Title: Attentional coordination in demonstrator-observer dyads facilitates learning and predicts performance in a novel manual task

Murillo Pagnotta[a], Kevin N. Laland[a], Moreno I. Coco[b,c]

a Centre for Social Learning and Cognitive Evolution, School of Biology, University of St Andrews, St Andrews, UK

b Faculdade de Psicologia, Universidade de Lisboa, Lisbon, Portugal

c School of Psychology, University of East London, London, UK

Abstract:

Observational learning is a form of social learning in which a demonstrator performs a target task in the company of an observer, who may as a consequence learn something about it. In this study, we approach social learning in terms of the dynamics of coordination rather than the more common perspective of transmission of information. We hypothesised that observers must continuously adjust their visual attention relative to the demonstrator's time-evolving behaviour to benefit from it. We eye-tracked observers repeatedly watching videos showing a demonstrator solving three manipulative puzzles before attempting at the task. The presence of the demonstrator's face and the availability of his verbal instruction in the videos were manipulated. We then used recurrence quantification analysis to measure the dynamics of coupling between the overt attention of the observers and the demonstrator's manipulative actions. Bayesian regression was applied to examine whether the observers' performance was predicted by such indexes of coordination, how performance changed as they accumulated experience, and if the availability of speech and intentional gaze of the demonstrator mediated it. Results showed that learners better able to coordinate their eye movements with the manipulative actions of the demonstrator had an increasingly higher probability of success in solving the task. The availability of speech was beneficial to learning, whereas the presence of the demonstrator's face was not. We argue that focusing on the dynamics of coordination between individuals may greatly improve understanding of the cognitive processes underlying social learning.

Keywords: observational learning; attentional synchronization; eye tracking; recurrence quantification analysis; Bayesian regression

Introduction

Throughout their lives, humans and nonhuman animals learn to perceive their surroundings and engage more or less skilfully with the different tasks they encounter. Within the behavioural sciences, a common distinction is made between individual (or asocial) learning and social learning (Galef, 1988; Heyes, 1994; Hoppitt & Laland, 2013; Whiten & Ham, 1992; Whiten, Horner, Litchfield, & Marshall-Pescini, 2004). The latter is defined as "learning that is facilitated by observation of, or interaction with, another individual (or its products)" and encompasses a wide range of processes (Hoppitt & Laland, 2013).

Here we focus on observational learning (a.k.a. 'production imitation'), which occurs when an observer acquires an action, or action sequence, after watching another individual perform it (Ashford, Bennett, & Davids, 2006; Carcea & Froemke, 2019; see Hoppitt & Laland, 2013, p. 4 and p. 64 for precise definitions). This type of learning occurs in formal settings such as in schooling, sports training, and apprenticeship, and it usually involves a 'demonstrator' (or 'model') and a 'learner' (or 'observer'). The demonstrator shows the learner the correct or normative way of performing the target task, either intentionally or unintentionally. The learner observes the demonstration and attempts the task. In this context, the dynamics of joint attention that underlies the execution and observation of the task may facilitate the development of the skills required to complete it effectively, as we argue below.

Our perspective is supported by the influential work of Tomasello and collaborators (Carpenter, Nagell, & Tomasello, 1998; Carpenter & Tomasello, 1995; Tomasello 1999, 2009; Tomasello, Kruger, & Ratner, 1993), who maintain that joint attention is critical to human social learning and social cognition. These authors suggest that both teaching and collaborative learning are critically reliant on human's ability to alternate perspective taking and to attend jointly to objects and events with others. Joint attention is thought to underlie the unique aspects of our species' social cognition skills, differentiating humans from other apes (Carpenter & Tomasello, 1995; Tomasello, 2009), scaffolding language learning and cognitive development (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998; Degotardi, 2017; Tomasello, 2003, 2009), and being a key deficit of individuals with autism spectrum disorders (Schertz, Odom, Baggett, & Sideris, 2013).

Observational learning has been extensively investigated in the context of motor control to understand, for example, how humans learn novel sequences of existing movement patterns (Bird & Heyes, 2005; Nissen & Bullemer, 1987), rhythmic patterns (Vogt, 1995), interlimb or whole-body coordination patterns (Casile & Giese, 2006; Hodges, Williams, Hayes, & Breslin, 2007), and how to adjust limb movements in novel environments (Mattar & Gribble, 2005). Given its intimate link with learning action sequences, observational learning has received considerable attention in the sport sciences; for example, to assess the effectiveness of demonstrations in facilitating skill acquisition (Horn, Williams, Hayes, Hodges, & Scott, 2007; Horn, Williams, Scott, & Hodges, 2005; Williams & Hodges, 2005).

Some of these studies have also examined the role played by overt attention during observational learning. (e.g., Breslin, Hodges, & Williams, 2009; D'Innocenzo, Gonzalez, Williams, & Bishop, 2016; Horn et al., 2005). For example, Breslin and colleagues (2009) examined how attending to different parts of the body of a demonstrator performing a novel cricket bowling action mediates how the action is acquired by the learners. Participants in this study underwent three practice blocks in which they first watched a demonstration video – which consisted of a point-light display film showing either the demonstrator's bowling arm, or his wrists, or his full body – five times and then had ten trials to replicate the action. On the following day, after a retention test, participants practiced another three blocks now watching the full-body point-light display film; and an additional retention test was performed on the third day. Measures of intralimb and interlimb coordination were used to compare the performance of learners with the demonstrator, and eye-tracking was used to examine learners' visual attention to the demonstration videos. When watching the full-body film, participants focused more on the bowling arm than on other body parts (e.g., the legs) suggesting learners prioritize the end effector of the action during observational learning. Most importantly, participants who saw the demonstrator's bowling arm on both days acquired an intralimb coordination profile more similar to the demonstrator compared to participants who saw his bowling arm only on day 2. Despite showing a very interesting relation between overt attention and task performance, this study did not explicitly assess it as the measures of overt attention used were aggregated over the entire trial (e.g., proportion of time spent on each area of interest), and thus they were unable to capture the dynamics of overt attention on a moment-by-moment basis. This aspect is at heart of the current study, which will examine precisely how learners must dynamically adapt their visual attention in order to stay 'in touch' (i.e. informationally coupled through active perception) with the relevant aspects of the task as they move in space and change over time; and how this attentional coordination is critically related to their task success.

To the best of our knowledge, only few studies have formally examined the association between overt attention and learning outcome, and these do not come from the field of social learning. Eye-movement coordination between speakers and listeners was, for example, found to be positively associated with discourse comprehension (Richardson & Dale, 2005), and emerge as a positive predictor of task success only when interlocutors could engage in a bi-directional conversation (Coco, Dale, & Keller, 2018). Other eye-movement studies have attempted to direct the learners' attention to specific aspects of the task by manipulating the saliency of visual stimuli and examined its effect on learning. Grant and Spivey (2003), for example, found that more learners arrived at the correct solution of a diagram-based insight task when presented with a diagram which highlighted a critical area, compared to a static diagram or a diagram which highlighted a non-critical area.

However, intentionally directing the observer's attention towards task-relevant aspects does not always facilitate learning (see van Gog, Jarodzka, Scheiter, Gerjets, & Paas, 2009, for counterevidence), which indicates that the relation between attentional coordination and performance may strongly depend on the demands of the task at hand and the specific context of demonstrator-observer interaction. Even if researchers in the field of social learning recognize the importance of joint attention, it is yet to be rigorously demonstrated that the time-evolving dynamics of coordination between demonstrators and learners are indeed predictive of their learning pattern.

