim. score – SAP sim. score	42	0.04	0.98
human observer sim. score – Luscinia sim. score	42	**0.57**	**< 0.01**
SAP sim. score - Luscinia sim. score	42	0.14	0.44

Similarity scores for the comparison between tutor and tutee songs
To find out whether the tutees had learned from the tutor, we checked whether their song was more similar to the tutor song than to an unfamiliar song. The human observer similarity scores for the tutor to tutee and tutee to tutor comparisons were significantly higher than the similarity scores for the comparisons with an unfamiliar song (model with 'model (tutor or unfamiliar)' was significantly better than null model, model to tutee comparison: N = 42, χ^2 = 5.39, p = 0.02, Table 8A, tutee to model comparison: N = 42, χ^2 = 4.75, p = 0.03, Table 8B). As this means that tutees' songs were more similar to their tutor's song than would be expected by random sharing in the colony, we assume that the tutees learned at least some aspects from their tutors. For all subsequent analyses, we proceed with comparisons between tutor and tutees only.

Table 8. Details of best models for the arcsine square-root transformed human observer similarity scores for the comparison of the model songs to the tutee songs (A) and the tutee songs to the model songs (B).

Human observer similarity scores

Response variable	***Model term***	***Level***	***Estimate***	***SE***	***t***
A. Model-tutee[1]	Intercept		0.52	0.02	21.63
	Model				
		Unfamiliar	-0.08	0.03	-2.34
B. Tutee-model[1]	Intercept		0.57	0.02	23.41
	Model				
		Unfamiliar	-0.08	0.03	-2.18

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

In the comparison of the syllables in the tutor's repertoire to those in the tutee's repertoire (tutor-tutee comparison), the human observer similarity scores differed between the treatment groups: these scores were higher in the Audio group than in the Audio-video group (model including 'treatment' was significantly better than null model, N = 42, χ^2 = 6.60, p = 0.04, see Table 9A (also

for post-hoc test results) and Figure 6). The tutor-tutee similarity scores did not differ between the tutoring schedules (model including 'schedule' was not significantly better than null model, N = 42, $\chi^2 = 3.34$, p = 0.19). In the comparison of the syllables in the tutee's repertoire to those in the tutor's repertoire (tutee-tutor comparison), human observer similarity scores were also highest in the Audio group (see Table 9A), but these similarity scores were not significantly affected by the different tutoring treatments (adding 'treatment' as fixed factor did not significantly improve the null model (N = 42, $\chi^2 = 4.72$, p = 0.09)). The tutee-tutor similarity scores also did not differ between the tutoring schedules (adding 'schedule' did not significantly improve the null model (N = 42, $\chi^2 = 2.27$, p = 0.32)).

The SAP similarity scores for the comparison of the tutor song to the tutee song (SAP similarity scores tutor-tutee) differed between the treatment groups and did not differ between the tutoring schedules: the tutor-tutee similarity scores were higher in the Audio-Pixel group than in the Audio group (model with 'schedule' was not significantly better than null model: N = 42, $\chi^2 = 2.89$, p = 0.24, while model with 'treatment' was significantly better than null model: N = 42, $\chi^2 = 8.73$, p = 0.01, see Table 9B (also for post-hoc test results) and Figure 6C). For the comparison of the tutee's songs with their tutor's song, the Sound Analysis Pro similarity scores (SAP similarity score tutee-tutor) did not differ between the tutoring schedules or the tutoring treatments (model with 'schedule' was not significantly better than null model: N = 42, $\chi^2 = 0.38$, p = 0.83, model with 'treatment' was not significantly better than null model: N = 42, $\chi^2 = 1.12$, p = 0.57, see Table 9B for best model).

The different treatment conditions affected the Luscinia similarity scores, but the post-hoc test did not detect any significant differences between two treatment groups (model including 'treatment' was significantly better than the null model, N = 42, $\chi^2 = 6.46$, p = 0.04, see Table 9C and Figure 6). Luscinia similarity scores were not affected by the different tutoring schedules (model including 'schedule' was not significantly better than the null model, N = 42, $\chi^2 = 0.89$, p = 0.64).

Overall, the similarity between tutor and tutee song was highest for the Audio tutees for all methods and comparisons, except for the SAP similarity scores for the tutor-tutee comparison (see Table 9 and Figure 6). For this comparison, similarity scores were highest in the Audio-pixel group.

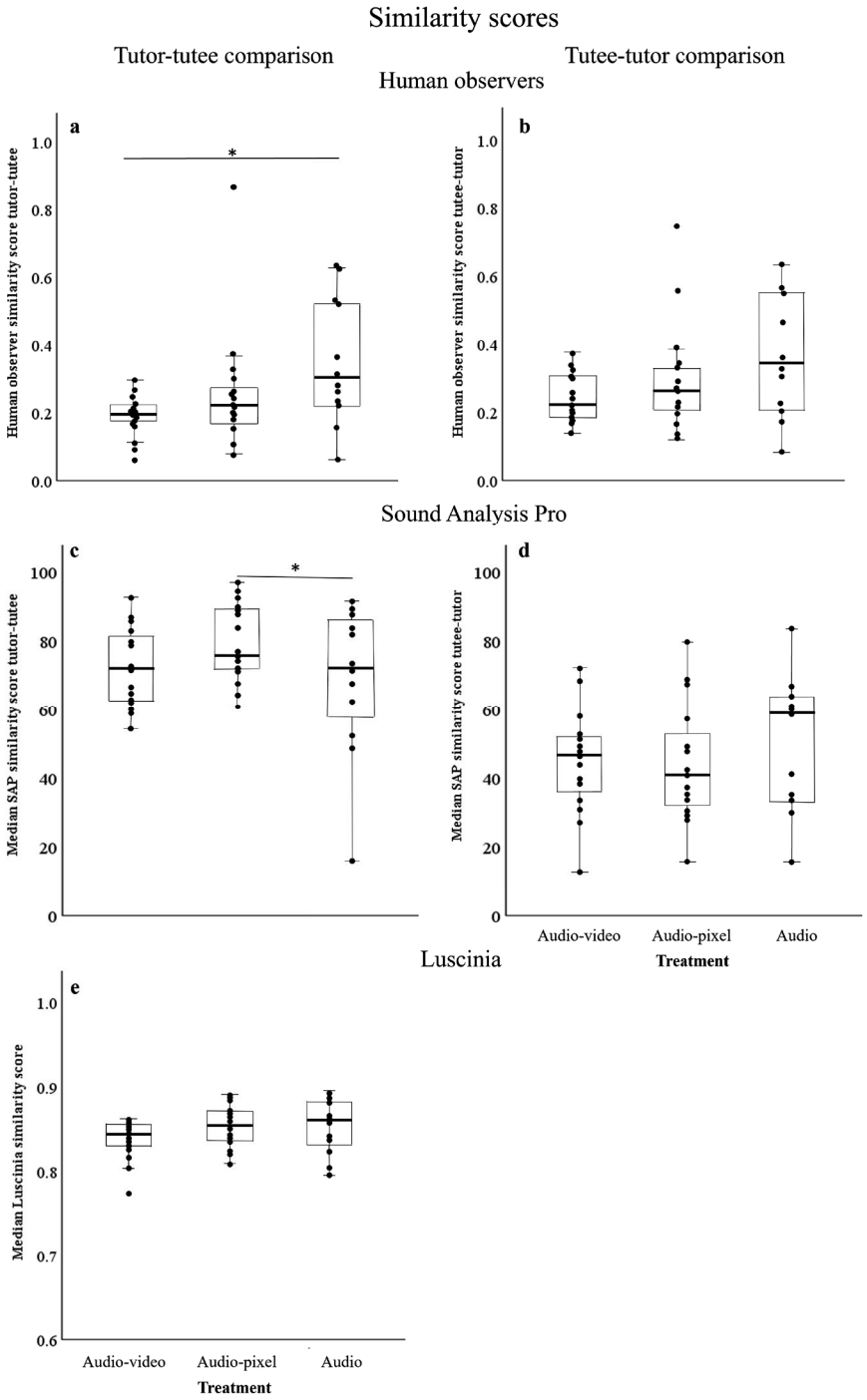

Figure 6. Graph showing the human observer similarity score for the tutor-tutee (a) and the

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). * indicates $p < 0.05$, LMMs see Table 9. NB Human observer and SAP similarity scores calculate how much of one signal can be found in another signal. Therefore, when comparing two signals, two different comparisons can be made (what proportion of the tutor motif is found in the tutee motif (tutor-tutee) and what proportion of the tutee motif is found in the tutor motif (tutee-tutor)). Luscinia does not calculate how much of one signal can be found in another signal, but calculates how dissimilar two signals are.

Table 9. Details of models with 'Treatment' as fixed factor for the arcsine square root transformed human observer similarity scores (A) and the best models for the arcsine square root transformed SAP (B) and Luscinia (C) similarity scores.

			Tutor-tutee			***Tutee-tutor***		
Response variable	***Model term***	***Level***	***Estim.***	***SE***	***t***	***Estim.***	***SE***	***t***
A Human observers sim. scores[1]	Intercept		0.62	0.05	12.18	0.64	0.05	14.05
	Treatment							
		Audio-video	-0.17	0.07	-2.58	-0.13	0.06	-2.16
		Audio-pixel	-0.10	0.07	-1.48	-0.07	0.06	-1.18
B SAP sim. scores[2]	Intercept		1.00	0.05	18.59	1.07	0.04	27.07
	Treatment							
		Audio-video	0.06	0.05	1.01			
		Audio-pixel	0.16	0.05	3.01			
C Luscinia sim. scores[3]	Intercept		1.19	0.01	109.69			
	Treatment							
		Audio-video	-0.024	0.01	-2.15			
		Audio-pixel	-0.001	0.01	-0.07			

[1] LMMs with random factor 'Tutor group'. Significant post-hoc comparison tutor-tutee: Audio vs. Audio-video: estimate: 0.17, SE: 0.07, t: 2.56, $p = 0.04$.
[2] LMMs with random factor 'Tutor group'. Significant post-hoc comparison tutor-tutee: Audio vs. Audio-pixel: estimate: -0.16, SE: 0.06, t: -2.99, $p = 0.02$. For the tutee-tutor comparison, 'treatment' was not included in the best model.
[3] LMMs with random factor 'Tutor group'.

SAP and Luscinia stereotypy scores

To test whether birds from the different treatments differed in how stereotyped they produced their motifs, we compared the 10 randomly selected tutee motifs to each other in SAP and Luscinia. There was no difference between the tutees from the different treatment groups in the SAP or Luscinia stereotypy scores

(model including 'treatment' was not significantly better than null model for the SAP stereotypy score (N = 42, $\chi^2 = 4.36$, p = 0.11, see Figure 7A, Table 10A) or the Luscinia similarity score (N = 42, $\chi^2 = 1.37$, p = 0.50, see Figure 7B, Table 10B). There was no difference between the birds raised with the different tutor song presentation schedules in the Luscinia stereotypy scores (model including 'schedule' was not significantly better than null model for these scores, N = 42, $\chi^2 = 2.99$, p = 0.22), but the schedules did affect the SAP stereotypy scores (model including 'schedule' was significantly better than null model, N = 42, $\chi^2 = 14.14$, p < 0.01). SAP stereotypy scores were higher for schedule 1 than for schedule 2 and 3.

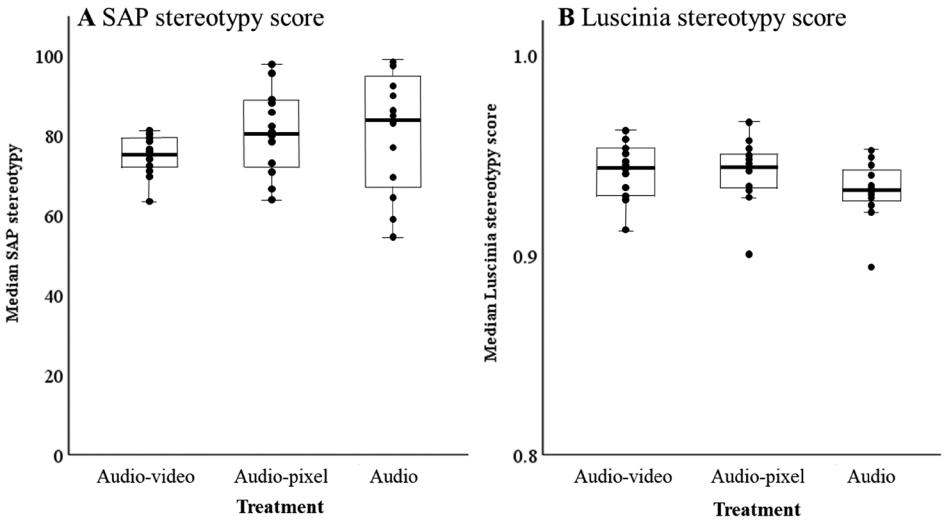

Figure 7. (a) SAP and (b) Luscinia stereotypy scores for the 10 randomly selected tutee motifs.

Table 10. Details of best models for the (arcsine square root transformed) SAP (A) and Luscinia (B) stereotypy scores.

Response variable[1]	***Model term***	***Level***	***Estimate***	***SE***	***t***
A. SAP stereotypy score	Intercept		1.23	0.04	30.15
	Schedule				
		Schedule 2	-0.15	0.05	-3.30
		Schedule 3	-0.23	0.06	-4.09

B. Luscinia stereotypy score	Intercept	1.32	0.005	249.4

[1] LMMs with random factor 'Tutor group'.

Conclusions and discussion

Multimodality can enhance stimulus salience, for instance because of an alerting function of one of its components or because components in different modalities interact and affect how they are perceived (Chen & Vroomen 2013; Feenders, Kato, Borzeszkowski, & Klump 2017; Partan & Marler 1999; Rowe 1999). Visual speech and song production cues alike might facilitate vocal learning (Kuhl and Meltzoff 1982; Slater et al. 1988; Beecher and Burt 2004; Teinonen et al. 2008; Derégnaucourt 2011; Lewkowicz and Hansen-Tift 2012; Tenenbaum et al. 2015). The aim of this study was to test whether visual exposure to a singing tutor through a high-quality video coupled with audio playback of the song has a facilitating effect on zebra finch song learning. Birds were tutored in three different conditions; audio only, audio with a video of the tutor and audio with a pixelated and reversed video. Song learning success was assessed when the juveniles had reached adulthood, using human observer visual spectrogram scoring and two automated song similarity assessment methods. We hypothesized that an auditory stimulus with concurrent visual stimulation would improve song learning compared to a unimodal auditory stimulus. Behavioural observations of the young birds showed that their engagement with the stimuli was highest in the condition where song presentation was combined with a tutor video. However, when looking at the learning outcomes, contrary to our expectations, the colour realistic video of a singing conspecific, albeit the most attractive stimulus for the tutees, did not show improved song learning compared to the birds that received audio-only playback in any of the song similarity assessment methods.