This approach is in line with the growing body of literature in the cognitive sciences arguing that behaviour and human interaction can be framed as multi-scale, self-organizing and dynamical phenomena (Chemero, 2009; Dale, Fusaroli, Duran, & Richardson, 2013; De Jaegher & Di Paolo, 2007; Haken, Kelso, & Bunz, 1985; Kelso, 1995, 2016; Schoner & Kelso, 1988; Schoner, Zanone, & Kelso, 1992). Important advances in the study of multi-modal coordination have, in fact, been possible through the application of non-linear methods of analysis such as recurrence quantification analysis (*RQA*) which can be used to quantify the temporal dynamics of two or more streams of data underlying human interaction, such as manipulative actions and eye-movement (Coco et al., 2017; Coco & Dale, 2014; Fusaroli, Konvalinka, & Wallot, 2014; Richardson, Dale, & Marsh, 2014; Wallot, Mitkidis, McGraw, & Roepstorff, 2016).

In the current study, we take inspiration from dynamical systems theory and borrow some of their methodological tools to examine social learning. We combined eye-tracking, *RQA*, and Bayesian hierarchical logistic regression analysis to investigate how learning rate in a novel manipulative task may depend on the patterns of attentional coordination that arise when learners watch a demonstrator performing task-specific actions. Learners were eye-tracked as they watched videos of a demonstrator showing them how to solve a manipulative construction puzzle (our target task, see Figure 1) and then attempted to solve the same puzzle on their own. Rather than running a single trial, we asked learners to watch the demonstration video and attempt the corresponding puzzle multiple times, so that we might monitor changes in their performance as a function of their accumulated experience.

We hypothesised that learners must adjust their overt attention dynamically and synchronously to the demonstrator's unfolding behaviour to benefit from it maximally. Specifically, we expected that if learners systematically time-locked their overt attention to the pieces being manipulated by the demonstrator, they might detect relevant aspects of the demonstration, such as the actions required to orderly and correctly assemble the pieces into the final structure. Thus, we predicted that higher attentional coordination of the learners to the manipulative actions of the demonstrator would result into increasingly better learning outcomes.

We acknowledge that the use of pre-recorded demonstrations imply that learners may dynamically adapt their allocation of overt attention to the manipulative actions displayed in the videos, but the demonstrator would always perform the same sequence of actions, and so, there is no dynamical interaction between the demonstrator and the learner. Hence, our use of the expressions `attentional coordination` or `synchronisation` must be interpreted as unidirectional (i.e., only the learner can dynamically adapt to the demonstrator).

Another important aspect of an intentional demonstration is gaze following, which is considered central to establishing and sustaining joint attention (e.g., Carpenter et al., 1998; Tomasello, Carpenter, Call, Behne, & Moll, 2005). However, it is also known that people shift their overt attention to objects just before reaching them and tend to look at them until the movement is completed (Johansson, Westling, Backstrom, & Flanagan, 2001; Land & Hayhoe, 2001). Thus, in the context of object manipulation, the objects being looked at may coincide with the objects being manipulated. This suggests that, during a manipulative task, joint attention could be achieved by either following the partner's gaze (the conventional gaze-following route) or the partner's hands (hand-eye coordination route).

Yu and Smith (2013), for example, provided eye-tracking evidence for this alternative route to joint attention by examining the attentional coordination of one-year-old children and their parents while playing together with toys. Given that seeing the partner's face might help direct one's own visual attention, and given that learning through (live or recorded) demonstration requires coordinating one's visual attention with the demonstrator, we examined whether the presence of the intentional gaze of the demonstrator helped (or not) to direct the attentional coordination of the learners and, especially, whether it improved (or not) their performance in the construction puzzle task. If gaze following is indeed required to establish joint attention, then we should expect that observers who could see the demonstrator's face (and thus could follow his gaze throughout the demonstration) would learn faster than those that could not (see Figure S1 in the electronic supplementary material for an example of the gaze manipulation and refer to demonstration videos available in the Open Science Framework at https://osf.io/jhtqb/). Conversely, if gaze following is not required for joint attention, then we should expect that observers seeing the demonstrator's face would not benefit from it as compared to those that did not see it.

The final aspect of an intentional demonstration on which our study focuses is that learners may or may not receive verbal instructions from the demonstrator. Psycholinguistics research has provided compelling evidence that sentence processing is tightly linked with other cognitive modalities such as visual attention: speakers tend to look at those objects that correspond with the words being spoken (Coco & Keller, 2012, 2015; Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998), and listeners also tend to look at those objects that correspond with the words being heard (Allopenna, Magnuson, & Tanenhaus, 1998; Coco, Keller, & Malcolm, 2016; Knoeferle & Crocker, 2006; Richardson & Dale, 2005). Moreover, systematic links between verbal and non-verbal (e.g., eye movement) behaviour extends to communicative dialogue, where speakers and listeners dynamically adapt their actions and vocalizations to the conversational partner as they go along in the dialogue (Clark & Krych, 2004; Fogel, 1993), and may even synchronize their eye-movement behaviour over time (Richardson, Dale, & Kirkham, 2007).

This literature clearly shows that listening to verbal communication can have a direct impact on one's visual attention, as well as on task performance. We therefore examined the impact of the demonstrator's verbal instruction on the learners' attentional coordination and on their performance at assembling the puzzle. Given the suggested role of speech in guiding the attention of listeners (e.g., Ingold, 2001; Tomasello, 2003), we predicted that learners who could listen to the demonstrator would learn faster than those ones that could not.

1. Methods

1.1. Design

We used a mixed factorial design with the type of demonstration video manipulated as a between-participant variable and with 3 repeated measures of task per participant and 5 repeated measures of iteration per task. Specifically, we crossed the visibility of the demonstrator's face (face visible or face blurred) with the availability of the demonstrator's verbal instructions (with audio or no audio), to produce four experimental conditions: face blurred and no audio (noFACE_noAUDIO); face visible and no audio (FACE_noAUDIO); face blurred with audio (noFACE_AUDIO); and face visible with audio (FACE_AUDIO). In addition, to discriminate between 'social' and 'individual' learning we ran two control conditions in which learners only saw a still image of the demonstrator and the puzzle pieces and could therefore not benefit from seeing his manipulative actions. In one condition, the still image was accompanied by the audio of the corresponding demonstration (noVIDEO_AUDIO) and hence learners could only benefit from the demonstrator's verbal instructions. In the other condition, the still image was shown without the audio (noVIDEO_noAUDIO), thus learners could not benefit in any way from the behaviour of the demonstrator. We report these two control conditions in the electronic supplementary material, as they were not central to the main arguments of our study.

Participants were randomly allocated to one of the six conditions and performed all three versions of the task (star, egg, and barrel). The order of the puzzles was counterbalanced between participants. At the start of each puzzle, the participants were asked to complete the puzzle without any instruction to obtain a baseline measure. They repeated the puzzle another five times, but each time they first watched the demonstration video before attempting the puzzle. This iterative procedure gives us repeated measures of performance (baseline plus 5), which could be used to construct a learning curve rather than a one-off success/failure outcome (see below for further details about how the data was modelled).

1.2. Participants

Fifty-three participants (32 female; age: range = [18, 50], median = 21, SD = 5.4) were recruited using the Experimenter Volunteer Panel of the University of Edinburgh. Forty participants did the four experimental conditions explained above and reported in what follows. Thirteen participants instead did the control conditions and, as mentioned, are reported only in the electronic supplementary material. All participants gave informed consent, had normal or corrected-to-normal vision, indicated no known learning disability, and were paid £7 as compensation for their time.

In addition, an experienced schoolteacher in Edinburgh (male, 33 years of age) was recruited to perform the role of the demonstrator in the video recordings used as stimuli and received £20 for his time. Prior to data collection, the study was approved by the University of St Andrews Teaching and Research Ethics Committee and by the Psychology Research Ethics Committee of the University of Edinburgh, in accordance with the British Psychological Society guidelines on ethics.

1.3. Material

The manipulative task was to solve construction puzzles, that is, to assemble sets of wooden pieces to form pre-defined structures. Each participant engaged with three puzzles (star, egg, and barrel, see Figure 1), which differed in the number of pieces (star: six pieces; egg: eight pieces; barrel: twelve pieces) and in the steps required to solve them. In the videos, the demonstrator shows and verbally describes the steps needed to assemble the different structures. The experimenter and the demonstrator scripted the verbal instructions beforehand so that the language used was standardised across the three puzzles (transcriptions of the verbal instructions are available in section 6 of the electronic supplementary material, and examples of the demonstration videos are available in the Open Science Framework page of this project).

A tripod-mounted camera positioned at eye level in front of the demonstrator was used to record the videos. The demonstrator was instructed to act naturally and to look at the camera from time to time, as if he were teaching an imaginary learner in front of him. The videos were captured in the portrait orientation and a lapel microphone was used to record the demonstrator's speech. Because the puzzles differed in the number of pieces, the demonstrations differed in duration (star: 40s, egg: 54s, barrel: 78s). We edited the videos to obtain the versions corresponding to the experimental conditions described above (i.e., face visible/face blurred, with audio/without audio) using the Wondershare Filmora software.

1.4. Experimental setup

Participants watched the videos while being eye-tracked on one desk and assembled the puzzles on another desk (see Figure 1B for a visualization of the workspace). They could easily move between the two desks by rotating 90 degrees on the chair. Videos were displayed on a 21’’ monitor in portrait orientation with a resolution of 1050 x 1680 pixels at a refresh rate of 100 Hz and a frame rate of 25 Hz. The audio was played on standard desktop speakers.

Eye-movements were tracked using a SR Research EyeLink 1000 with Desktop Mount at a sampling rate of 1000Hz. We only tracked the dominant eye, which was assessed using a parallax test. A forehead-and-chin rest was used to stabilize the participant’s head movement. The monitor covered 35 degrees of visual angle vertically and 22 degrees horizontally, and the distance between the headrest and the top of the monitor was 74 cm. Nine-point calibration routines were performed before watching the video for the first time for each puzzle, and a drift check was performed before each subsequent attempt. Experiment Builder (SR Research) was used to implement the experiment. All sessions were also video recorded using two tripod-mounted cameras, but these images were used only to double check the validity of the measures of success manually coded by the experimenter during each session.

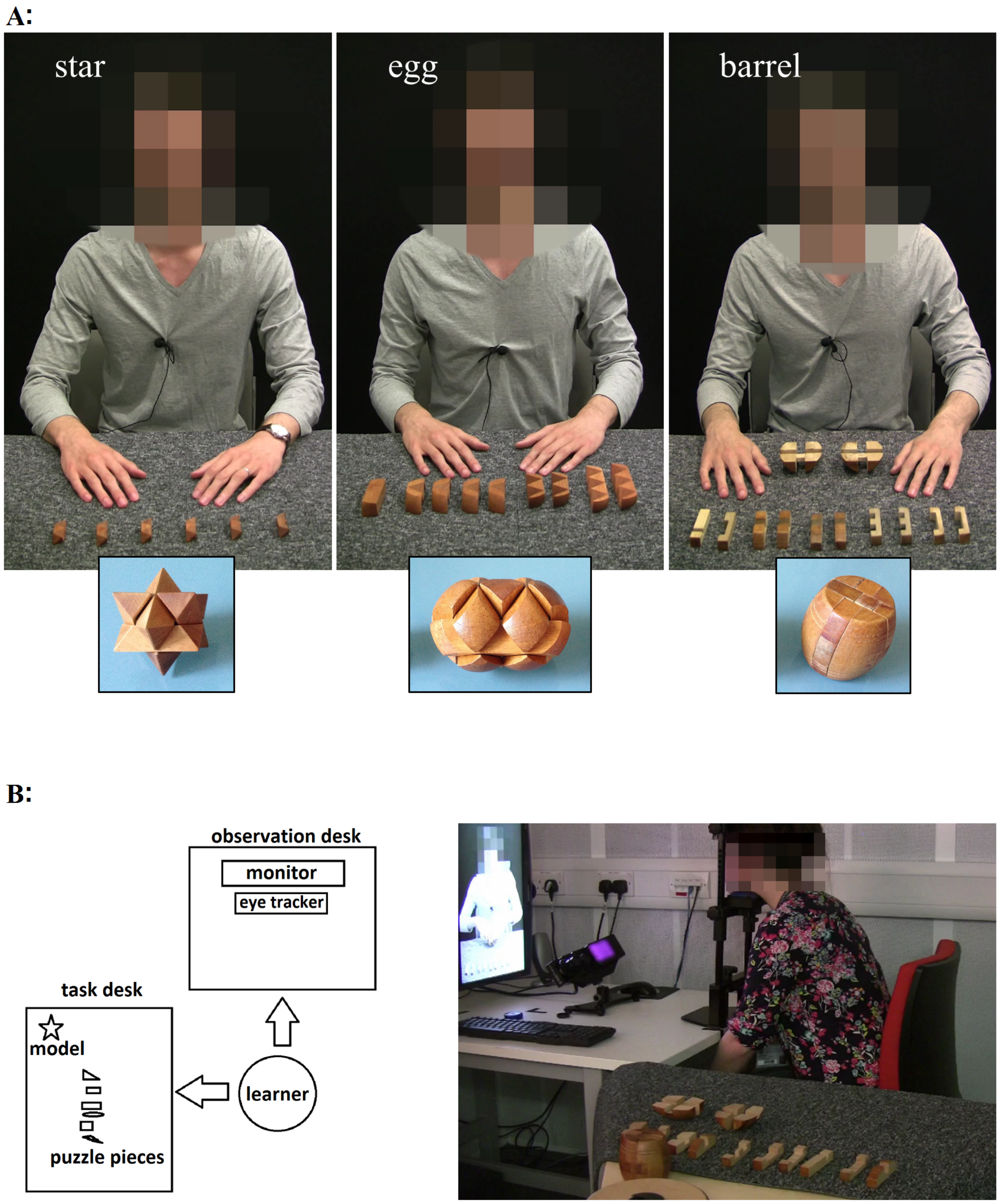

Figure 1 The experimental setup. **A:** Examples of the starting frames of the demonstration videos for the three puzzle tasks (star, egg, and barrel) in which the demonstrator has his face blurred. The insets show the corresponding solved puzzles. **B:** Plan diagram and photo of the workspace. The learner is at the eye-tracking desk watching the demonstration video and to her left is the task desk with the pieces of a barrel puzzle as well as an assembled model.

1.5. Procedure

The experimenter told the participants that they would alternate between watching the demonstration videos and attempting the task, and that this procedure would be repeated five times for each of the three puzzles, yielding a total of 15 trials per participant. At the start of each puzzle, the participant was shown all pieces of the puzzle and a correctly finished model and was asked whether she or he had seen it before. If the participant knew the puzzle, the experimenter would skip it and move on to the next (only one participant was familiar with one puzzle). Then, the experimenter asked the participant to produce a copy of the finished model to assess her or his initial ability to solve the puzzle (i.e., before watching the demonstration for the first time) and obtain a baseline score. Participants had a fixed time interval to solve the task (star: 90s, egg: 90s, barrel: 120s) corresponding to twice the time required by the demonstrator to solve it at a comfortable pace. During this period, participants could manipulate their own pieces and visually inspect the finished model but not touch it. The experimenter kept track of the time and interrupted the learner after the time-out, prompting her or him to turn to the eye-tracking desk. After the calibration and validation procedure, the participant watched the demonstration video corresponding to one, out of the four, experimental conditions while being eye-tracked. During this period, the experimenter disassembled the puzzle and re-arranged the pieces on the task desk to prepare for the participant's next attempt. After watching the video for the first time, the participant turned to the task desk and had another attempt at solving the puzzle, thus yielding the first performance measure after the baseline. The participant then turned back to the eye-tracking desk and, after a drift check, watched the demonstration video a second time before the next attempt. This sequence of steps (baseline test plus five iterations of watching the demonstration and attempting the task) was repeated for each of the three puzzles.