Our prediction that visual exposure to a singing tutor improves vocal learning arose from empirical and theoretical evidence in the literature (van Kampen and Bolhuis 1991, 1993; Adret 1992; Hultsch et al. 1999; Rowe 1999). The puzzling results found in this study raise two possibilities – either our design or our assumptions were inappropriate. We will first discuss which methodological confounds can be excluded and then the wider implications of these findings regarding video tutoring.

Could it be that song learning success in this study was not affected by the visual stimulus due to the video being of insufficient video quality? Owing to technical and theoretical advancements, our study improved on potential

technical shortfalls of earlier video tutoring studies such as unrealistic colours, too slow refresh rates or poor sound quality that have been a worry for animal studies in general (Oliveira et al. 2000; Ware et al. 2015; Chouinard-Thuly et al. 2017), and for an earlier video tutoring study in this species (Adret, 1997). Here, we adapted our videos to the specific colour vision and flicker fusion frequency of the zebra finch visual system, using colour realistic imagery (Tedore and Johnsen 2017), high-speed cameras and monitors with high refresh rates. However, while this meant state-of-the-art stimulus preparation, video recordings and playbacks (other than high quality audio playbacks) run risk of artefacts as they are not playbacks of the original stimuli, but only emulate those stimulus properties triggering the percepts associated with particular stimuli. Besides, even though we used the highest current standards, there could still be other video properties, such as deviations from real birds' visual appearances in brightness, interference from electromagnetic fields (Pinzon-Rodriguez and Muheim, 2017) or artefacts arising from the conditions during filming the singing tutors (e.g. the choice of background colour or filming the singing tutors through a layer of Plexiglas). It is also possible that the distance between the screen and the loudspeaker affected whether the birds perceived the auditory and visual stimulation as originating from the same location, which might have negatively affected potential facilitating effects of the visual stimulation on vocal learning. Any of the above or other reasons unknown to us, might have negatively affected the birds' acceptance of the videos as a conspecific tutor. However, the behavioural data show that the birds were attracted to the videos and that they did discriminate the animated conspecific from the pixelated abstract animation: during song presentations tutees spent substantially more time close to the stimulus showing the singing male than the video showing the same bird animation but pixelated and reversed. Tutees not only used the perch near the video with the singing male more than the other perches, but they also actively flew more to the screen than tutees exposed to the pixelated video. In this context, it is important to note that the pixelated video differed from the normal tutor video in at least two aspects: the pixels were randomized and the frames were presented in reversed order. We therefore cannot tell whether the difference in tutee behaviour in response to the pixelated compared to the normal tutor videos resulted from the lack of synchrony between auditory and visual stimulation or from the lack of seeing a conspecific bird on the screen in the pixelated videos. Without being able to pin down the exact mechanism, we can state from the behavioural data that the tutor video was attractive to the birds and that they were interested in it. These observations also suggest that pairing an interesting moving visual stimulus with auditory song exposure does not necessarily lead to improved song learning. A similar observation was

made by Houx and ten Cate (1999): zebra finch tutees spent more time on the perch next to a visual stimulus in form of a taxidermic mount of an adult male zebra finch during than before its exposure. The visual stimulus, however, did not affect song learning success.

Song exposure frequency remains another debated influence on song learning (Chen et al. 2016; Derégnaucourt et al. 2013; Tchernichovski et al. 1999). In our experiment, exposure frequency varied between the different schedules used for different tutor groups, but it was always the same for the three treatments within one tutor group. This therefore seems unlikely to have systematically biased the outcome concerning the differences between treatment groups unless a ceiling or floor effect had masked treatment effects. This does not seem very probable given that there were three different tutor song presentation schedules with pronounced differences in song exposure frequencies. These ranged from 30 to 192 tutor song presentations daily which is comparable with previous playback studies where some have used comparably low song exposure frequencies and still showed some learning from the song playback (20 songs daily: Derégnaucourt et al. 2013; Funabiki & Funabiki 2009 and 40 songs daily in the operant playback study first reporting a potential negative effect of overexposure: Tchernichovski et al. 1999). Besides, the similarity scores obtained by all three similarity assessment methods did not differ between the tutoring schedules. Only two song parameters (sequence linearity and stereotypy assessed by Sound Analysis Pro) differed between the tutoring schedules. These two parameters are both related to how stereotyped a tutee produces its motifs and were lower in the schedules with more daily song exposure. This finding might support the hypothesis that a low song exposure frequency can have positive effects on song learning outcomes in zebra finch tutees (Chen et al. 2016; Tchernichovski et al. 1999).

It is always possible that our song analysis methods did not pick up any subtle difference in song learning. However, because we wanted to be able to compare our data with old and recent song learning studies, we used the three most common and established similarity assessment methods: human observers, SAP and Luscinia (and to the best of our knowledge, these three methods have not previously been used on the same data set). The overall main result that the audio-visually tutored birds did not show improved song learning was the same for all three methods. Perhaps not surprisingly, given the differences in how bioacoustic parameters are weighed in the different approaches, the three methods differed in which between group differences they detected. Most likely, the different algorithms used by the automated methods for calculating

similarity picked up different parameters of song similarity than human observers assessing visual representations of the sounds. Owing to human visual perceptions principles, humans will have recognised shared patterns rather than single parameters. We used ten randomly selected motifs per tutee to calculate similarity with the automated methods SAP and Luscinia, but used only one full motif per tutee for the human observer method, which might explain why we here found a lower correlation between each automated method and the human observers than has previously been found (Luscinia: Lachlan et al. 2010; SAP: Tchernichovski et al. 2000). However, we also found a low correlation between SAP and Luscinia although these scores were based on exactly the same 10 motifs per individual. The differences between the three methods clearly deserve further attention. Note, however, that both automated methods were validated using visual scoring by human observers and that visual scoring is considered an objective suitable method for assessing song similarity as long as multiple independent observers blind to the expected outcome of the comparisons are used as judges (Jones, ten Cate, & Bijleveld 2001). Regarding the test of our main hypothesis that audio-visual exposure should improve song learning, the similarity scores of all three methods did not show such an effect: they were never significantly higher in the Audio-video group than in the Audio-pixel or Audio group despite the higher engagement the tutees showed with these stimuli.

A possible interpretation of these findings is therefore that multimodal stimulus presentation might increase tutee's attention during presentation, but might not affect zebra finch song learning success. Previous studies have, however, demonstrated increased learning of an audio signal in birds when it was paired with visual stimulation (Hultsch et al. 1999; van Kampen & Bolhuis 1991, 1993), despite the use of a less naturalistic visual stimulus than in our study and several earlier ones (Bolhuis et al. 1999; Houx and ten Cate 1999b). Perhaps the sudden appearance of a social stimulus captured the attention of the zebra finch tutees in a different way than a non-specific movement and that and/or the scramble competition between the male and female juvenile we sometimes saw for the positions on perch 1 distracted them from the auditory stimulus. As demonstrated by the behavioural observations, males and females were equally attracted to the visual stimuli. It might be that the excitement of the companion by the visual social stimulation was more salient to the male tutees than the auditory song stimulus. This might also explain why the birds raised with the pixelated video had higher SAP tutor-tutee similarity scores than the birds raised with the tutor video, as both young birds seemed more excited by the tutor video than by the pixelated video (i.e. spending more time close

to it and approaching it more). The pixelated video was probably less socially meaningful to the tutees than the tutor video. In future studies we would have to test if other stimulus presentation schemes, e.g. more ongoing visual stimulus exposure instead of only very limited (sudden) exposure may lead to better song learning performance. It is also possible that the young females influenced males' song development by reinforcing particular song structures or encouraging a particular singing style or practicing (Jones and Slater 1993; Kojima and Doupe 2011; Ruploh et al. 2013; Carouso-Peck and Goldstein 2019). Female zebra finches do not sing themselves, but in mixed-age social rearing, they normally develop socially learned song preferences for the adult male song(s) they are exposed to as subadults (Miller 1979; Clayton 1988; Riebel 2000, 2003; Riebel et al. 2002; Holveck and Riebel 2014). Females could have learned from the tutor and then 'coached' the male tutees. If they learned equally well from the different tutoring methods, they might thereby have reduced the difference between treatment groups. However, if females, like the males in this study, learned rather poorly from the model, they might have learned from their male peers instead (as documented in Honarmand et al. 2015), and in turn reinforced aspects of their peers' songs. Through iteration of this process, both female preference and male song might have moved further away from the model song. Much will depend on how uni- versus multimodal tutoring affects female preference learning. We are not aware of any study directly investigating this question (but see Holveck & Riebel, 2014, for demonstrating that live and tape-tutored females develop preferences based on early song experiences). Whether song preference learning is differentially affected by multi- compared to unimodal tutoring will thus have to be explored further in the future. Even with the careful control of the stimulus preparations, it remains possible that the filming context of the videos was suboptimal. We presented audio and video stimuli of tutors recorded when alone and singing undirected song. Zebra finch adults can, however, produce pupil-directed song towards juvenile conspecifics, which differs from undirected and adult female-directed song in several acoustic parameters (Chen et al. 2016). As female-directed and undirected song also differ in the accompanying body movements (Sossinka and Böhner 1980), it is possible (but to our knowledge not yet tested) that specific visual components proceed, accompany or follow the production of pupil-directed song and that therefore tutoring with audio or audio-visual pupil-directed song might lead to better song learning outcomes compared to tutoring with undirected song. It would be interesting to repeat the current experiment using videos of tutors producing pupil-directed song to test this idea.

It is, however, also important to stress that although video playback can pro-

vide audio-visual stimulation, it remains to be seen whether a 2-dimensional tutor can ever replace a 3-dimensional live bird, as a video provides no depth and this might mean that a substantial part of the singing movements are not visible to the bird. It is also possible that not the multimodal cues per se but the social and interactive qualities of a live tutor need to be emulated in such a setup. For instance, operant tape-tutoring, where song playback is contingent on specific tutee behaviour, can lead to better learning outcomes than passive tape-tutoring, where tutees cannot predict when song playback will occur (Adret 1993; Derégnaucourt et al. 2013, but see Houx and ten Cate 1999b). Besides, behaviour or stimuli contingent on immature song production can positively affect song learning outcomes (Carouso-Peck and Goldstein 2019; Carouso-Peck et al. 2020). With respect to the role of behaviour and social interactions as important drivers for learning to take place, there are clear parallels between song learning and imprinting processes. For zebra finches, it has been shown that mere visual exposure to a stuffed bird (which might be compared to exposure to audio only playback), or even exposure to a live bird behind a wire had no or limited effect on being used as a model for sexual imprinting compared to when behavioural interactions could occur (ten Cate 1984; ten Cate et al. 1984). In a filial imprinting experiment, quail chicks exposed to a live hen behind a transparent screen developed a strong filial attachment, much stronger than chicks exposed to a moving stuffed hen, while exposure to a non-moving stuffed hen did not result in a measurable attachment (ten Cate 1989). Follow-up studies using animated three-dimensional visual stimulation, for instance in a Virtual Reality context or using robots, are necessary to further investigate the effect of presenting song production-related visual cues in addition to passive playback of tutor song on song learning as a first step and comparing such stimulation in interactive versus a non-responsive mode as a subsequent step.

In conclusion, in this study, although young birds were more attracted to and spent more time engaging with the audio-visual than the audio-only tutors, video presented visual cues related to sound production did not show a facilitating effect on vocal learning in zebra finches. Future studies with methodological adaptations are necessary to further investigate the influence of meaningful visual cues on the vocal learning process.

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. We want to

thank Cynthia Tedore for very helpful advice on video color adjustments and screen calibration and Carel ten Cate, an anonymous reviewer and the editor for comments on earlier versions of this manuscript.

References

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

Adret P (1997) Discrimination of video images by zebra finches (*Taeniopygia guttata*): direct evidence from song performance. J Comp Psychol 111:115–125. https://doi.org/10.1037/0735-7036.111.2.115

Adret P (1992) Imitation du chant chez les diamants mandarins: voir, entendre et int eragir. Ann la Fond Fyssen 7:73–82

Baptista LF, Gaunt SLL (1997) Social interaction and vocal development in birds. In: Snowdon CT, Hausberger M (eds) Social influences on vocal development. Cambridge, Cambridge University Press, pp 23–40

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

Beecher MD, Burt JM (2004) The role of social interaction in bird song learning. Curr Dir Psychol Sci 13:224–228. https://doi.org/10.1111/j.0963-7214.2004.00313.x

Bischof H-J, Böhner J, Sossinka R (1981) Influence of external stimuli on the quality of the song of the zebra finch. Z Tierpsychol 57:261–267. https://doi.org/10.1111/j.1439-0310.1981.tb01927.x

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

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

Bolhuis JJ, Okanoya K, Scharff C (2010) Twitter evolution: Converging mechanisms in birdsong and human speech. Nat Rev Neurosci 11:747–759. https://doi.org/10.1038/nrn2931

Bowmaker JK, Heath LA, Wilkie SE, Hunt DM (1997) Visual pigments and oil droplets from six classes of photoreceptor in the retinas of birds. Vision Res 37:2183–2194. https://doi.org/10.1016/S0042-6989(97)00026-6

Brainard MS, Doupe AJ (2002) What songbirds teach us about learning. Nature 417:351–358. https://doi.org/10.1038/417351a

Carouso-Peck S, Goldstein MH (2019) Female social feedback reveals non-imitative mechanisms of vocal learning in zebra finches. Curr Biol 29:631–636. https://doi.org/10.1016/j.cub.2018.12.026

Carouso-Peck S, Menyhart O, DeVoogd TJ, Goldstein MH (2020) Contingent parental responses are naturally associated with zebra finch song learning. Anim Behav 165:123–132. https://doi.org/10.1016/j.anbehav.2020.04.019

Catchpole CK, Slater PJB (1995) How song develops. In: Catchpole CK, Slater PJB (eds) Bird Song: Biological Themes and Variations. Cambridge: Cambridge University Press., pp 45–69

Chen L, Vroomen J (2013) Intersensory binding across space and time: A tutorial review. Attention, Perception, Psychophys 75:790–811. https://doi.org/10.3758/s13414-013-0475-4

Chen Y, Matheson LE, Sakata JT (2016) Mechanisms underlying the social enhancement of vocal learning in songbirds. Proc Natl Acad Sci 201522306. https://doi.org/10.1073/pnas.1522306113

Chouinard-Thuly L, Gierszewski S, Rosenthal GG, et al (2017) Technical and conceptual considerations for using animated stimuli in studies of animal behavior. Curr Zool 63:5–19. https://doi.org/10.1093/cz/zow104

Clayton NS (1988) Song discrimination learning in zebra finches. Anim Behav 36:1016–1024. https://doi.org/https://doi.org/10.1016/S0003-3472(88)80061-7.