2. Analysis

2.1. Data processing

Demonstrator's manipulation data. We coded the demonstrator's manipulative actions from the demonstration videos into categorical time series at a sample rate of one observation every 25 milliseconds using the free software Solomon version beta 17.03.22 (Péter, 2016). Solving the puzzle requires joining pieces together, thus producing compounds (i.e., the partially-solved puzzle) along the way. In each 25ms temporal window, we used unique categorical labels to code the individual pieces, the compound being manipulated, or to indicate that the demonstrator was not holding any piece. When the demonstrator had a compound in one hand and a piece-to-be-added in the other hand, we used the label for the new piece and, after it was incorporated, the label for the newly-formed compound (see Figure 2A for an illustration of the resulting time series).

Learner's eye-movement data. Fixations and saccades events were extracted from the raw gaze data using the SR Research Data Viewer software, which performs saccade detection based on velocity and acceleration thresholds of $30°s^{-1}$ and $9{,}500°s^{-2}$, respectively. The eye-movement coordinates were mapped against dynamic Areas Of Interest (AOI), which were defined for each demonstration video using the same labels for pieces and compounds described in the previous paragraph and a label for 'other' to indicate when the participant was looking anywhere else on the screen. We used a customized algorithm written in the *R* programming language (R Core Team, 2016) to aggregate the eye-movement data into windows of 25ms and assign the label of the AOI that was fixated most of the time within such interval. We therefore obtained categorical time series indicating the sequence of objects fixated by the observers (scan-patterns) in each trial, with length and labels matching the categorical time series indicating the demonstrator's manipulative actions. To avoid very small differences in length that occurred during eye-tracking data collection among participants (star: SD = 6ms, range [1573ms, 1643ms]; egg: SD = 13ms, range [2000ms, 2159ms]; barrel: SD = 4ms, range [3078, 3114]), we normalized the length of the scan-patterns and manipulative actions in each puzzle to the same number of bins (star: 1,500 bins, egg: 2,000 bins, barrel: 3,000 bins).

Learner's performance data. At the end of each trial, the experimenter coded the learners' performance as either a success (i.e. the puzzle was assembled correctly before time-out) or a fail (i.e. the puzzle was not assembled before the time-out), and validated this data by watching the video recordings of the sessions.

Data exclusion. The initial dataset included 600 trials (40 participants x 3 puzzles x 5 iterations). From these, 5 trials were excluded due to one participant knowing the puzzle, 3 due to one participant inadvertently moving away from the eye tracker, 2 due to the participant accidentally moving the desk during data collection (perturbing the eye tracking system), and 124 due to the eye tracking data not being acquired properly. The final dataset comprised of 36 participants and 466 trials (condition noFACE_noAUDIO: 10 participants and 131 trials; FACE_noAUDIO: 8 participants and 109 trials; noFACE_AUDIO: 8 participants and 100 trials; and FACE_AUDIO: 10 participants and 126 trials).

2.2. Recurrence Quantification Analysis (*RQA*)

We examined the coordination dynamics between the scan-patterns of the learners (i.e. the sequence of pieces learners looked at while watching the demonstration videos) and the manipulative actions of the demonstrator (i.e. the sequence of pieces the demonstrator manipulated in the demonstration videos) using Recurrence Quantification Analysis or *RQA* (Marwan & Kurths, 2002; Marwan, Romano, Thiel, & Kurths, 2007; Shockley, Butwill, Zbilut, & Webber, 2002; Webber & Zbilut, 2005; Zbilut, Giuliani, & Webber, 1998). In particular, we produced cross-recurrence plots (*CRP*), from which we computed joint-recurrence plots (*JRP*) across the five trials of each puzzle to better capture the iterative process of the task. We used the *crqa* package (version 1.0.9) developed by Coco and Dale (2014) in the *R* software (R Core Team, 2016) to run our analyses using parameter values appropriate for categorical data: *delay* = 1, *embedding* = 1, and *radius* = 0.001.

In Figure 2B and Figure 2C, we illustrate how *CRP*s and *JRP*s were computed for a participant attempting the star puzzle across five iterations after the baseline test. For each trial, we had two time series: one for the manipulative actions of the demonstrator and the other for the scan-pattern of the learner watching the demonstration. Note that the time series for the demonstrator is the same across all five trials (because the demonstration video is the same) but the time series of the learner is different in each trial (because learners are free to move their eyes differently each time).

We produced a *CRP* for each trial by pairing the demonstrator (horizontal axis) with the learner (vertical axis). Conceptually, when the labels of the two time series match in some combination of time-points $[x_i, y_i]$ (i.e., if the puzzle piece being manipulated by the demonstrator at time x_i is the one being looked at by the learner at time y_i), this returns a cross-recurrence point for that entry. When the labels do not match, there is no cross recurrence (see Dale, Warlaumont, & Richardson, 2011, for an extensive explanation of *RQA* applied to categorical time series).

We then obtained joint-recurrence plots (*JRP*s) by simply multiplying the *CRP* of each iteration with all previous iterations on the same puzzle (see Figure 2C). Conceptually, only if all *CRP*s multiplied have a value of 1 in some entry $[x_i, y_i]$ (thus indicating cross-recurrence at that delay in all *CRP*s), then the resulting *JRP* will also have a value of 1 in that same entry, otherwise, the value will be zero. For the first iteration, we just kept the corresponding *CRP*, as there is no previous iteration to multiply it with. For iteration 2, we multiplied the two *CRP*s obtained for iterations 1 and 2. For iteration 3, we multiplied the three *CRP*s obtained for iterations 1, 2, and 3; and similarly for iterations 4 and 5. Therefore, the resulting *JRP*s reflect the dynamics of coordination between demonstrator's action and observer's gaze that is consistently found across the trials with each puzzle.

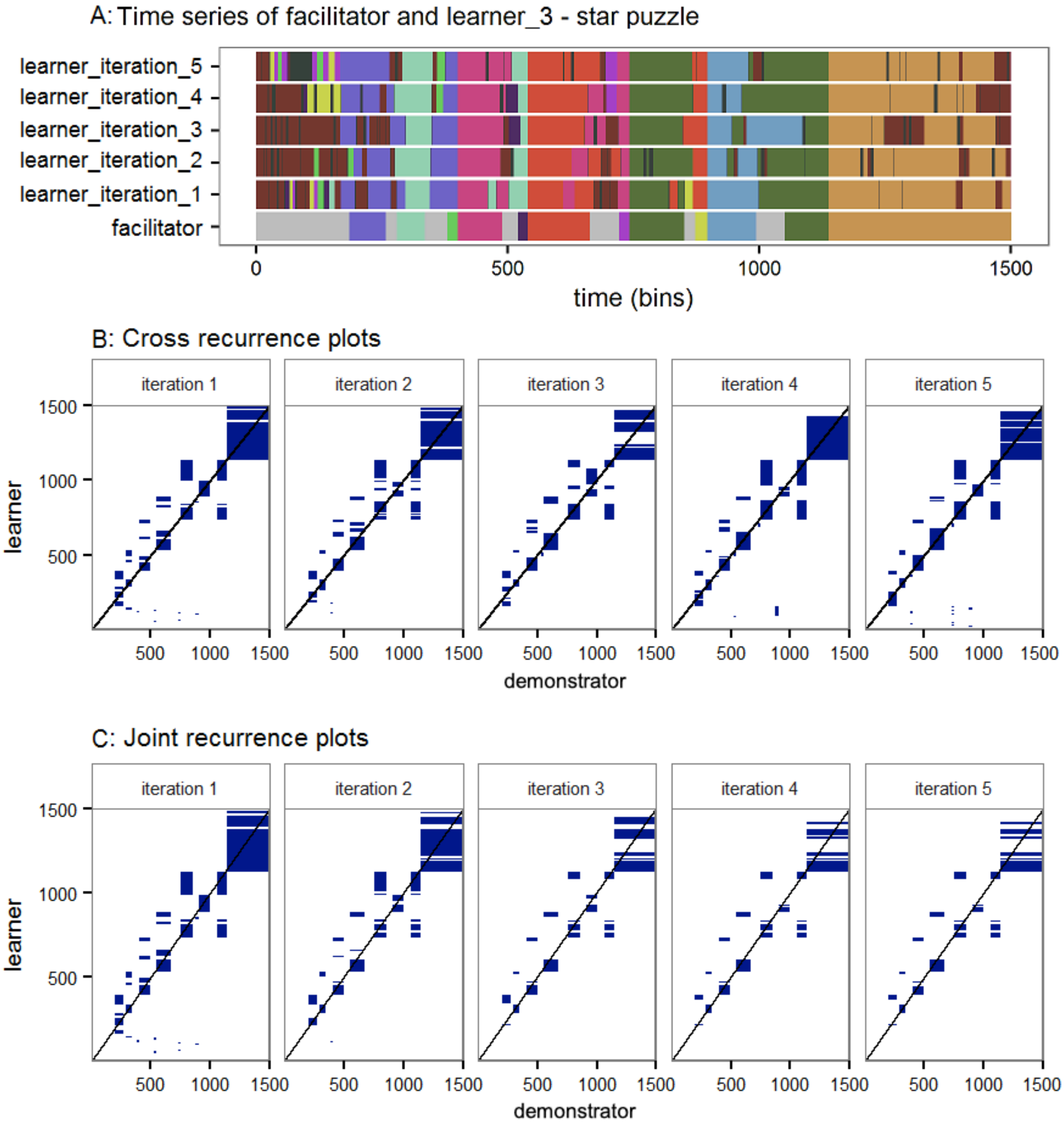

Figure 2. A: A single time series of the demonstrator manipulating the pieces of the star puzzle and five time series of one of the learners watching the corresponding video across the five iterations. The colours indicate either a single piece or the partially assembled puzzle being manipulated/looked at. The grey colour in the demonstrator's time series represents the moments in which he was not manipulating any piece. **B:** Cross recurrence plots (*CRP*) of the demonstrator's manipulative actions (horizontal axis) and the learner watching them (vertical axis). The line of synchrony, i.e., lag 0, is shown in black, and cross recurrence points are shown in blue. **C:** Joint recurrence plots (*JRP*) produced from the *CRP*s shown in B. For each iteration, the *JRP* is

produced by multiplying the *CRP* of that iteration with all previous ones, which leaves in only the recurrence points that consistently occur across iterations.

From each *JRP*, we computed three recurrence measures reported in the main analysis. The *recurrence rate* (*RR*), which is the proportion of cross-recurrence points in the *JRP,* corresponds mathematically to the cross-correlation sum (Kantz, 1994) and reflects the degree of shared activity or coordination between the two time series. The *determinism* (*DET*), which is the proportion of cross-recurrence points that form continuous diagonal lines (longer than a predefined threshold defined with the parameter *mindiagline* in the crqa package) and reflects the degree of synchronization between the two time series. The *mean line length* (*L*), which is the average length of the diagonal lines (longer than the threshold), reflects the average time in which the two time series remain synchronized.

To compute *DET* and *L* it is necessary to define the threshold parameter (*mindiagline* in the crqa package) because it indicates the minimum length of the diagonal lines in the recurrence plots, i.e. it defines the number of consecutive time-points needed to consider whether the two time series (e.g., the demonstrator and the observer) are in the same state (e.g., manipulating/attending to the same target). In our study, we obtained this threshold empirically by: (1) examining a range of possible threshold values, (2) plotting the resulting *DET* values as a function of the different threshold values examined, (3) visually inspecting these plots and (4) choosing the parameter value that counters ceiling effects (i.e., that leads *DET* values to vary rather than be concentrated at 100%). We obtained a minimum diagonal length threshold value of 30 data-points, which corresponds to a period of 750ms in the raw time series data. In other words, only synchronized attention and manipulative action that was longer than 750ms counted towards the values of *DET* and *L*.

Additionally, we computed measures of recurrence across the vertical line structures of the JRPs: the laminarity (*LAM*) and the trapping time (*TT*) and obtained largely corroborating results of those observed on the diagonal lines (i.e., *RR*, *DET* and *L*) reported in the main text. These additional analyses are explained and reported in section 6 of the electronic supplementary material.

2.3. Statistical analysis

RQA measures are descriptive in nature and, therefore, comparisons among cases (e.g., conditions, participants, or appropriate baselines) are required to draw inferences and examine specific predictions (Marwan et al., 2007; Shockley et al., 2002). Thus, we examined the relation between the learners' performance, the *RQA* measures of attentional coordination, and the design variables using Bayesian hierarchical logistic regression modelling and the framework of model comparison (Gelman et al., 2014; McElreath, 2016). This allowed us to adequately capture the complexity of our mixed design with repeated measures while improving the estimation of the effects with relatively small samples (e.g., Baldwin & Fellingham, 2013; Depaoli & van Schoot, 2015). Bayesian regression models were fit in the probabilistic programming language STAN (B. Carpenter et al., 2017) using the *map2stan* function, and compared using the *compare* function, both from the *rethinking* package (McElreath, 2016) in the *R* software. We used Markov Chain Monte Carlo (MCMC) simulation to obtain samples from the posterior distribution of the unknown parameters for which summary statistics were then computed (e.g., mean, credible intervals, differences, or the proportion of positive values). For all models, we used weakly informative priors (i.e., they were not completely flat but had little influence on the estimated posterior distributions) to obtain a wide range of sensible parameter values and yet avoid unreasonable values (Gelman et al., 2014; McElreath, 2016). We used normal priors with mean 0 and standard deviation of 10 for all non-constrained parameters, and we used half-Cauchy priors with location 0 and shape 5 for the variance parameters.

Our core question is whether attentional coordination, operationalized through the independent variables *RR*, *DET*, and *L,* is predictive of learners' performance across trials. We first fitted to the performance data our base model, a hierarchical logistic model (logit link) predicting the probability of task success (Eq. 1). The predictors are the parameters modelling the experimental conditions, i.e. *face* (indicating whether learners could see the demonstrator's face or if it was blurred) and *audio* (indicating whether learners could listen to the demonstrator's verbal instruction or not), *iteration*, and the interaction between condition and *iteration*. Both *face* and *audio* were dummy coded and modelled as between-participant fixed effects, whereas *iteration* was coded numerically from 0 to 4 (i.e., the five trials with each puzzle after the baseline test) and modelled as a within-participant fixed effect. The model also included indicators of the *task* (three levels: star, barrel and egg) and *participant* (36 levels) as varying intercepts (also called fully-crossed random effects). None of the participants solved any of the tasks during the baseline test, therefore we did not include the baseline score as a covariate. This base model captures how performance varies across iterations (i.e. the steepness of the learning curves) for the different experimental conditions and does not include any coordination variable. More formally, the base model can be represented as:

$$logit(p) = b_0 + b_1 * face + b_2 * audio + b_3 * face * audio + (b_4 + b_5 * face + b_6 * audio + b_7 * face * audio) * iteration + 1|task + 1|participant \quad (1)$$

We then fitted three additional models, each including one of the coordination variables, which were z-scored (i.e. subtracted from the mean and divided by the standard deviation), as a main (i.e. additive) effect. These models can be represented as:

$$logit(p) = base\_model + b_8 * RR \quad (2A)$$

$$logit(p) = base\_model + b_8 * DET \quad (2B)$$

$$logit(p) = base\_model + b_8 * L \quad (2C)$$

Lastly, we fitted three additional models including the interaction between the experimental condition and the respective coordination variable, thus allowing the effect of coordination (if there was any) to vary across conditions. These models can be represented as:

$$logit(p) = base\_model + (b_8 + b_9 * face + b_{10} * audio + b_{11} * face * audio) * RR \quad (3A)$$

$$logit(p) = base\_model + (b_8 + b_9 * face + b_{10} * audio + b_{11} * face * audio) * DET \quad (3B)$$

$$logit(p) = base\_model + (b_8 + b_9 * face + b_{10} * audio + b_{11} * face * audio) * L \quad (3C)$$

For each coordination variable, we compared the base model and the two additional models using the Widely Applicable Information Criterion or WAIC (Gelman et al., 2014; McElreath, 2016) to examine whether adding the coordination variable, either only as a main effect or also in interaction with condition, improves model prediction accuracy (the results of the model comparison are reported in section 4 in the electronic supplementary material). Lower values of WAIC indicate better predictive accuracy than higher values. We also examined the Akaike weights, which are rescaled values of WAIC where a total weight of 1 is partitioned among the models under consideration, thus indicating relative predictive accuracy among them (McElreath, 2016). Including *RR* as a main effect improved model accuracy but its interaction with the experimental conditions did not improve it further. Thus, we report model 2A. With respect to *DET* and *L*, including them both as main effect and in interaction with the experimental conditions improved the prediction accuracy over the base model. Thus, we report models 3B and 3C.