Cuthill IC, Hart NS, Partridge JC, et al (2000) Avian colour vision and avian video playback experiments. Acta Ethol 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*). Interact Stud 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. J Physiol 107:210–218. https://doi.org/10.1016/j.jphysparis.2012.08.003

Deshpande M, Pirlepesov F, Lints T (2014) Rapid encoding of an internal model for imitative learning. Proc R Soc London Ser B Biol Sci 281:20132630. https://doi.org/10.1098/rspb.2013.2630

Doupe AJ, Kuhl PK (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 LA (1989) The influences of visual and vocal interaction on song learning in zebra finches. Anim Behav 37:507–508. https://doi.org/10.1016/0003-3472(89)90097-3

Feenders G, Kato Y, Borzeszkowski KM, Klump GM (2017) Temporal ventriloquism effect in european starlings: evidence for two parallel processing pathways. Behav Neurosci 131:337–347. https://doi.org/10.1037/bne0000200

Fleishman LJ, Endler JA (2000) Some comments on visual perception and the use of video playback in animal behavior studies. Acta Ethol 3:15–27. https://doi.org/10.1007/s102110000025

Funabiki Y, Funabiki K (2009) Factors limiting song acquisition in adult zebra finches. Dev Neurobiol 69:752–759. https://doi.org/10.1002/dneu.20738

Galoch Z, Bischof HJ (2007) Behavioural responses to video playbacks by zebra finch males. Behav Processes 74:21–26. https://doi.org/10.1016/j.beproc.2006.09.002

Goldstein MH, King AP, West MJ (2003) Social interaction shapes babbling: testing

parallels between birdsong and speech. Proc Natl Acad Sci U S A 100:8030–5. https://doi.org/10.1073/pnas.1332441100

Griffith SC, Buchanan KL (2010) The zebra finch : the ultimate Australian supermodel. Emu 110:v–xii. https://doi.org/10.1071/MUv110n3_ED

Guillette LM, Healy SD (2016) The roles of vocal and visual interactions in social learning zebra finches: A video playback experiment. Behav Processes 139:43–49. https://doi.org/10.1016/j.beproc.2016.12.009

Guillette LM, Healy SD (2019) Social learning in nest-building birds watching live-streaming video demonstrators. Integr Zool 14:204–213. https://doi.org/10.1111/1749-4877.12316

Halfwerk W, Varkevisser J, Simon R, et al (2019) Toward testing for multimodal perception of mating signals. Front Ecol Evol 7:2013–2019. https://doi.org/10.3389/fevo.2019.00124

Higham JP, Hebets EA (2013) An introduction to multimodal communication. Behav Ecol Sociobiol 67:1381–1388. https://doi.org/10.1007/s00265-013-1590-x

Holveck MJ, Riebel K (2014) Female zebra finches learn to prefer more than one song and from more than one tutor. Anim Behav 88:125–135. https://doi.org/10.1016/j.anbehav.2013.11.023

Honarmand M, Riebel K, Naguib M (2015) Nutrition and peer group composition in early adolescence: impacts on male song and female preference in zebra finches. Anim Behav 107:147–158. https://doi.org/10.1016/j.anbehav.2015.06.017

Houx BB, ten Cate C (1999a) Song learning from playback in zebra finches: is there an effect of operant contingency? Anim Behav 57:837–845. https://doi.org/10.1006/anbe.1998.1046

Houx BB, ten Cate C (1999b) Do stimulus-stimulus contingencies affect song learning in zebra finches (*Taeniopygia guttata*)? J Comp Psychol 113:235–242. https://doi.org/10.1037/0735-7036.113.3.235

Hultsch H, Schleuss F, Todt D (1999) Auditory-visual stimulus pairing enhances perceptual learning in a songbird. Anim Behav 58:143–149. https://doi.org/10.1006/anbe.1999.1120

Ikebuchi M, Okanoya K (1999) Male zebra finches and bengalese finches emit directed songs to the video images of conspecific females projected onto a TFT display. Zoolog Sci 16:63–70. https://doi.org/10.2108/zsj.16.63

James LS, Fan R, Sakata JT (2019) Behavioural responses to video and live presentations of females reveal a dissociation between performance and motivational aspects of birdsong. J Exp Biol 222:jeb206318. https://doi.org/10.1242/jeb.206318

Jesse F, Riebel K (2012) Social facilitation of male song by male and female conspecifics in the zebra finch, *Taeniopygia guttata*. Behav Processes 91:262–266. https://doi.org/10.1016/j.beproc.2012.09.006

Jones AE, Slater PJB (1993) Do young male zebra finches prefer to learn songs that are familiar to females with which they are housed. Anim Behav 46:616–617. https://doi.org/10.1006/anbe.1993.1233

Jones AE, ten Cate C, Bijleveld CCJH (2001) The interobserver reliability of scoring

sonagrams by eye: A study on methods, illustrated on zebra finch songs. Anim Behav 62:791–801. https://doi.org/10.1006/anbe.2001.1810
Kojima S, Doupe AJ (2011) Social performance reveals unexpected vocal competency in young songbirds. Proc Natl Acad Sci U S A 108:1687–92. https://doi.org/10.1073/pnas.1010502108
Kozak EC, Uetz GW (2016) Cross-modal integration of multimodal courtship signals in a wolf spider. Anim Cogn 19:1–9. https://doi.org/10.1007/s10071-016-1025-y
Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. Science (80-) 218:1138–1141. https://doi.org/10.1126/science.7146899
Lachlan RF, van Heijningen CAA, ter Haar SM, ten Cate C (2016) Zebra finch song phonology and syntactical structure across populations and continents-a computational comparison. Front Psychol 7:1–19. https://doi.org/10.3389/fpsyg.2016.00980
Lachlan RF, Verhagen L, Peters S, ten Cate C (2010) Are there species-universal cate gories in bird song phonology and syntax? A comparative study of chaffinches (*Fringilla coelebs*), zebra finches (*Taenopygia guttata*), and swamp sparrows (*Melospiza georgiana*). J Comp Psychol 124:92–108. https://doi.org/10.1037/a0016996
Lenth R, Singmann H, Love J, et al (2018) Emmeans: estimated marginal means, aka least-squares means
Lewkowicz DJ, Hansen-Tift AM (2012) Infants deploy selective attention to the mouth of a talking face when learning speech. Proc Natl Acad Sci U S A 109:1431–1436. https://doi.org/10.1073/pnas.1114783109
Lombardo SR, MacKey E, Tang L, et al (2008) Multimodal communication and spatial binding in pied currawongs (*Strepera graculina*). Anim Cogn 11:675–682. https://doi.org/10.1007/s10071-008-0158-z
Mello C V (2014) The zebra finch, *Taeniopygia guttata*: An avian model for investigating the neurobiological basis of vocal learning. Cold Spring Harb Protoc 2014:1237–1242. https://doi.org/10.1101/pdb.emo084574
Miller DB (1979) Long-term recognition of father's song by female zebra finches. Nature 280:389–391
Narins PM, Grabul DS, Soma KK, et al (2005) Cross-modal integration in a dart-poison frog. Proc Natl Acad Sci U S A 102:2425–2429. https://doi.org/10.1073/pnas.0406407102
Nelson D (1997) Social interaction and sensitive phases for song learning: A critical review. In: Snowdon CT, Hausberger M (eds) Social influences on vocal development. Cambridge, Cambridge University Press, pp 7–22
Oliveira RF, Rosenthal GG, Schlupp I, et al (2000) Considerations on the use of video playbacks as visual stimuli: the Lisbon workshop consensus. Acta Ethol 3:61–65. https://doi.org/10.1007/s102110000019
Partan S, Marler P (1999) Communication goes multimodal. Science (80-) 283:1272–1274. https://doi.org/0.1126/science.283.5406.1272
Patterson ML, Werker JF (1999) Matching phonetic information in lips and voice

is robust in 4.5-month-old infants. Infant Behav Dev 22:237–247. https://doi.org/10.1016/S0163-6383(99)00003-X

Phan ML, Pytte CL, Vicario DS (2006) Early auditory experience generates long-lasting memories that may subserve vocal learning in songbirds. Proc Natl Acad Sci U S A 103:1088–1093. https://doi.org/10.1073/pnas.0510136103

Pinzon-Rodriguez A, Muheim R (2017) Zebra finches have a light-dependent magnetic compass similar to migratory birds. J Exp Biol 220:1202–1209. https://doi.org/10.1242/jeb.148098

Price PH (1979) Developmental determinants of structure in zebra finch song. J Comp Physiol Psychol 93:260–277. https://doi.org/10.1037/h0077553

Ręk P (2018) Multimodal coordination enhances the responses to an avian duet. Behav Ecol 29:411–417. https://doi.org/10.1093/beheco/arx174

Riebel K (2003) Developmental influences on auditory perception in female zebra finches - is there a sensitive phase for song preference learning? Anim Biol 53:73–87

Riebel K (2000) Early exposure leads to repeatable preferences for male song in female zebra finches. Proc R Soc London Ser B Biol Sci 267:2553–8. https://doi.org/10.1098/rspb.2000.1320

Riebel K, Smallegange IM, Terpstra NJ, Bolhuis JJ (2002) Sexual equality in zebra finch song preference: evidence for a dissociation between song recognition and production learning. Proc R Soc London Ser B Biol Sci 269:729–33. https://doi.org/10.1098/rspb.2001.1930

Rowe C (1999) Receiver psychology and evolution of multicomponent signals. Anim Behav 58:921–931. https://doi.org/10.1006/anbe.1999.1242

Ruploh T, Bischof HJ, von Engelhardt N (2013) Adolescent social environment shapes sexual and aggressive behaviour of adult male zebra finches (Taeniopygia guttata). Behav Ecol Sociobiol 67:175–184. https://doi.org/10.1007/s00265-012-1436-y

Scharff C, Nottebohm F (1991) A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: Implications for vocal learning. J Neurosci 11:2896–2913. https://doi.org/10.1523/JNEUROSCI.11-09-02896.1991

Slater PJB, Eales LA, Clayton NS (1988) Song learning in zebra finches (*Taeniopygia guttata*): progress and prospects. Adv study Behav 18:1–34. https://doi.org/10.1016/S0065-3454(08)60308-3

Solomon S, Lennie P (2007) The machinery of colour vision. Nat Rev Neurosci 8:276–286. https://doi.org/10.1038/nrn2094

Soma MF (2011) Social factors in song learning: a review of Estrildid finch research. Ornithol Sci 10:89–100. https://doi.org/10.2326/osj.10.89

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

Swaddle JP, McBride L, Malhotra S (2006) Female zebra finches prefer unfamiliar males but not when watching noninteractive video. Anim Behav 72:161–167. https://doi.org/10.1016/j.anbehav.2005.12.005

Taylor RC, Klein BA, Stein J, Ryan MJ (2011) Multimodal signal variation in space and time: how important is matching a signal with its signaler? J Exp Biol 214:815–820. https://doi.org/10.1242/jeb.043638

Tchernichovski O (2011) SAP User Manual (downloaded from: http://soundanalysis-pro.com/manual-1/manual-pdf/view)

Tchernichovski O, Lints T, Mitra PP, Nottebohm F (1999) Vocal imitation in zebra finches is inversely related to model abundance. Proc Natl Acad Sci U S A 96:12901–4. https://doi.org/10.1073/pnas.96.22.12901

Tchernichovski O, Mitra PP (2002) Towards quantification of vocal imitation in the zebra finch. J Comp Physiol A 188:867–878. https://doi.org/10.1007/s00359-002-0352-4

Tchernichovski O, Nottebohm F, Ho CE, et al (2000) A procedure for an automated measurement of song similarity. Anim Behav 59:1167–1176. https://doi.org/10.1006/anbe.1999.1416

Tedore C, Johnsen S (2017) Using RGB displays to portray color realistic imagery to animal eyes. Curr Zool 63:27–34. https://doi.org/10.1093/cz/zow076

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

ten Cate C (1984) The influence of social relations on the development of species recognition in zebra finch males. Behaviour 91:263–285. https://doi.org/10.1163/156853984X00100

ten Cate C (1989) Stimulus movement, hen behaviour and filial imprinting in Japanese quail (*Coturnix coturnix japonica*). Ethology 82:287–306

ten Cate C (1991) Behaviour-contingent exposure to taped song and zebra finch song learning. Anim Behav 42:857–859. https://doi.org/10.1016/S0003-3472(05)80131-9

ten Cate C, Los L, Schilperoord L (1984) The influence of differences in social expe rience on the development of species recognition in zebra finch males. Anim Behav 32:852–860

Tenenbaum EJ, Sobel DM, Sheinkopf SJ, et al (2015) Attention to the mouth and gaze following in infancy predict language development. J Child Lang 42:1173–1190. https://doi.org/10.1017/S0305000914000725

van Kampen HS, Bolhuis JJ (1993) Interaction between auditory and visual learning during filial imprinting. Anim Behav 45:623–625. https://doi.org/10.1006/anbe.1993.1074

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

Ware E, Saunders DR, Troje NF (2015) The influence of motion quality on responses towards video playback stimuli. Biol Open 4:803–811. https://doi.org/10.1242/bio.011270

Williams H (1990) Models for song learning in the zebra finch: fathers or others? Anim Behav 39:745–757. https://doi.org/10.1016/S0003-3472(05)80386-0

Appendix

Output of the MATLAB script provided by Tedore and Johnson (2017) with the correction factors (new calculations with ASUS adjustments)

```
Suppl_2_Natural_Radiance_to_RGB_Coordinates
How many photoreceptor classes does your study species have? 3
Please specify the 0-255 brightness level to which you would like to
set the G phosphor of your background color. 85
background best fit b1 factor = 7.3454e-10, b2 factor = 7.3657e-10, b3
factor = 7.3566e-10
best possible RGB background values: R = 77, G = 85, B = 98
Calculating optimal RGB rendering for color patch spectra, please
wait...
target patch 1 Q1 = 0.16807, Q2 = 0.33968, Q3 = 1
best possible RGB patch 1 Q1 = 0.16642, Q2 = 0.33695, Q3 = 1
best possible RGB patch 1 values: R = 119, G = 72, B = 65
target patch 2 Q1 = 0.15426, Q2 = 0.51055, Q3 = 1
best possible RGB patch 2 Q1 = 0.15392, Q2 = 0.51063, Q3 = 1
best possible RGB patch 2 values: R = 107, G = 86, B = 51
target patch 3 Q1 = 0.22301, Q2 = 0.50781, Q3 = 1
best possible RGB patch 3 Q1 = 0.2228, Q2 = 0.51751, Q3 = 1
best possible RGB patch 3 values: R = 92, G = 74, B = 58
Elapsed time is 335.758313 seconds.
```

Table 11 Results for post-hoc comparison of proportion of time that tutees spent in different areas of the cage (corrected for the perch length in that area). Significant p-values are indicated in bold.

Contrast	Estimate	SE	t	p
Area 1 Audio vs Area 1 Audio-video	-0.32	0.05	-6.72	**<0.01**
Area 1 Audio vs Area 1 Audio-pixel	-0.14	0.05	-2.95	0.09
Area 1 Audio vs Area 2 Audio	0.07	0.05	1.56	0.82
Area 1 Audio vs Area 2 Audio-video	0.27	0.05	5.60	**<0.01**
Area 1 Audio vs Area 2 Audio-pixel	0.16	0.05	3.33	**0.03**
Area 1 Audio vs Area 3 Audio	0.17	0.05	3.61	**0.01**
Area 1 Audio vs Area 3 Audio-video	0.36	0.05	7.47	**<0.01**
Area 1 Audio vs Area 3 Audio-pixel	0.24	0.05	5.13	**<0.01**
Area 1 Audio-video vs Area 1 Audio-pixel	0.18	0.05	3.77	**<0.01**
Area 1 Audio-video vs Area 2 Audio	0.40	0.05	8.28	**<0.01**
Area 1 Audio-video vs Area 2 Audio-video	0.59	0.05	12.32	**<0.01**
Area 1 Audio-video vs Area 2 Audio-pixel	0.48	0.05	10.05	**<0.01**
Area 1 Audio-video vs Area 3 Audio	0.49	0.05	10.33	**<0.01**

Area 1 Audio-video vs Area 3 Audio-video	0.68	0.05	14.19	**<0.01**
Area 1 Audio-video vs Area 3 Audio-pixel	0.57	0.05	11.85	**<0.01**
Area 1 Audio-pixel vs Area 2 Audio	0.22	0.05	4.51	**<0.01**
Area 1 Audio-pixel vs Area 2 Audio-video	0.41	0.05	8.55	**<0.01**
Area 1 Audio-pixel vs Area 2 Audio-pixel	0.30	0.05	6.28	**<0.01**
Area 1 Audio-pixel vs Area 3 Audio	0.31	0.05	6.56	**<0.01**
Area 1 Audio-pixel vs Area 3 Audio-video	0.50	0.05	10.42	**<0.01**
Area 1 Audio-pixel vs Area 3 Audio-pixel	0.39	0.05	8.08	**<0.01**
Area 2 Audio vs Area 2 Audio-video	0.19	0.05	4.04	**<0.01**
Area 2 Audio vs Area 2 Audio-pixel	0.08	0.05	1.77	0.70
Area 2 Audio vs Area 3 Audio	0.10	0.05	2.05	0.51
Area 2 Audio vs Area 3 Audio-video	0.28	0.05	5.90	**<0.01**
Area 2 Audio vs Area 3 Audio-pixel	0.17	0.05	3.57	**0.02**
Area 2 Audio-video vs Area 2 Audio-pixel	-0.11	0.05	-2.27	0.37
Area 2 Audio-video vs Area 3 Audio	-0.09	0.05	-1.98	0.56
Area 2 Audio-video vs Area 3 Audio-video	0.09	0.05	1.87	0.64
Area 2 Audio-video vs Area 3 Audio-pixel	-0.02	0.05	-0.46	0.99
Area 2 Audio-pixel vs Area 3 Audio	0.01	0.05	0.28	1.00
Area 2 Audio-pixel vs Area 3 Audio-video	0.20	0.05	4.14	**<0.01**
Area 2 Audio-pixel vs Area 3 Audio-pixel	0.09	0.05	1.80	0.68
Area 3 Audio vs Area 3 Audio-video	0.18	0.05	3.85	**<0.01**
Area 3 Audio vs Area 3 Audio-pixel	0.07	0.05	1.52	0.84
Area 3 Audio-video vs Area 3 Audio-pixel	-0.11	0.05	-2.33	0.33

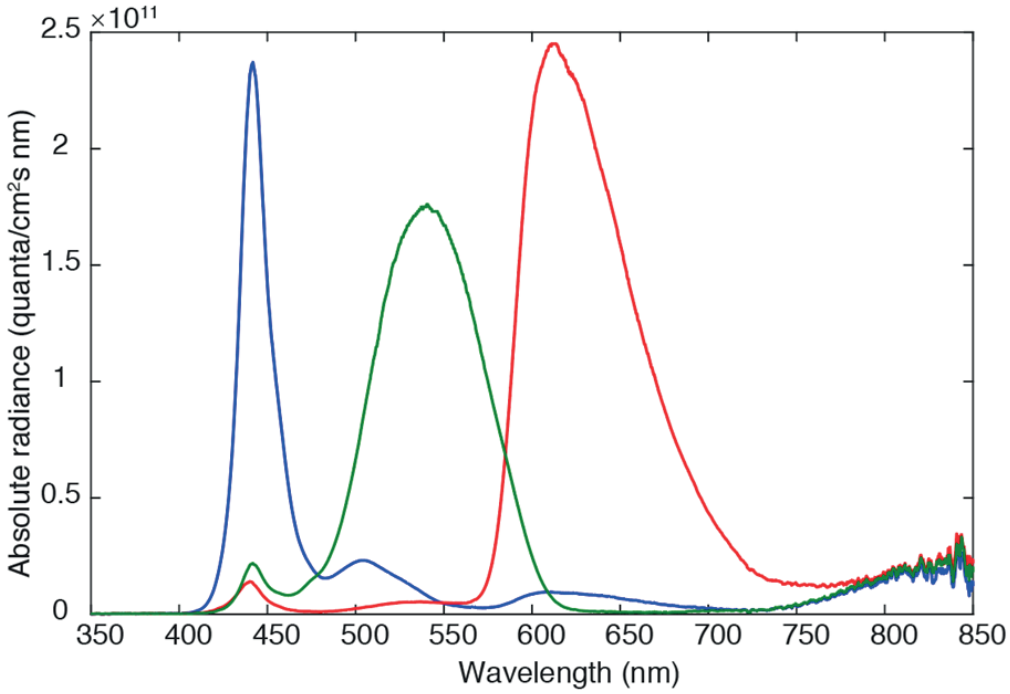

Figure 8. Absolute radiance of the ASUS gaming monitors used for the stimuli presentation.

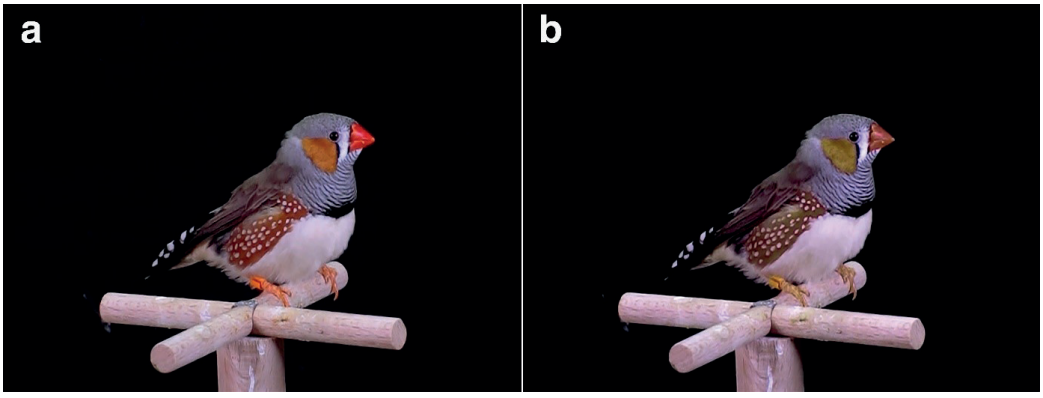

Figure 9. Example frames from a video stimulus. (a) Original frame before colour adjustment. (b) Colour adjusted frame which was used for stimulus presentation. Note that the colours were adjusted for presentation on a particular screen (VG248QE, ASUS, Taipei, Taiwan) and that colours might deviate if shown on a different screen or in a printed version.

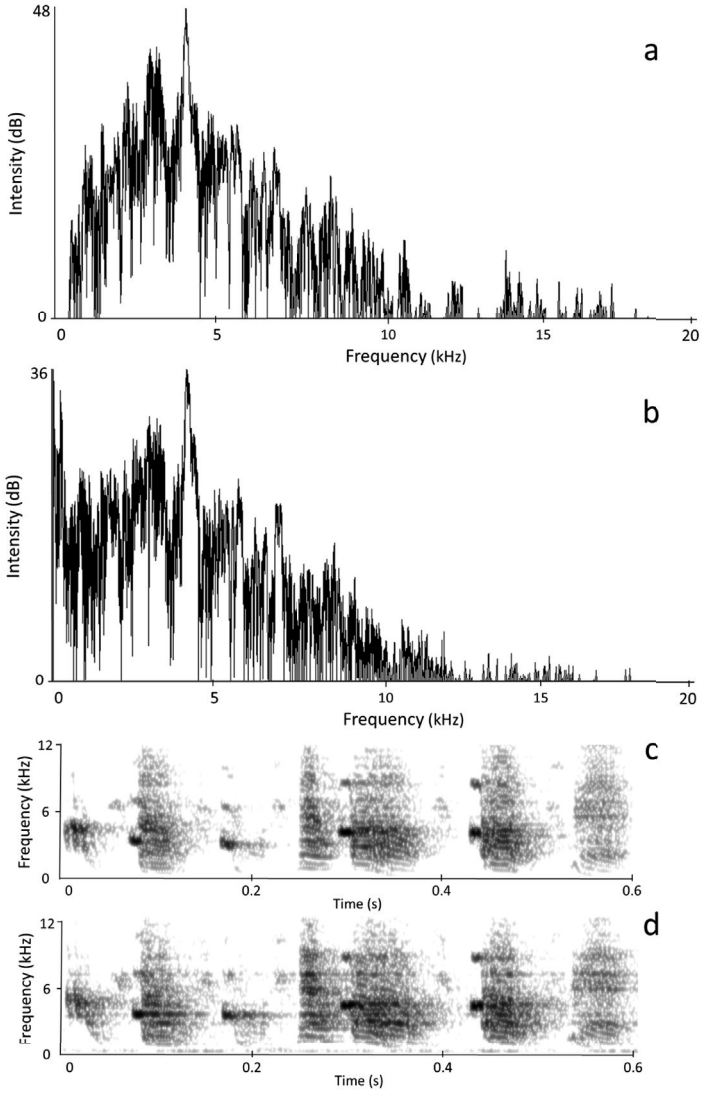

Figure 10. Power spectra of one motif of one of the tutors in the original recording (a) and re-recorded after playback in the experimental set-up (b). Spectrogram of the same original recording (c) and the re-recorded playback in the experimental set-up (d).

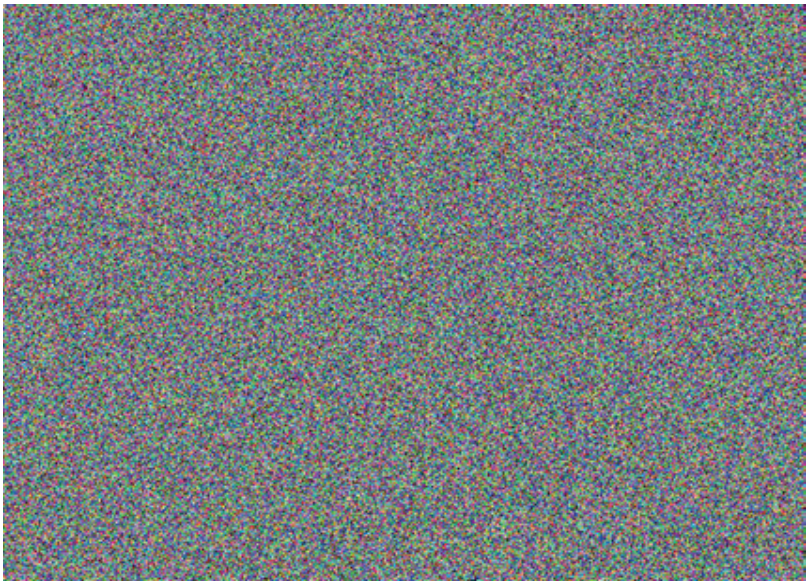

Figure 11. Image of the random pixels used for the displacement filter to generate the pixelated videos.

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.