We ran 2000 iterations (including 1000 warmup iterations) on three chains for each model to ensure the robustness of the results, and report estimates of the posterior distributions from a total of 3,000 samples after warmup. All STAN models converged and mixing of the independent MCMC chains was good, as indicated by inspecting the trace plots and the number of effective sample sizes, and checking the *Rhat* values of the parameters were no higher than 1.01. More details can be found in the Open Science Framework page of this project where we provide a tutorial with the data and scripts to fit and compare the models, as well as to interpret the final models by computing the effects reported in Table 1 and replicating Figures 3 and 4. Unless otherwise indicated, we report the mean and 95% central credible interval of the estimated parameters from the fitted models. A strong evidence for an effect is when the 95% credible interval excludes 0, and weak evidence when the 95% credible interval includes 0 but the 90% does not.

In section 2 of the electronic supplementary material we report two more models, one examining the performance of the learners across the experimental and the additional control conditions to provide further evidence that learning is indeed facilitated by the demonstrator (in other words, that this is a case of 'social' learning), and the other examining the proportion of fixation of the learners to the demonstrator face vs. pieces to obtain clearer insights on the effect of intentional gaze.

3. Results

Table 1 shows the parameter estimates and odds ratios of the three fit logistic models chosen for interpretation (*RR*: model 2A; *DET*: model 3B; *L*: model 3C). Take, for example, the model including *RR* (i.e. the first four rows in Table 1). We observe an odds ratio of 3.17 for the effect of iteration in the noFACE_noAUDIO condition, which means the odds of solving the puzzle increases 217% from one iteration to the next. Similarly, we observe an odds ratio of 2.48 for the effect of *RR* across all conditions (as there is no interaction between experimental conditions and *RR* in the model), which means the odds of solving the puzzle increases 148% for each unit increase in *RR*.

To help interpretation, we simulated data from the fitted models. To do this, we must decide how to deal with the random effects. We could simulate them too and doing this would increase the variation obtained for the simulated outcome. However, this is unhelpful here as we are not so much interested in the differences among tasks or among participants, but rather in the systematic differences among the experimental conditions. To focus on this aspect, we declared the random effects as zero in the simulations, which corresponds to simulating for an 'average' task and 'average' participant. Figure 3 shows simulations from the three final models to illustrates the effect of *RR*, *DET*, and *L* on the probability of success across conditions, averaging over the effect of iteration. Figure 4, instead, focuses on model 2A (with *RR*) to illustrate also the effect of iteration, and the corresponding figures for *DET* and *L* can be found in the electronic supplementary material, section 5.

We will interpret the results of each model in turn and start with model 2A (i.e., *RR*). In line with our main prediction, we found strong evidence that the coordination variable *RR* was positively associated with the probability of success across all experimental conditions (see effect of coordination on Table 1, Figure 3 top row, and Figure 4), which indicates that attentional coordination is beneficial for observational learning. Furthermore, the effect of iteration was positive in all conditions, i.e., learners get progressively better at solving the puzzle.

In order to test whether the effect of iteration (i.e. learning rates) differs across conditions, we examined the posterior distribution from the fitted model. For each sample of the posterior distribution, we computed the difference between the effect of iteration estimated for different conditions (say, FACE_AUDIO and FACE_noAUDIO). This process generates a vector of estimated differences, which we summarised by computing the mean and 95% credible intervals. This summary statistics can be used as evidence (or lack thereof) for a systematic difference between conditions (Gelman et al., 2014). A credible interval crossing zero suggests that the difference between the estimates is not systematic (or, in a frequentist terminology, 'not significant'). If the credible interval instead does not cross the zero, this suggests that the difference is indeed systematic or 'significant'. Moreover, a positive difference means the first term of the difference has a higher estimate, and a negative difference means the second term has a higher estimate.

We found that the effect of iteration was larger in the condition FACE_AUDIO than FACE_noAUDIO (difference between the estimates: 1.14 [0.4, 1.97]) and noFACE_AUDIO than noFACE_noAUDIO (difference between the estimates: 0.88 [0.15, 1.56]). This indicates that learners who could listen to the demonstrator learned faster than those that could not. We found no difference between the effect of iteration for conditions FACE_AUDIO and noFACE_AUDIO: -0.04 [-0.8, 0.76]; and for conditions FACE_noAUDIO and noFACE_noAUDIO: -0.3 [-0.97, 0.39]). This result instead indicates that the performance of learners did not benefit from seeing the demonstrator's face.

Table 1 Estimated mean values and a 95% CI (unless a 90% CI is otherwise indicated) for the relative effects of iteration and coordination on the probability of task success across conditions, computed for the three final models (one for each coordination variable, *RR*, *DET*, and *L*). Values indicating strong or weak evidence of an effect are in bold to aid reading.

Coordination variable in the model	Condition	Effect of iteration		Effect of coordination	
		Estimate	Odds ratio	Estimate	Odds ratio
RR		**1.11 [0.55, 1.63]**	**3.04 [1.74, 5.11]**	**0.91 [0.02, 1.78]**	**2.48 [1.02, 5.93]**
		**0.81 [0.26, 1.37]**	**2.25 [1.30, 3.95]**	**0.91 [0.02, 1.78]**	**2.48 [1.02, 5.93]**
		**2.00 [1.34, 2.67]**	**7.36 [3.83, 14.41]**	**0.91 [0.02, 1.78]**	**2.48 [1.02, 5.93]**
		**1.95 [1.26, 2.65]**	**7.04 [3.52, 14.21]**	**0.91 [0.02, 1.78]**	**2.48 [1.02, 5.93]**
DET		**1.39 [0.70, 2.12]**	**4.03 [2,01, 8.30]**	**1.14 [0.01, 2.29]**	**3.13 [1.01, 9.91]**
		0.15 [-0.66, 0.91]	1.17 [0.52, 2.47]	-1.32 [-3.03, 0.45]	0.27 [0.05, 1.57]
		**2.70 [1.76, 3.68]**	**14.89 [5.80, 39.70]**	**2.14 [0.81, 3.68]**	**8.50 [2.25, 39.49]**
		**1.44 [0.78, 2.15]**	**4.23 [2.18, 8.57]**	**-1.11 [-2.12, -0.17]**	**0.33 [0.12, 0.85]**
L		**1.73 [0.98, 2.55]**	**5.66 [2.67, 12.86]**	**2.05 [0.82, 3.28]**	**7.76 [2.28, 26.45]**
		0.01 [-0.77, 0.77]	1.01 [0.46, 2.16]	**-1.82 90% CI [-3.42, -0.24]**	**0.16 90% CI [0.03, 0.79]**
		**2.20 [1.37, 3.07]**	**9.07 [3.93, 21.63]**	**1.39 90% CI [0.02, 2.71]**	**4.00 90% CI [1.02, 14.96]**
		**1.58 [0.97, 2.28]**	**4.87 [2.65, 9.79]**	-0.41 [-1.11, 0.30]	0.66 [0.33, 1.35]