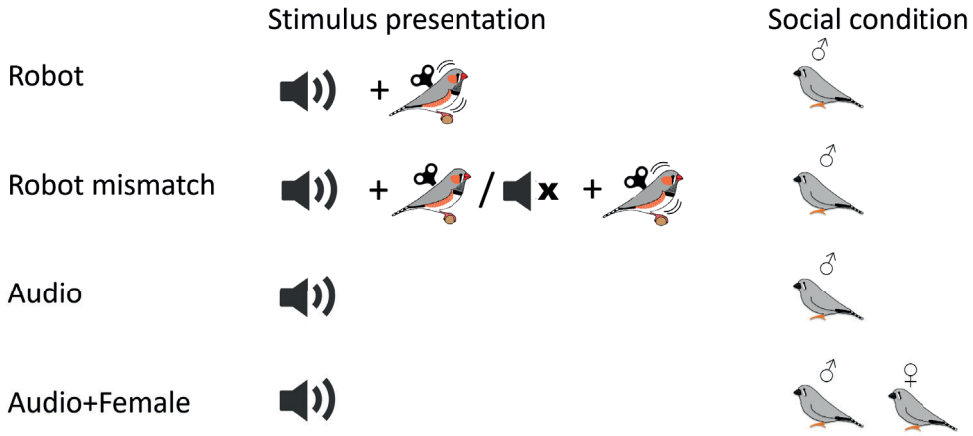

Figure 1. Schematic overview of the four different tutoring treatments. The normal loudspeaker symbol represents auditory song playback, while the loudspeaker symbol with a cross represents a situation without auditory song playback. Lines next to the RoboFinch icon (zebra finch with winding key on its 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).

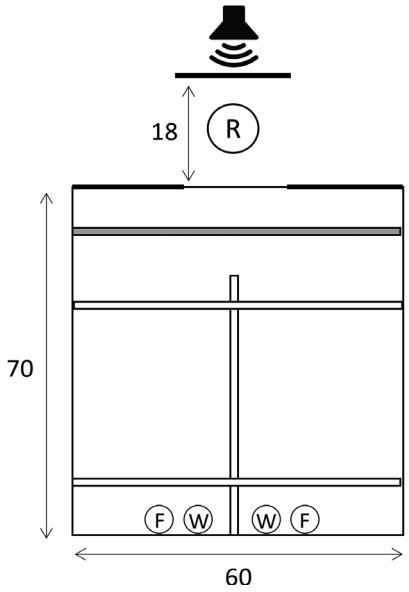

Figure 2. Schematic top view of the experimental set-up with perches. R = location of the RoboFinch in the Robot and Robot mismatch treatment, F = food, W = water. The loudspeaker was placed behind a panel covered in black loudspeaker cloth. All measurements are in cm.

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.

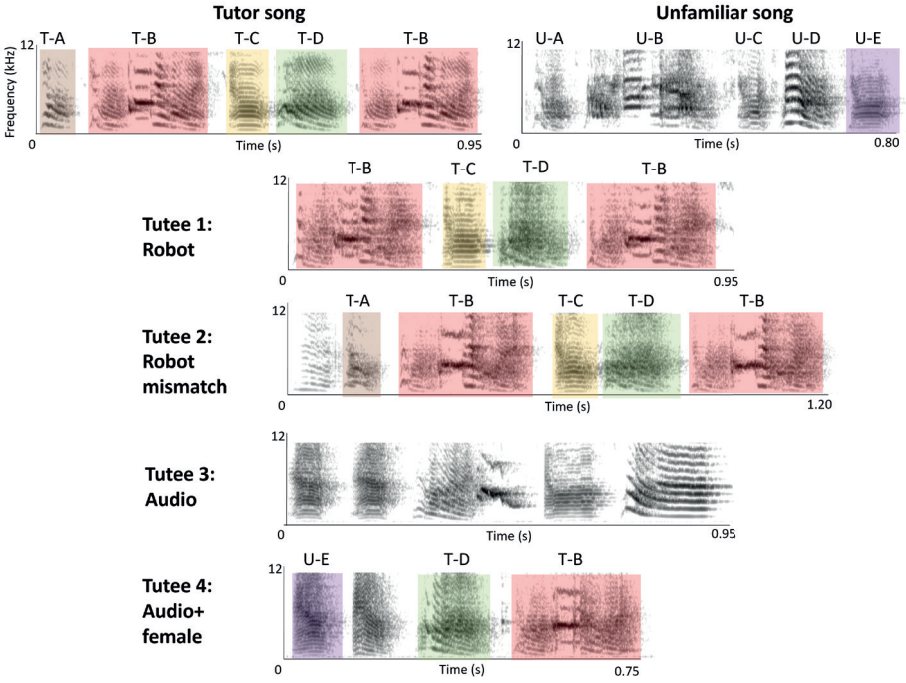

Figure 3. Spectrograms of full motif of tutor, unfamiliar full motif of another adult male and three tutees from one tutor group. Letters above spectrograms of tutor and unfamiliar song indicate how syllables were labelled with letters for further analyses. Human observer similarity between tutors and 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).

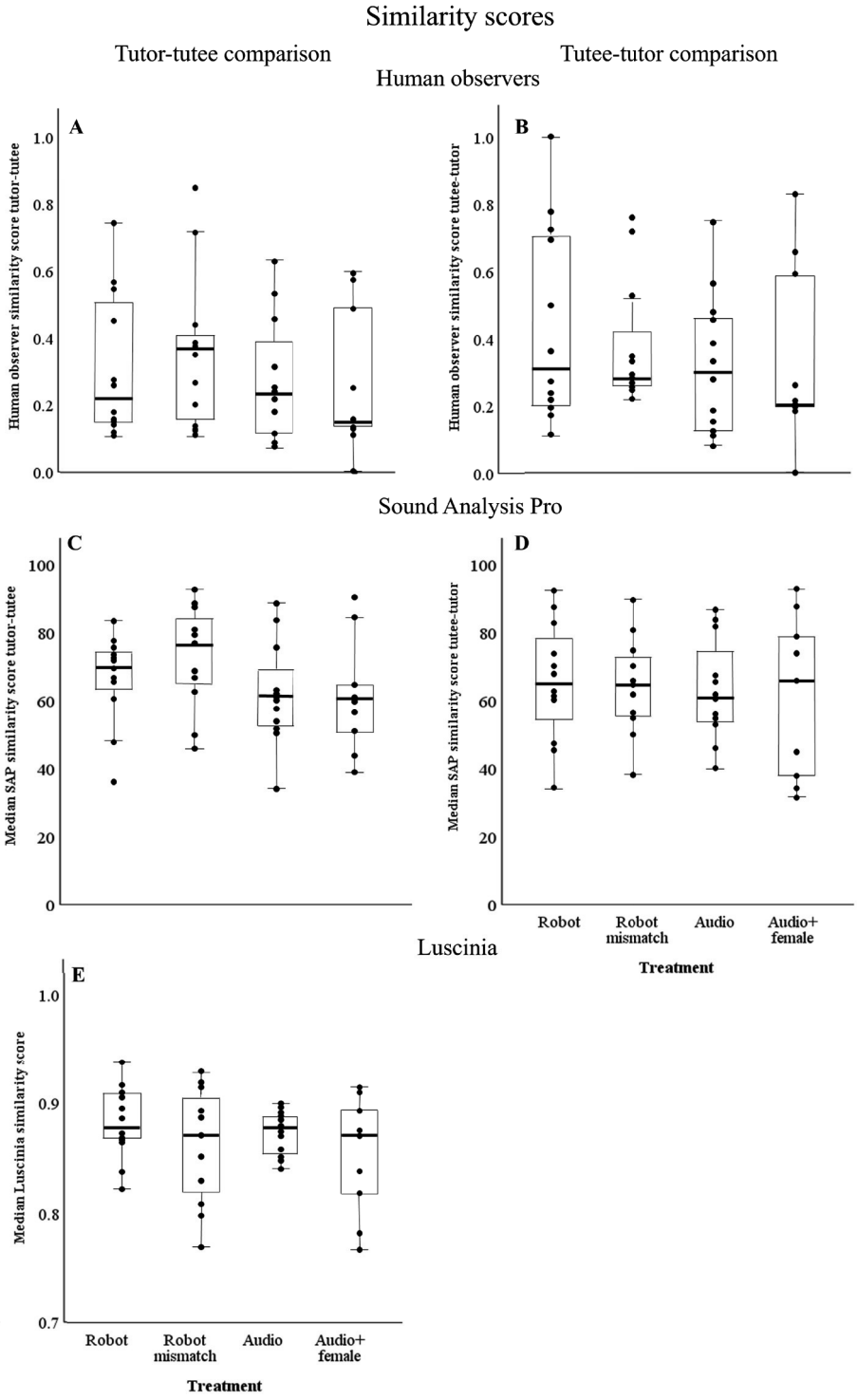

Figure 4. Graph showing the human observer similarity score for the tutor-tutee (a) and the tutee-tutor comparison (b), the SAP similarity score for the tutor-tutee (c) and the tutee-tutor (d) comparison and the Luscinia similarity score for the symmetric tutee and tutor comparison (e).

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).

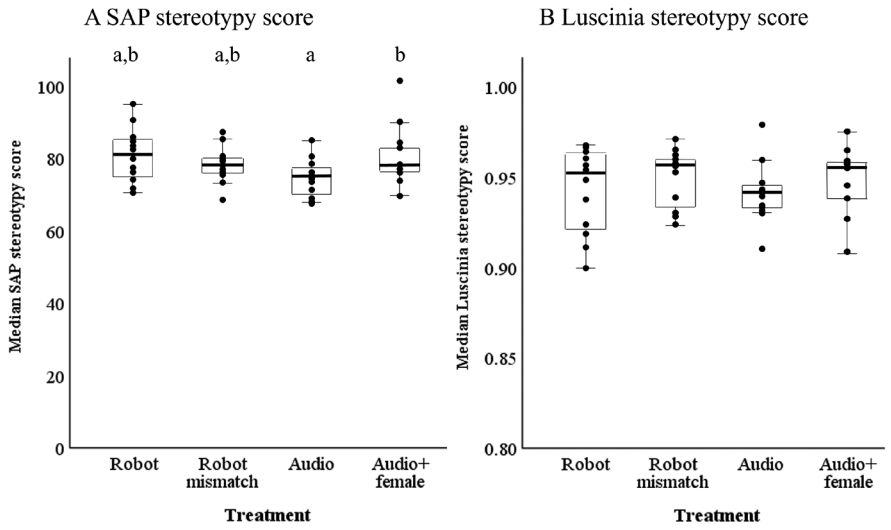

Figure 5. SAP (A) and Luscinia (B) stereotypy scores for the 10 randomly selected tutee motifs. Different letters above boxes in 5A indicate a significant difference of $p < 0.05$ according to post-hoc test, LMM see Table 6A.

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			

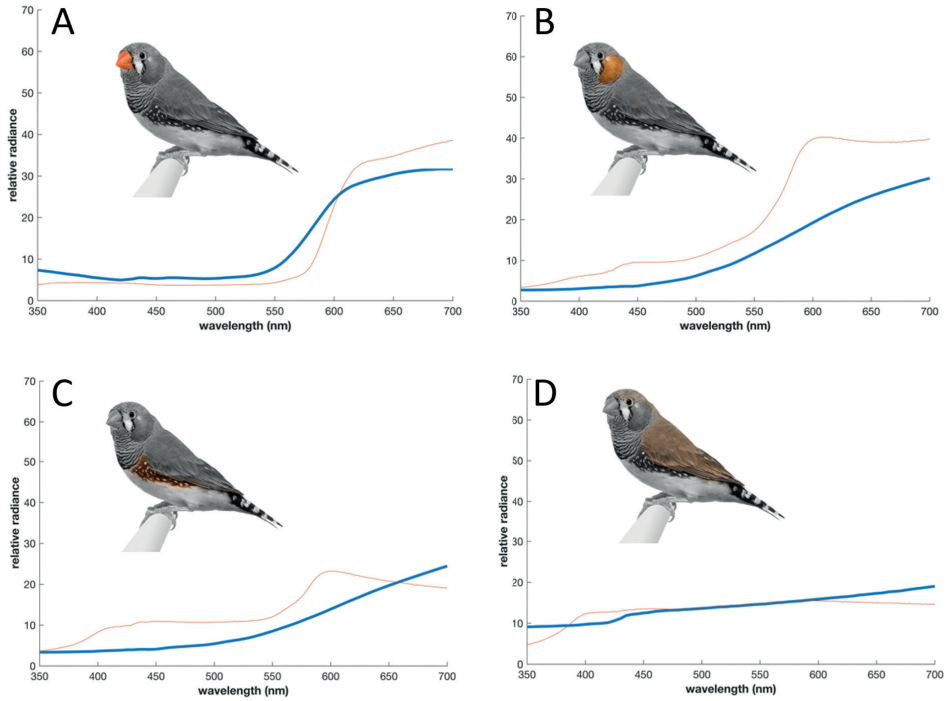

Figure A1. Colouring of the RoboFinch (red line) in comparison to real feathers/beaks of zebra finches (blue line). (a) beak, (b) cheeks, (c) sides, (d) back.

Chapter 6

Thesis summary and general discussion

Bird song is one of the most thoroughly studied animal examples of a vocally learned signal (Catchpole and Slater 1995; Bradbury and Vehrencamp 2011) and often used as a model system for human speech development, because of the many parallels between speech and bird song (Doupe and Kuhl 1999; Bolhuis et al. 2010). Several songbird species learn less well from audio-only tutor song exposure (so called 'tape tutoring') than from live social tutors (reviewed in Baptista & Gaunt, 1997; Soma, 2011). This might be because live tutoring, unlike audio-only tutoring, enables social tutor-tutee interactions, which are thought to play an important role in the vocal learning process (e.g. Beecher & Burt, 2004; Goldstein, King, & West, 2003; Kuhl, 2003, 2007, but also see Nelson 1997). It is unclear, however, whether and to what extent live tutoring also facilitates song learning because it results in multimodal exposure to a tutor, as tutees can both see and hear their tutor, while audio-only tutoring results in unimodal exposure. In this thesis, I investigated the effect of audio-visual compared to audio-only exposure to a tutor on song learning in zebra finches, a songbird species often-cited for learning less well from audio-only tutors than from live tutors (Eales 1989; Derégnaucourt et al. 2013; Chen et al. 2016). In this chapter, I will summarize and discuss the results of the studies described in this thesis and indicate what future research can further improve our knowledge on the effect of multimodal tutor exposure on vocal learning.

Open issues from previous (zebra finch) song tutoring studies

To get more insight into the factors playing a role in the vocal learning process, the effect of different tutoring paradigms on birdsong learning has been studied extensively, especially in zebra finches. Like multiple other songbird species, zebra finches learn more from a social, live tutor than from audio-only exposure to tutor song (Eales 1989; Derégnaucourt et al. 2013; Chen et al. 2016). Several studies have investigated the effect of specific dimensions on zebra finch song learning in order to find out what facilitates song learning from live tutors compared to audio-only song exposure (e.g. Adret, 1993; Bolhuis, van Mil, & Houx, 1999; Houx & ten Cate, 1999a). Based on the outcomes of these studies, it is now often hypothesized that social interaction with a tutor is the key facilitating aspect of live compared to audio-only tutoring (e.g. Chen, Matheson, & Sakata, 2016; Derégnaucourt, Poirier, Kant, & Linden, 2013; Slater, Eales & Clayton, 1988). In Chapter 2, previous zebra finch song tutoring studies were reviewed to find out whether they have systematically controlled for multi- and unimodal tutoring while studying the importance of social interaction for zebra finch song learning. In almost all studies, tutees with multimodal tutor exposure could socially interact with their tutor, while tutees with unimodal tutor exposure could not socially interact with their tutor (Chapter 2). Studies

thus usually confounded 'multimodal' and 'social' tutoring. Social tutoring tends to lead to improved song learning compared to non-social tutoring, but as social and multimodal were confounded, this might partly be due to a facilitating effect of multimodal exposure to a tutor. Another systematic difference between live and audio-only tutoring studies was the social environment of the tutees during tutoring. While audio-only tutored birds were usually housed in social isolation during tutoring, live tutored birds had the tutor as a social companion. This makes it unclear whether the lower amount of song copying from audio-only tutors might partly be attributed to an adverse effect of social isolation on song learning in the tape tutored tutees (Chapter 2). The song tutoring experiments described in this thesis were therefore aimed at testing the effect of multi- versus unimodal tutor exposure, while tutees in the different tutoring conditions were housed in comparable social environments during tutoring.

Song tutoring experiments comparing audio and audio-visual tutor exposure
The first tutoring experiment of this thesis, described in Chapter 3, was designed to investigate whether multi- compared to unimodal exposure to a live tutor would facilitate zebra finch song learning. To this end, zebra finch tutees were offered visual exposure to an adult tutor through a one-way mirror, in addition to auditory tutor exposure. Song learning in these tutees was compared to that in tutees that were raised in the same cage as the tutor and in tutees that were only auditorily exposed to the tutor. All tutees in this experiment were housed with a female companion. The tutees with multimodal exposure were expected to show improved tutor song copying compared to the tutees with unimodal exposure. The song analysis suggested that the unimodally tutored tutees had copied less tutor song than the tutees from the other groups, although the difference was not significant. I also found that the multimodally tutored tutees differed in their song ontogeny from the unimodally tutored tutees: more changes occurred after 65 days post-hatching in the song of the audio-only tutored birds than in that of the live tutored birds, while the audio-visually tutored birds did not differ from the live tutored birds. Although these results could be interpreted to support that multimodal tutor exposure facilitates song learning, an alternative explanation could be that visual feedback from the tutor in response to the tutees' vocalizations had facilitated song learning. To offer multimodal tutor exposure without the possibility of the tutor providing visual feedback to the tutees, I used artificial tutors in the other tutoring experiments described in chapter 4 and 5 of this thesis.

The tutoring experiment described in Chapter 4, investigated song learning in tutees that could see a video of the tutor singing the song that they were at the

same time auditorily exposed to. I compared these tutees to tutees that were only auditorily exposed to song and to tutees that heard song while they were exposed to the same tutor video, but here the pixels were randomized and the frames were played in reversed order. Again, all tutees were housed with a female companion. I expected that the tutees that were exposed to the normal video in addition to song playback would show improved song learning compared to the audio with the pixelated video and audio-only tutoring conditions. The results, however, did not show that the tutor videos led to improved song learning, even though the tutees in the condition with the normal tutor video were attracted most by the stimulus presentation. The videos used in this experiment were adjusted to zebra finch vision with state-of-the-art techniques, but it might be that certain properties of the videos, such as the brightness, negatively affected the birds' acceptance of the videos as conspecific tutors. Additionally, the lack of three-dimensionality in the videos might have made the visual cues less salient. Therefore a three-dimensional robotic zebra finch (Robo- Finch) was used for the visual stimulation in Chapter 5.

In the experiment described in Chapter 5, I investigated song learning in tutees that were exposed to the playback of pre-recorded tutor song, while a RoboFinch was simultaneously producing the beak and head movements that normally accompany the production of this song. These tutees were compared to tutees exposed to the same tutor song without a RoboFinch present and to tutees exposed to a RoboFinch that started moving after song playback had finished. These tutees were all housed in social isolation, and to investigate whether that affected their song learning outcomes, I also included a condition in which tutees were housed with a female companion while they were only auditorily exposed to the tutor song. The expectation was that the visual cues that were synchronized with the auditory song presentation would lead to improved song learning compared to the other tutoring conditions. However, I did not find any significant effects of the visual cues on song learning success. There was an effect of the social companion during tutoring on song learning outcomes: the 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.

In the following paragraphs, I will discuss what the results of these song tutoring experiments suggest about the effect of a social companion and the effect of audio-visual versus audio-only tutor exposure on song learning.

The effect of having a social companion during tutoring on song learning
In previous zebra finch song tutoring studies, significantly higher song learning success was found in live than in audio-only tutored tutees. However, in these studies the live tutored tutees had the tutor as a social companion, while the audio-only tutored tutees where housed in social isolation (e.g. Eales 1989; Derégnaucourt et al. 2013; Chen et al. 2016). In the song tutoring experiment described in Chapter 3, I compared song learning in audio-only and live tutored tutees that were both socially housed with a female companion (who does not sing) during song tutoring. With all tutees socially housed, the song of the live tutored tutees was not significantly more similar to the tutor song than the song of the tutees that were only auditorily exposed to the tutor. This suggests that the social isolation of the audio-only tutored tutees in previous studies might have contributed to the difference in song learning success between audio-only and live tutored tutees. The tutoring conditions did not lead to significant differences between the groups, but out of the three tutoring conditions in Chapter 3, the tutees from the live tutoring condition copied most from the tutor, which is in line with previous studies showing more learning from live than audio-only tutors and which suggests that the previously found difference between live and audio-only tutored tutees cannot solely be attributed to the difference in the social environment of the tutees during tutoring.

In the experiment described in Chapter 5, song learning from pre-recorded audio-only song playback was compared in male tutees that were housed in social isolation and in male tutees that were housed with a female companion during the tutoring period. The amount of tutor song copied did not differ between these tutees. Between-motif stereotypy, however, was higher in the tutees tutored with a female companion than in the tutees that were tutored in social isolation (Chapter 5). Song learning outcomes can thus be affected by whether zebra finches are housed with a social companion or in social isolation during tutoring. In future studies, it is therefore important to make sure that birds tutored in different tutoring conditions are housed in comparable social environments during the tutoring phase.

Comparing audio-visual and audio-only tutoring conditions
To investigate song learning from audio-visual and audio-only tutors, three tutoring experiments were conducted in which tutees in an audio-only condition were presented with tutor song auditorily only, while tutees in an audio-visual condition received the exact same song exposure auditorily while being visually exposed to either the live tutor producing this song (Chapter 3), a two-dimensional video of the tutor producing this song (Chapter 4) or a three-dimen-

sional robot tutor producing the beak and head movements accompanying the production of this song (Chapter 5). The birds that thus received audio-visual tutoring were unable to have visual social interactions with their tutors, and therefore the effect of audio-visual tutor exposure could be investigated independent of the effect of social tutor-tutee interactions.

In these tutoring studies, song learning success in the different treatments was assessed by comparing the adult song of the tutees to the song of their tutor. The findings in the experiment described in Chapter 3 suggested that tutees with audio-visual exposure to a live tutor tended to have a higher tutor song learning success than tutees with audio-only exposure to a live tutor: the song of the audio tutees tended to show the lowest, and the song of the live tutees the highest similarity with the tutor song, while the audio-visual tutees showed an intermediate level of similarity. Conversely, the audio group tended to show the highest similarity with the song of their father, which they were exposed to before the experimental tutoring. In the tutoring experiments described in Chapter 4 and Chapter 5, the audio-visual tutoring conditions did not lead to improved tutor song copying compared to the audio-only tutoring conditions. Across the three experimental methods described in this thesis, multimodal exposure to a live tutor thus seemed to have induced higher song learning success than unimodal exposure, while multimodal exposure to artificial tutors did not lead to improved song learning success compared to unimodal exposure.

To study the effect of audio-visual or audio-only tutoring on the timing of song learning, tutee song was recorded at two different moments in time: once at 65 days post-hatching, which is still during song development, and once after 100 days post-hatching, when song is normally crystallized (Gobes et al. 2017). To find out whether tutor groups differed in how 'developed' song already was at 65 days post-hatching when compared to song at 100 days, I recorded the motifs produced by the tutees at these two moments. In the tutoring experiment described in Chapter 3, more changes after 65 days occurred in the audio-only tutored birds than in live tutored birds, while the audio-visually tutored birds did not differ from live tutored birds in this respect. This is in line with an earlier finding demonstrating that zebra finch tutees that were exposed to a tutor only auditorily change their song up to a later age than tutees reared together with tutors in aviaries (Morrison & Nottebohm, 1993). The conclusion of this earlier study was that the closing of the sensitive period depends on whether a bird was able to have visual social interaction with a tutor. In the experiment in Chapter 3, however, I did not find a difference in the amount of changes between the live and the audio-visually tutored group, even though the audio-vis-

ually tutored group could not have visual social interaction with the tutor. This suggests that the timing of song development might not only be influenced by visual social interaction, but also by mere visual exposure to the tutor. The song produced by the tutees from the experiments in Chapter 4 and 5 were unfortunately still too variable at 65 days to use it for further analyses. This might have had to do with the tutoring through passive play-back of pre-recorded tutor song in these studies instead of the tutoring by a live conspecific in Chapter 3 that enabled vocal tutor-tutee interaction, which might affect song learning (Chapter 2). It is possible that the tutoring conditions in Chapter 4 and 5 did not lead to differences in the amount of tutor song copied by the tutees, but did affect the course of song development. Unfortunately, the current data do not allow a conclusion on whether this was the case. The effect of multi- versus unimodal tutoring on the timing of song development found in Chapter 3 shows that it is worthwhile to record both subadult and adult song of zebra finch tutees in tutoring experiments, as it can demonstrate whether different tutoring conditions affect the time course of vocal development. Future studies should address this, using a method that can assess how developed highly variable subadult song is.

In addition to effects on song learning, possible effects of uni- or multimodal tutoring on tutee behaviour were investigated. In Chapter 3, I tested how much time tutees in the audio and audio-visual condition spent in the observation huts, which was the location from which the tutees in the audio-visual, but not in the audio-only condition could see the tutor. Overall, tutees spent a higher proportion of time in the observation huts than expected. However, the tutees from the audio-visual and audio group spent a comparable amount of time in the huts. The possibility to see the tutor thus did not lead to an increase in hut visits. Although the video tutoring experiment described in Chapter 4 did not demonstrate a difference in song learning, it did show that tutee behaviour during song presentation was affected by the different tutoring conditions. The condition where auditory song presentation was accompanied by a video of the tutor producing the song was most salient to the tutees, but this did not lead to increased song learning, as mentioned before, possibly due to insufficient quality of the visual stimulus. In another tutoring study, zebra finch tutees spent more time on the perch next to a visual stimulus (a taxidermic mount of an adult male zebra finch) during than before its exposure, but the presentation of this visual stimulus also did not facilitate song learning (Houx and ten Cate 1999a). This suggests that visual stimulation presented together with auditory song presentation affects tutee behaviour, but not necessarily song learning success.

In the experiments described in Chapter 4 and 5, a control condition was included in which tutees were raised with visual stimulation that had no rhythmic correspondence with the auditory stimulation. These conditions were included to investigate whether non-social, non-sound-contingent visual stimulation would affect song learning differently than sound-contingent visual stimulation (namely the beak and body movements normally accompanying song production). In Chapter 4, for this condition a video was used in which the pixels were randomized and the frames were played in reversed order. In Chapter 5, tutees in this condition were raised with a RoboFinch that started moving after auditory song playback had finished. We expected that the synchronized audio-visual conditions would improve song learning more than these control conditions, for instance because nightingales show improved learning from song playbacks presented with a synchronously flashing stroboscope (Hultsch et al. 1999). Contrary to our expectation, the control conditions did not affect song learning outcomes differently from the 'normal' audio-visual conditions. However, in Chapter 4, tutees spent more time close to the normal tutor video than to the pixelated and reversed video, suggesting that social, sound-specific visual stimulation might be more salient to tutees than non-social, non-sound-specific visual stimulation. Likewise, in humans, sound-specific motor gestures have been found to attract the attention of infants more than unspecific gestures (Kuhl and Meltzoff 1982; Patterson and Werker 1999) and several other studies in animals have shown effects of correctly synchronized visual and acoustic information on perceptual salience (e.g. Taylor et al. 2011; Rȩk 2018).

Effects of audio-visual tutoring on vocal learning

Based on the literature, I hypothesized that audio-visual tutor exposure would lead to improved song learning compared to audio-only tutor exposure (reviewed in Chapter 2). Although the results of Chapter 3 were in line with this hypothesis, this hypothesis was not supported by the results of Chapter 4 and 5. However, Chapter 3 used a live conspecific tutor, while Chapter 4 and 5 used artificial tutors that had not been used in previous tutoring studies. It is thus unclear to what extent methodological decisions, such as the amount and timing of song playback and the stereotypy of song presentation, affected song learning outcomes. It is also unclear whether the visual quality of these tutors was sufficient to affect song learning. However, in the context of imprinting, learning of an auditory signal in chickens was enhanced when simultaneously with the presentation of the auditory signal a rotating box was shown (van Kampen and Bolhuis 1991, 1993). Moreover, young nightingales learn songs from audio playbacks combined with stroboscope light flashes better than

songs presented as audio-only playbacks (Hultsch et al. 1999). This suggests that other bird species can show improved learning of auditory signals when these are paired with any moving visual stimulation. Another difference between the experiment in Chapter 3 on the one hand, and the experiments described in Chapter 4 and 5 on the other, is that the latter experiments were carried out in sound attenuated chambers in which tutees did not hear anything except for their own vocalizations and the tutor song. One of the advantages of multi- compared to unimodal signalling is that multimodal signals are more likely to be detected by receivers than unimodal signals (reviewed in Rowe, 1999). In the experiments in the sound attenuated chambers, it was very unlikely that the tutees did not detect the tutor song. It might thus be that the facilitating effect of the visual cues in addition to auditory song presentation would have been stronger in a noisier environment, in which the detection probability of tutor song would be lower. Likewise, for human speech, visual exposure to speakers' mouth movements contributes to speech intelligibility especially in noisy environments (Sumby and Pollack 1954; Middelweerd and Plomp 1987).