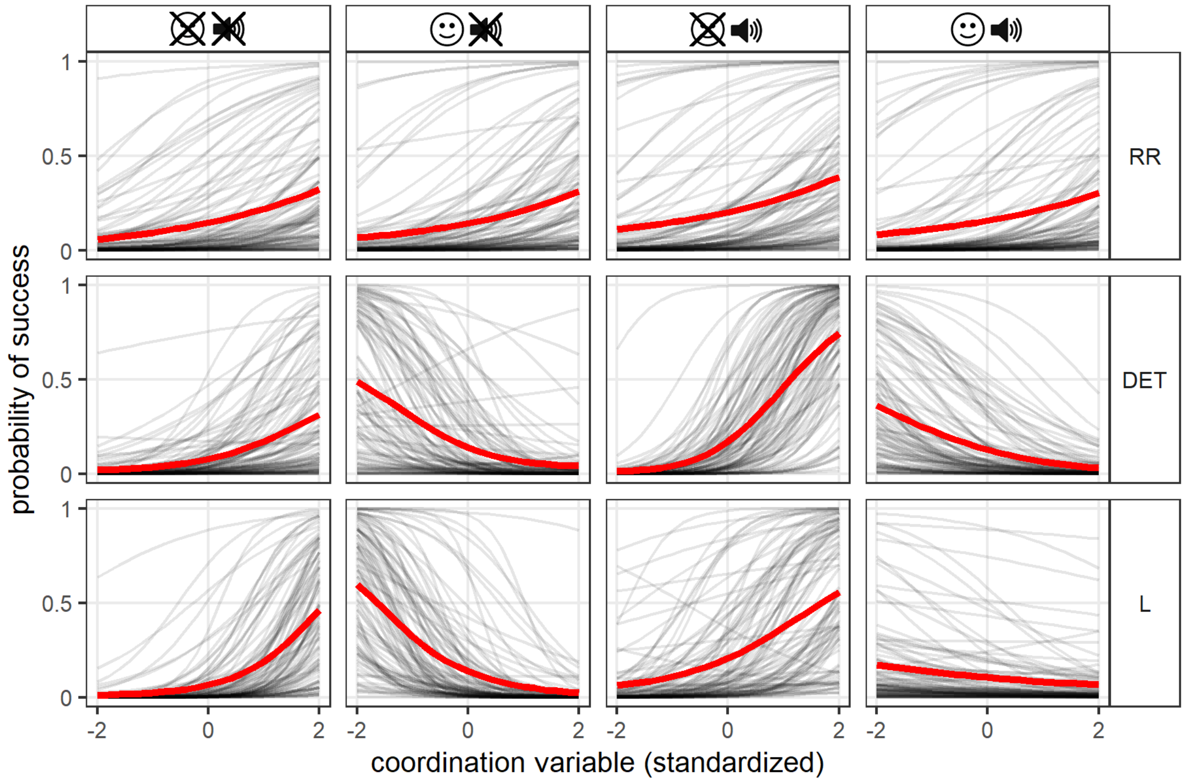

Figure 3. Posterior predictions of the three final logistic models showing the probability of success (vertical axis) as a function of coordination (horizontal axis) as captured by the *RQA* variables (*RR*, top row; *DET*, middle row; *L*, bottom row) across the four experimental conditions organized along the columns. Coordination variables are standardized (z-scored) with -2 corresponding to 2 SD below the average (low coordination); 0 corresponding to the average value; and 2 corresponding to 2 SD above the average (high coordination). These simulations are for an average task and average participant. The shaded black lines represent 100 simulations and the thick red lines represent the mean of all simulations within each plot.

The estimated parameters just discussed reflect the relative effects of iteration and coordination on the probability of successfully assembling the puzzle. In order to visualize and interpret their joint contribution, we simulated outcome values (probability of success) from the fitted model. We fixed the parameter for *RR* at either the average value, a low value (2 *sd* below the average), or a high value (2 *sd* above the average) and generated 100 predictions for the probability of success for an average task and average participant. The simulated outcome, reported in Figure 4, clearly shows how the performance of hypothetical learners (vertical axes) increases as a function of iterations (horizontal axes), varies for the different experimental conditions (across columns) and is modulated by the degree of attentional coordination (across rows). A comparison between the three plots within each column in Figure 4 shows that the learning curves are shifted upwards from low to high values of attentional coordination. This illustrates that learning is faster among learners who could coordinate their overt attention with the demonstrator's manipulations more consistently across trials (i.e. those with higher values of coordination computed from the *JRPs*). In addition, the learning curves are steeper in column 3 compared with those in column 1, and in column 4 compared to column 2, which confirms that learning was faster for those individuals who could listen to the verbal instructions as compared to those that could not. Finally, the learning curves in column 2 are not systematically different from those in column 1, and those in column 4 are also not different from those in column 3, which confirms that seeing the demonstrator's face did not seem to facilitate learning.

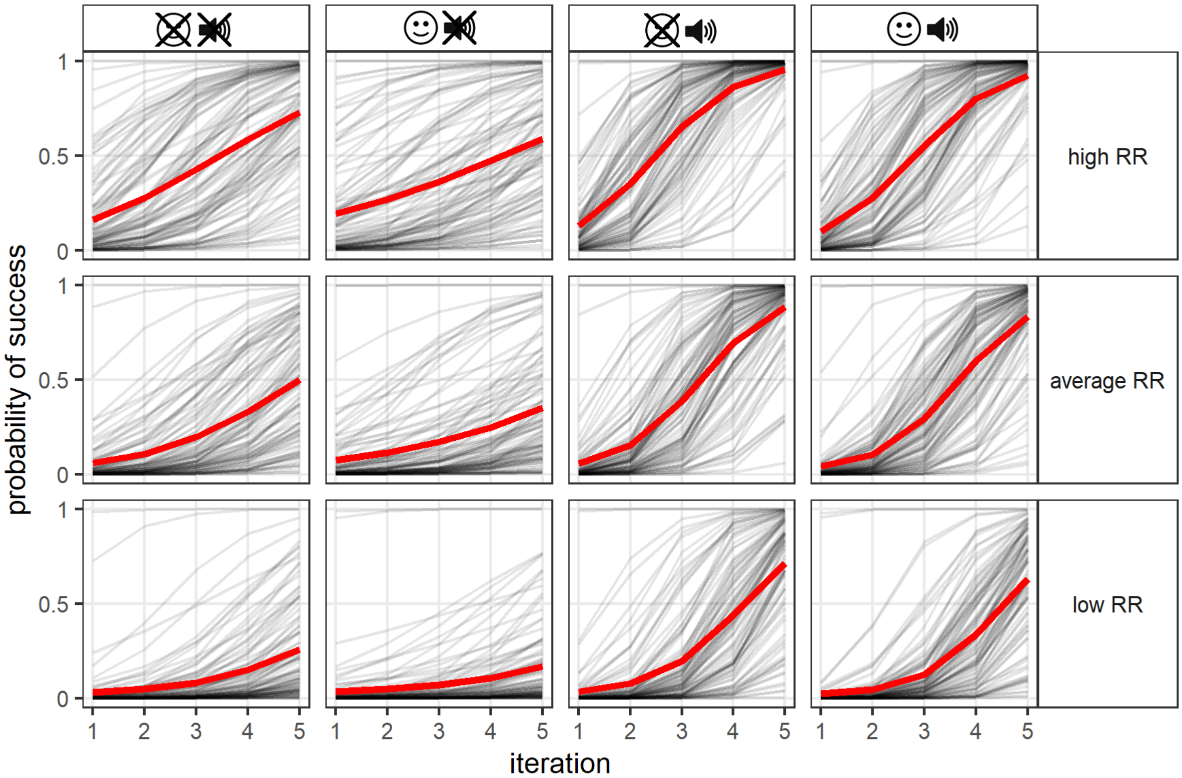

Figure 4. Posterior predictions of the final logistic model with the coordination variable *RR* (model 2A) showing the probability of success (vertical axis) as a function of iterations (horizontal axis) across conditions (columns), while holding *RR* at either 2 *sd* below the average (low *RR*, bottom row), at the average value (average *RR*, middle row), or at 2 *sd* above the average (high *RR*, top row). These simulations are for an average task and average participant. The shaded black lines represent 100 simulations and the thick red lines represent the mean of all simulations within each plot. To see the effect of the different values of *RR* on performance, the reader should compare the three plots within each column. To see the effect of seeing the demonstrator's face compared to face blurred, the reader should compare the plots in column 1 with those in column 2, and the plots in column 3 with 4. To see the effect of listening to the demonstrator's speech compared to no audio, the reader should compare the plots in column 1 with those in column 3, and the plots in column 2 with 4.