The results of Chapter 3, however, suggest that visual exposure to a tutor can affect song development. From this chapter it is unclear by which mechanism visual tutor exposure might have affected song learning. For instance, it is possible that the tutor gave visual feedback to tutee vocalizations. In other studies, visual feedback contingent on tutee vocalizations was found to improve zebra finch song development (Carouso-Peck and Goldstein 2019). It is, however, also possible that exposure to the visual cues accompanying song production, such as beak and throat movements, affected song learning. For instance, exposure to these visual cues might have drawn the tutee's attention to the auditory signal, as the detectability of a signal can be enhanced if it is presented at the same time as an additional stimulus in another sensory modality (Feenders, Kato, Borzeszkowski, & Klump 2017; reviewed in Rowe 1999). Likewise, in second language learning in human adults, audio-visual training, where mouth and lip movements associated with unfamiliar speech sounds are visible, improves the perception and production of these speech sounds more than audio-only training (e.g. Badin, Tarabalka, Elisei, & Bailly, 2010; Hazan, Sennema, Iba, & Faulkner, 2005; Hirata & Kelly, 2010; Liu, Massaro, Chen, Chan, & Perfetti, 2007; Wang, Hueber, & Badin, 2014). Unlike tape tutors, live tutors can provide visual feedback to tutee vocalizations and provide exposure to sound-production accompanying visual cues. This suggests that besides social tutor-tutee interaction, other mechanisms might play a role in the vocal learning process and might contribute to the difference in song learning suc-

cess from live and audio-only tutors.

Several songbird species learn less well from audio-only than from live social tutors and in many taxonomic groups, the simultaneous presentation of two stimuli in different modalities has been shown to improve signal perception compared to the presentation of one stimulus (reviewed in Rowe, 1999). This suggests that in general, audio-visual exposure to a vocalizing tutor might facilitate vocal learning compared to audio-only exposure. It is important to note, however, that not all songbird species learn less well from tape tutors than from live tutors (reviewed in Baptista & Gaunt, 1997). Future research could investigate whether there is a correlation between the ecology or song characteristics of different songbird species and whether these species learn less well from audio-only playback than from live tutors. For instance, as suggested by Slater et al. (1988), visual cues might be mainly of importance in species with quiet vocalizations, that can only be perceived when tutees are close to a tutor. This type of research might help in forming hypotheses concerning why certain species learn equally well from audio-only exposure to vocalizations as from live tutors, while others do not.

Suggestions for further research

During this research, I identified several open questions that I think should be addressed in further studies. First of all, in the tutoring studies described in this thesis, song learning success in the different tutoring conditions was assessed by determining the similarity between tutee and tutor song. This similarity was calculated using three different methods (visual spectrogram comparisons by human observers and similarity assessment by Luscinia and Sound Analysis Pro software), that all have previously been used to assess song learning success in zebra finches. Up till now, however, these three methods had not been used and compared with the same dataset. The results of the different methods were not very highly correlated, suggesting that the methods pick up different aspects of song similarity. Future research should look into these differences and aim to find out which method best represents sound similarity perception by zebra finches. In future song tutoring studies, that method should then be used to assess song learning success. This thesis mainly focussed on the effect of multi- or unimodal tutor exposure on the auditory component of song production. Future studies could investigate whether multi- or unimodal tutoring affects the visual component of song production. For instance, it could be assessed whether the previously found similarity between the beak movements of tutees and tutors (Williams, 2001) is affected by whether tutees had audio-only or audio-visual exposure to their tutor during the sensitive phase for song

learning.

Second, the artificial tutoring paradigms used in this thesis offer many possibilities for future research. However, first more research is needed into the effect of different methodological choices concerning these artificial tutors on song learning outcomes. For instance, the RoboFinch used in the robot tutoring experiment (Chapter 5) offers many possibilities for further research into multimodal communication and social interactions. The robotic zebra finch can also be used to study the process of multimodal integration in zebra finches, by offering a slight spatial or temporal mismatch between the auditory and visual information and investigating whether this affects zebra finch behaviour compared to a situation without a mismatch (as has been done in dart-poison frogs: Narins, Grabul, Soma, Gaucher, & Hödl, 2005 and pied currawongs: Lombardo, MacKey, Tang, Smith, & Blumstein, 2008).

This thesis focussed on the effect of visual cues on song production learning in male zebra finches. Song production learning only occurs in males, but both male and female zebra finches develop a preference for songs heard early in life over unfamiliar songs, no matter whether they have heard this song from a live (Riebel, Smallegange, Terpstra, & Bolhuis, 2002) or tape tutor (Holveck & Riebel, 2014; Houx & ten Cate, 1999a, b; Riebel, 2000). So far, however, no studies have investigated whether visual cues that are presented in addition to auditory song presentation affect song preference learning, for instance when it comes to the strength of the preference for a particular song. Carrying out the experiments described in this thesis with both male and female tutees, and assessing both song production and preference learning, can shed light on whether visual cues affect both processes equally.

Conclusions

To conclude, the studies in this thesis have demonstrated that multi- versus unimodal exposure to a live tutor can affect the timing of vocal development and possibly also the amount of vocal learning. Multimodal exposure to artificial tutors affected tutee behaviour and made stimulus presentation more salient, but did not affect the song learning outcomes assessed in the experiments in this thesis. These were, however, the first studies using these artificial tutors and future studies should, therefore, further investigate how properties of these artificial tutors affect song learning. I also found that song learning outcomes can be affected by the social environment in which tutees are housed during tutoring. Multi- versus unimodal tutoring and social housing versus social isolation during tutoring might have played a role in the difference in song learning out-

comes found in previous studies comparing live and tape tutoring paradigms. Future studies should be aware of the possible influences of multimodal tutor exposure and the social context on vocal development.

References

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

Badin P, Tarabalka Y, Elisei F, Bailly G (2010) Can you "read" tongue movements? Evaluation of the contribution of tongue display to speech understanding. Speech Commun 52:493–503. https://doi.org/10.1016/j.specom.2010.03.002

Baptista LF, Gaunt SLL (1997) Social interaction and vocal development in birds. In: Snowdon CT, Hausberger M (eds) Social influences on vocal development. Cambridge, Cambridge University Press, pp 23–40

Beecher MD, Burt JM (2004) The role of social interaction in bird song learning. Curr Dir Psychol Sci 13:224–228. https://doi.org/10.1111/j.0963-7214.2004.00313.x

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

Bolhuis JJ, Okanoya K, Scharff C (2010) Twitter evolution: Converging mechanisms in birdsong and human speech. Nat Rev Neurosci 11:747–759. https://doi.org/10.1038/nrn2931

Bradbury JW, Vehrencamp SL (2011) Principles of animal communication. Sinauer Associates, Sunderland

Carouso-Peck S, Goldstein MH (2019) Female social feedback reveals non-imitative mechanisms of vocal learning in zebra finches. Curr Biol 29:631–636. https://doi.org/10.1016/j.cub.2018.12.026

Catchpole CK, Slater PJB (1995) How song develops. In: Catchpole CK, Slater PJB (eds) Bird Song: Biological Themes and Variations. Cambridge: Cambridge University Press., pp 45–69

Chen Y, Matheson LE, Sakata JT (2016) Mechanisms underlying the social enhancement of vocal learning in songbirds. Proc Natl Acad Sci 201522306. https://doi.org/10.1073/pnas.1522306113

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. J Physiol 107:210–218. https://doi.org/10.1016/j.jphyspar is.2012.08.003

Doupe AJ, Kuhl PK (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 LA (1989) The influences of visual and vocal interaction on song learning in zebra finches. Anim Behav 37:507–508. https://doi.org/10.1016/0003-3472(89)90097-3

Feenders G, Kato Y, Borzeszkowski KM, Klump GM (2017) Temporal ventriloquism effect in european starlings: evidence for two parallel processing pathways. Behav Neurosci 131:337–347. https://doi.org/10.1037/bne0000200

Gobes SMH, Jennings RB, Maeda RK (2017) The sensitive period for auditory-vocal learning in the zebra finch: consequences of limited-model availability and multiple-tutor paradigms on song imitation. Behav Processes 163:5–12. https://doi.org/10.1016/j.beproc.2017.07.007

Goldstein MH, King AP, West MJ (2003) Social interaction shapes babbling: testing parallels between birdsong and speech. Proc Natl Acad Sci U S A 100:8030–5. https://doi.org/10.1073/pnas.1332441100

Hazan V, Sennema A, Iba M, Faulkner A (2005) Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. Speech Commun 47:360–378. https://doi.org/10.1016/j.specom.2005.04.007

Hirata Y, Kelly SD (2010) Effects of lips and hands on auditory learning of second-language speech sounds. J Speech Lang Hear Res 53:298–310. https://doi.org/10.1044/1092-4388(2009/08-0243)

Holveck MJ, Riebel K (2014) Female zebra finches learn to prefer more than one song and from more than one tutor. Anim Behav 88:125–135. https://doi.org/10.1016/j.anbehav.2013.11.023

Houx BB, ten Cate C (1999a) Do stimulus-stimulus contingencies affect song learning in zebra finches (*Taeniopygia guttata*)? J Comp Psychol 113:235–242. https://doi.org/10.1037/0735-7036.113.3.235

Houx BB, ten Cate C (1999b) Song learning from playback in zebra finches: is there an effect of operant contingency? Anim Behav 57:837–845. 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. Anim Behav 58:143–149. https://doi.org/10.1006/anbe.1999.1120

Kuhl PK (2007) Is speech learning "gated" by the social brain? Dev Sci 10:110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kuhl PK (2003) Human speech and birdsong: communication and the social brain. Proc Natl Acad Sci U S A 100:9645–9646. https://doi.org/10.1073/pnas.1733998100

Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. Science (80-) 218:1138–1141. https://doi.org/10.1126/science.7146899

Liu Y, Massaro DW, Chen TH, et al (2007) Using visual speech for training chinese pronunciation: an in-vivo experiment. SLaTE Work Speech Lang Technol Educ ISCA Tutor Res Work Summit Inn, Farmington, Pennsylvania USA 29–32

Lombardo SR, MacKey E, Tang L, et al (2008) Multimodal communication and spatial binding in pied currawongs (*Strepera graculina*). Anim Cogn 11:675–682. https://doi.org/10.1007/s10071-008-0158-z

Middelweerd MJ, Plomp R (1987) The effect of speechreading on the speech-recep-

tion threshold of sentences in noise. J Acoust Soc Am 82:2145–2147. https://doi.org/10.1121/1.395659
Narins PM, Grabul DS, Soma KK, et al (2005) Cross-modal integration in a dart-poison frog. Proc Natl Acad Sci U S A 102:2425–2429. https://doi.org/10.1073/pnas.0406407102
Nelson D (1997) Social interaction and sensitive phases for song learning: A critical review. In: Snowdon CT, Hausberger M (eds) Social influences on vocal development. Cambridge, Cambridge University Press, pp 7–22
Patterson ML, Werker JF (1999) Matching phonetic information in lips and voice is robust in 4.5-month-old infants. Infant Behav Dev 22:237–247. https://doi.org/10.1016/S0163-6383(99)00003-X
Ręk P (2018) Multimodal coordination enhances the responses to an avian duet. Behav Ecol 29:411–417. https://doi.org/10.1093/beheco/arx174
Riebel K (2000) Early exposure leads to repeatable preferences for male song in female zebra finches. Proc R Soc London Ser B Biol Sci 267:2553–8. https://doi.org/10.1098/rspb.2000.1320
Riebel K, Smallegange IM, Terpstra NJ, Bolhuis JJ (2002) Sexual equality in zebra finch song preference: evidence for a dissociation between song recognition and production learning. Proc R Soc London Ser B Biol Sci 269:729–33. https://doi.org/10.1098/rspb.2001.1930
Rowe C (1999) Receiver psychology and evolution of multicomponent signals. Anim Behav 58:921–931. https://doi.org/10.1006/anbe.1999.1242
Slater PJB, Eales LA, Clayton NS (1988) Song learning in zebra finches (*Taeniopygia guttata*): progress and prospects. Adv study Behav 18:1–34. https://doi.org/10.1016/S0065-3454(08)60308-3
Soma MF (2011) Social factors in song learning: a review of Estrildid finch research. Ornithol Sci 10:89–100. https://doi.org/10.2326/osj.10.89
Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215. https://doi.org/10.1121/1.1907309
Taylor RC, Klein BA, Stein J, Ryan MJ (2011) Multimodal signal variation in space and time: how important is matching a signal with its signaler? J Exp Biol 214:815–820. https://doi.org/10.1242/jeb.043638
van Kampen HS, Bolhuis JJ (1993) Interaction between auditory and visual learning during filial imprinting. Anim Behav 45:623–625. https://doi.org/10.1006/anbe.1993.1074
van Kampen HS, Bolhuis JJ (1991) Auditory learning and filial imprinting in the chick. Behaviour 117:303–319. https://doi.org/10.1163/156853991X00607
Wang X, Hueber T, Badin P (2014) On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French. 10th Int Semin Speech Prod 449–452

Nederlandse samenvatting

‘Seeing’ voices: de rol van multimodale kenmerken in vocaal leren.

Zowel mensen als zangvogels leren hun soortspecifieke vocalisaties (spraak bij mensen en zang bij zangvogels) vroeg in hun leven van volwassen soortgenoten. Bij mensen en verschillende zangvogelsoorten hebben jonge individuen meer moeite met het leren van de vocalisaties wanneer ze deze alleen kunnen horen, bijvoorbeeld via luidsprekers, dan wanneer ze deze horen terwijl ze sociale interactie kunnen aangaan met een vocaliserende soortgenoot. Vaak wordt op basis hiervan geconcludeerd dat sociale interactie met een volwassen soortgenoot belangrijk is voor het leren van vocalisaties. Het is echter onduidelijk of vocaal leren van een sociale leermeester ook gemakkelijker is doordat individuen een sociale leermeester zowel kunnen horen als zien, in plaats van alleen horen.

Verschillende studies hebben aangetoond dat het tegelijkertijd aanbieden van stimuli in verschillende sensorische modaliteiten een positief effect kan hebben op de perceptie ervan. Bij zowel spraak als vogelzang worden gelijktijdig meerdere zintuigen gestimuleerd. In beide gevallen is sprake van een auditief signaal, maar ook van visuele kenmerken in de vorm van bijvoorbeeld lip- of snavelbewegingen. Dit maakt spraak en vogelzang multimodale signalen, oftewel signalen die via meerdere zintuigsystemen waargenomen kunnen worden. Observationeel onderzoek in mensen suggereert dat het zowel kunnen zien als horen van een spreker een positief effect kan hebben op de spraakontwikkeling en voor zangvogels zijn aanwijzingen gevonden dat visuele stimulatie een effect kan hebben op de zangontwikkeling.

Zangvogels, en met name zebravinken, worden vaak als modelsysteem gebruikt voor spraakontwikkeling, omdat vogelzang- en spraakontwikkeling veel overeenkomsten vertonen. Voor zebravinken is het gemakkelijker om zang te leren die ze horen van een soortgenoot waar ze sociale interactie mee aan kunnen gaan dan om zang te leren die ze alleen kunnen horen, maar het is nog onduidelijk welk effect het kunnen zien van een soortgenoot heeft op het leren van zang. In dit proefschrift is experimenteel onderzocht of audiovisuele, vergeleken met alleen auditieve, blootstelling aan een tutor een positief effect heeft op de zangontwikkeling bij zebravinken, onafhankelijk van een mogelijk effect van sociale interactie.

In Hoofdstuk 2 van dit proefschrift is bekeken of eerdere zangleerstudies met zebravinken de hypothese ondersteunen dat multimodale ten opzichte van unimodale blootstelling aan een tutor het zangleren bevordert. In dit hoofdstuk

werd duidelijk dat de meeste studies concluderen dat vooral sociale interactie belangrijk is voor vocaal leren. Deze conclusie is gebaseerd op studies waarin een situatie met een sociale tutor wordt vergeleken met een situatie waarin zang alleen auditief wordt aangeboden en waaruit blijkt dat zebravinken beter leren van de sociale tutor. Het is echter onduidelijk in hoeverre dit ermee te maken heeft dat de sociale tutor zowel gezien als gehoord kan worden, terwijl in de andere situatie zang alleen gehoord kan worden. Daarnaast bleek in dit hoofdstuk dat dit verschil in zangleersucces ook deels ermee te maken zou kunnen hebben dat de vogels die alleen auditief aan zang werden blootgesteld, in sociale isolatie werden gehuisvest tijdens het zangleren, terwijl de vogels met de sociale tutor gezelschap hadden van de tutor.

In Hoofdstuk 3 is een zangleerstudie beschreven waarin jonge vogels werden gehuisvest in een van drie verschillende condities: in dezelfde kooi als een volwassen tutor, in een aangrenzende kooi waar ze deze zelfde tutor konden horen door luidsprekerdoek of in een aangrenzende kooi waar ze de tutor konden horen, maar ook konden zien door een eenrichtingsspiegel. In deze laatste conditie konden de jonge vogels de tutor dus horen en zien, maar was geen visuele sociale interactie mogelijk, omdat de tutor de jonge vogels niet kon zien. Alle jonge vogels waren gehuisvest met een vrouwtje, zodat geen van de vogels in sociale isolatie opgroeide. De zang van de jonge vogels werd opgenomen toen zij 65 dagen oud waren (wanneer de zang van zebravinken nog in ontwikkeling is) en toen ze ouder dan 100 dagen waren (wanneer de zang normaal gesproken ontwikkeld is). De zang van de vogels die alleen auditief aan de tutor werden blootgesteld, vertoonde meer veranderingen na 65 dagen dan de zang van de vogels met multimodale blootstelling aan de tutor. Dit suggereert dat multimodale blootstelling aan een tutor invloed kan hebben op het verloop van de zangontwikkeling. Daarnaast leken de vogels die in dezelfde kooi als de tutor opgroeiden, het meest van de tutor geleerd te hebben en de vogels die alleen auditief aan de tutor waren blootgesteld, het minst. De vogels met audiovisuele tutorblootstelling leken hier tussenin te zitten. Hoewel deze resultaten suggereren dat het kunnen zien en horen van een tutor een positief effect heeft op het zangleren, is het bijvoorbeeld ook mogelijk dat de resultaten voortkomen uit visuele feedback die de tutor gaf op de vocalisaties van de jonge vogels en die alleen waargenomen kon worden door de vogels die de tutor konden zien. In de volgende hoofdstukken werd daarom gebruikgemaakt van artificiële tutors om het effect van het kunnen zien en horen van een tutor te onderzoeken los van mogelijke effecten van visuele feedback van de tutor.

In Hoofdstuk 4 is onderzocht of het leren van vooraf opgenomen tutorzang

die door luidsprekers wordt afgespeeld, vergemakkelijkt wordt wanneer tegelijkertijd video's van de zingende tutor worden afgespeeld. Deze video's waren aangepast aan het visuele systeem van zebravinken, dat op verschillende punten afwijkt van dat van mensen. In dit hoofdstuk werd zang ofwel alleen auditief, ofwel in combinatie met de video van de tutor aangeboden. Daarnaast was er een conditie waarin zang werd afgespeeld in combinatie met dezelfde video, maar dan gepixeleerd en in omgekeerde volgorde afgespeeld. Het doel van deze conditie was te onderzoeken of niet-sociale en niet-specifieke visuele stimulatie een ander effect heeft op het leren van zang dan sociale, geluidspecifieke visuele stimulatie. De jonge vogels brachten veel tijd door in de buurt van de video's en vonden de audiovisuele stimulatie met de normale video's in dit opzicht aantrekkelijker dan de audiovisuele stimulatie met de gepixeleerde video's en de auditieve stimulatie. Er werd echter geen effect van de verschillende condities op zangleersucces gevonden. Dit zou ermee te maken kunnen hebben dat een video slechts een tweedimensionale representatie van de tutor geeft. Dit beïnvloedt mogelijk het effect van visuele stimulatie op het zangleren. In Hoofdstuk 5 is daarom gebruikgemaakt van een driedimensionale robot als tutor.

In Hoofdstuk 5 werd gebruikgemaakt van zogenoemde 'RoboFinches', ge-3D-printe robotzebravinken die hun snavel en hoofd kunnen bewegen. In één conditie werden jonge zebravinken auditief blootgesteld aan vooraf opgenomen tutorzang, terwijl een RoboFinch tegelijkertijd de bijbehorende snavel- en hoofdbewegingen maakte. Deze conditie werd vergeleken met een conditie waarin zang alleen auditief werd aangeboden en een conditie waarbij de RoboFinch pas begon te bewegen nadat het afspelen van de tutorzang was afgelopen. De vogels in deze condities werden gehuisvest in sociale isolatie. Om te onderzoeken of dit een negatief effect had op het zangleren, was er ook een conditie waarin de jonge vogels alleen auditief werden blootgesteld aan de tutorzang, terwijl ze gehuisvest waren met een vrouwtje. De verschillende tutorcondities bleken geen effect te hebben op de hoeveelheid zang die de jonge vogels leerden van de tutor. De mannetjes die in sociale isolatie gehuisvest waren terwijl ze alleen auditief de zang aangeboden kregen, bleken met minder stereotypie te zingen dan de mannetjes die een vrouwtje als gezelschap hadden terwijl ze auditief aan de tutorzang werden blootgesteld. Dit suggereert dat de sociale omgeving waarin jonge vogels zich bevinden tijdens de sensitieve periode voor zang leren, een effect heeft op de zangontwikkeling.

De resultaten van Hoofdstuk 3 ondersteunen de hypothese dat multimodale blootstelling aan een tutor een positief effect heeft op zang leren. Dit is niet het geval voor de resultaten van Hoofdstuk 4 en 5. Hierbij moet echter opgemerkt

worden dat dit de eerste keer is dat deze artificiële tutors werden gebruikt in een zangleerstudie. Vervolgstudies zouden moeten onderzoeken hoe verschillende methodologische keuzes met betrekking tot deze tutors het zangleren beïnvloeden. Hierbij kan bijvoorbeeld gedacht worden aan de hoeveelheid zang die aangeboden wordt. Daarnaast kon in dit proefschrift niet vastgesteld worden of de leercondities in Hoofdstuk 4 en 5 een effect hadden op de mate waarin de zang veranderde tussen dag 65 en dag 100, omdat de zang van de vogels in deze experimenten met 65 dagen nog zeer variabel was. Het is dus mogelijk dat de leercondities in deze experimenten geen effect hadden op de hoeveelheid zang die de jonge vogels kopieerden van de tutor, maar wel op het verloop van hun zangontwikkeling.

Geconcludeerd kan worden dat de experimentele studies in dit proefschrift hebben aangetoond dat multi- ten opzichte van unimodale blootstelling aan een levende tutor de timing en mogelijk de mate van vocaal leren bij zebravinken beïnvloedt. Multimodale blootstelling aan een artificiële tutor had een effect op het gedrag van jonge vogels en maakte de stimuluspresentatie aantrekkelijker voor de jonge vogels, maar had geen effect op de zangleerparameters die in dit proefschrift onderzocht werden. Zangontwikkeling bleek beïnvloed te worden door de sociale omgeving waarin de jonge vogels zich bevonden tijdens de sensitieve periode voor zang leren. Eerdere zangleerstudies in zebravinken hebben vaak geen rekening gehouden met een mogelijk effect van multi- versus unimodale tutorblootstelling en sociale huisvesting versus sociale isolatie tijdens het zangleren, maar dit proefschrift heeft aangetoond dat deze factoren de zangontwikkeling kunnen beïnvloeden.

Acknowledgements

First of all, I would like to thank my supervisor, Katharina Riebel, for her support, encouragement and enthusiasm. I would also like to thank my promotor, Carel ten Cate, for his support and helpful advice.

I want to thank Ralph Simon for all his work on this project and for the fun we had while working together. I also want to thank the other members of the 'seeing voices' consortium, Wouter Halfwerk, Ezequiel Mendoza and Constance Scharff, for the interesting and insightful discussions we had about the experiments. Thanks to Dré Kampfraath, Rogier Elsinga, Peter Wiersma and Wesley Delmeer for their help with the RoboFinch.

A big thank you to all students who were involved in my PhD project: Bas, Brandon, David, Eva, Esmee, Femke, Idse, Jens, Kamiel, Maëva, Qiaoyi and Rozanda. Thanks for thinking along and for the effort that you put into this project.

Many thanks to my colleagues at the behavioural biology group, especially Annebelle, Fabian, Hans, Inge, Jeroen, Jing, Meike, Merel, Michelle, Ning and Temp, for always being so interested and encouraging. I want to thank Peter Snelderwaard for his technical assistance and assistance with animal care and I want to thank the animal care takers: Michelle Chan, Michelle Geers and Roy van Swetselaar.

Thanks to my parents, my brother and his girlfriend, my family and my friends for their support and for distracting me from work every now and then. Finally, thank you Daniël for supporting me in everything I do, which even meant spending your time off building bird cages and preparing experimental set-ups with me. Thank you as well for being my work-from-home colleague during the covid lockdown and for always looking after my mental wellbeing.

Curriculum vitae

Judith Varkevisser was born in 1992 in Leidschendam, the Netherlands. From 2004 to 2010, she followed a vwo-gymnasium program at College het Loo in Voorburg. In 2013, she obtained her Bachelor's degree (cum laude) in linguistics at Leiden University. During her Bachelor, she did an internship at Stichting Plotsdoven, studying video glasses as a means of communication for post-lingually deafened adults. In 2015, she completed a research Master in linguistics (cum laude) at Leiden University. During her Master, she worked as a student assistant at the Leiden University Centre of Linguistics in a project on prosody in whispered speech and in a project on speech production and perception in Brazilian learners of English. She did an internship at the Institute of Biology Leiden studying acoustic changes in the crystallized song of zebra finches. For her Master thesis, she conducted a comparative analysis of sex differences in birdsong.

In 2016, Judith started her PhD trajectory as part of the HFSP funded project 'Seeing voices: the role of multimodal cues in vocal learning' at the Institute of Biology Leiden supervised by dr. Katharina Riebel and prof. dr. Carel ten Cate. During her PhD trajectory, she participated in teaching courses on behavioural biology and supervised several student projects. Judith is currently working as a lecturer and developer of teaching material on research methodology and statistics at the University of Amsterdam, as a student writing coach at the Hogeschool Rotterdam and as a self-employed student writing coach and thesis editor.

Publications

Varkevisser, J., Mendoza, E., Simon, R. et al. (2022). Multimodality during live tutoring is relevant for vocal learning in zebra finches. Animal Behaviour, 187, 263-280. https://doi.org/10.1016/j.anbehav.2022.03.013

Varkevisser, J., Simon, R., Mendoza, E. et al. (2021). Adding colour-realistic video images to audio playbacks increases stimulus engagement but does not enhance vocal learning in zebra finches. Animal Cognition, 25, 294-274. https://doi.org/10.1007/s10071-021-01547-8

Simon, R., **Varkevisser, J.,** Mendoza, E., Hochradel, K., Scharff, C., Riebel, K., & Halfwerk, W. (2019). Development and application of a robotic zebra finch (RoboFinch) to study multimodal cues in vocal communication. PeerJ Preprints 7:e28004v2 https://doi.org/10.7287/ peerj.-preprints.28004v2

Halfwerk, W., **Varkevisser, J.,** Simon, R., Mendoza, E., Scharff, C., & Riebel, K. (2019). Towards testing for multimodal perception in mating signals. Frontiers in Ecology and Evolution, 7, 124.

Post da Silveira, A., & **Varkevisser, J.** (2019). Sex differences in vocalic duration production in L1 and in L2. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.) Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019 (pp. 3358-3362).