Model 3B (i.e., with coordination variable *DET*) and model 3C (with *L*) show similar patterns, albeit with some interesting differences (Table 1, Figure 3 middle and bottom rows, see Figures S5 and S6 in the electronic supplementary material for the visualization of posterior predictions). When the demonstrator's face was blurred, both *DET* and *L* were positively associated with probability of success, which confirms that learners who synchronized their eye-movement for longer with the demonstrator's actions learned faster than those synchronising for shorter period of time.

However, when the demonstrator's face was visible, the probability of success was actually reduced for increasing values of *DET* and *L*. This is illustrated in Figure 3 (middle and bottom rows), which shows that the probability of success declines for higher values of *DET* and *L* in the conditions FACE_noAUDIO and FACE_AUDIO. Accordingly, Figures S5 and S6 in the electronic supplementary material show that the learning curves shift downward as we move from low to high values of *DET* and *L*. This suggests that seeing the demonstrator's face, compared to face blurred, was detrimental to learning. This result is confirmed by the strong evidence that iteration has a smaller effect on the probability of success when comparing FACE_noAUDIO with noFACE_noAUDIO for both *DET* and *L* (difference between the estimates for *DET*: -1.24 [-2.31, -0.22]; for *L*: -1.72 [-2.75, -0.63]); and comparing FACE_AUDIO with noFACE_AUDIO for *DET* but not for *L* (difference between the estimates for *DET*: -1.26 [-2.37, -0.14]; for *L*: -0.62 [-1.72, 0.39]).

We speculate that the presence of the demonstrator's face attracted the attention of learners to it, distracting them from the actual manipulation task without providing any benefit. Additional analyses reported in the electronic supplementary material (section 3) corroborate this suggestion by confirming that learners looked more at the demonstrator's face when it was visible compared to blurred (difference in the mean estimates of the proportion of fixation time between FACE_noAUDIO and noFACE_noAUDIO: 3.14% [0.5%, 10.3%]), between FACE_AUDIO and noFACE_AUDIO: 5.6% [0.8%, 17.9%]), and even more so when they could listen to his speech (difference between FACE_AUDIO and FACE_noAUDIO: 2.9%, 90% CI [0.2%, 8.0%]).

4. Discussion

Observational learning (or production imitation) is a time-evolving process involving a demonstrator (or model), a learner (or observer), and a target task. In this study, we borrowed the conceptual and analytical framework of dynamical system theory as applied and developed in the cognitive sciences (e.g., Coco et al., 2017; Dale, et, al. 2013; Fusaroli, et al., 2014) to investigate the role of attentional coordination in the 'passing on' or re-construction of knowledge. Researchers in diverse fields have claimed that learning through observation benefits from a constant interaction and tight attentional coupling between the learner and the resources made available by the demonstrator (e.g., M. Carpenter et al., 1998; Mundy & Newell 2009; Tomasello, 2009). However, the experimental support for this claim has lacked both temporal and spatial resolution – for example, because studies used manual annotations of gaze directions from video footage (e.g., M. Carpenter et al 1998), or used eye-tracking measures that aggregate data over time, such as number of fixations, which provides little insight about how attention unfolds over time (e.g., Breslin et al., 2009).

In the current study, we combined eye-tracking with sophisticated computational analyses (*RQA* and Bayesian hierarchical regression) and provided evidence that learners better able to coordinate their overt attention with the manipulative actions of the demonstrator had an increasingly higher probability of success in solving a construction puzzle task. Through this dynamical interaction with the demonstrator's unfolding actions, learners discovered object affordances and the sequence of actions required to successfully complete the task more quickly than if they were learning alone.

In this study, we also investigated how the availability of verbal instruction and intentional gaze interacts with attentional coordination and mediate the learning outcomes. Speech and overt attention are known to synchronise strongly during language comprehension, language production, and even dialogue tasks (e.g., Coco & Keller, 2012; Knoeferle & Crocker, 2006; Richardson et al., 2007). We therefore expected that the availability of verbal instruction would improve task performance and be associated with better coordination between overt attention and manipulative actions. Indeed, we found evidence that speech helps cognitive processes to align and plays an important role in the passing on of knowledge, as shown by the stronger improvement of performance compared to when speech was not available.

The availability of intentional gaze is considered important to build joint attention (e.g., Tomasello et al., 2005) and we therefore expected that being able to see the demonstrator's face (as opposed to his blurred face) would improve the learning outcome of our participants in the manipulative task. However, we found that the availability of the demonstrator's face, and hence of his intentional gaze, were instead detrimental to learning. Learners tended to look more often at the demonstrator's face when it was visible (compared to blurred) and even more often when they could also hear him speaking. These bouts of attention away from the manipulative actions of the demonstrator and towards his face have likely distracted learners and hence negatively impacted on their learning. We note, however, that our study utilises pre-recorded videos and that, in cases of live interaction, the behaviour of looking at the partner's eyes is likely to play important roles, such as to indicate engagement or request the partner's attention, and hence may be beneficial to learning. Regardless, it is interesting to observe that learners coordinated their visual attention with the demonstrator's actions even when his face was blurred. This result is consistent with the "hand-eye coordination" route to joint attention (Yu & Smith, 2013) rather than the more widely acknowledged gaze-following route and suggests that this alternative route may play an important role in the processes of social learning which has received little attention.

Using pre-recorded demonstrations enabled us to achieve greater control when measuring the attentional coordination across learners, because they all watched the same videos. While demonstration videos are commonly used in studies of observational learning, this is arguably one of the main limitations of this design. Most cases of observational learning occur during face-to-face encounters, thus it would be important to examine demonstrator-learner dyads interacting live using the same paradigm. Another important limitation of this study is the relatively small number of participants. The novel manipulative task we conceived was particularly time-consuming, as it not only involved eye-tracking (while participants watched the demonstrations) but also required manual performance (to measure success in every trial) and was iterative (to measure changes in performance across trials, i.e. learning), requiring a total of 15 trials for each participant. To overcome the resulting time constraint, we manipulated the experimental conditions (i.e. type of demonstration video) between participants, which limited the sample size in each. Even though Bayesian statistics is more robust in the context of small sample sizes (see Gelman et al. 2014; van de Schoot et al., 2014) and despite finding systematic differences across conditions, the results must be interpreted as exploratory and might be used as an important foundation for future research interested in similar research questions and deploying a similar methodology. The results from the current study can constitute a solid basis for power analyses estimating effect size statistic in designs aimed at replicating our findings or extending in other ways our innovative experimental approach.

This study did not seek to address how the ability to identify and track the relevant aspects of the demonstration develops. Further work might use a similar paradigm to examine dyads from different age groups, and we expect that measures of attentional coordination will be positively correlated with age. In principle, similar methods could be applied to the study of social learning in nonhuman animals, allowing researchers to explore whether coordination is central to social learning more generally, or a species-specific feature of human social learning.

One methodological contribution of our study is to show that the combination of eye-tracking methods, *RQA*, and hierarchical modelling, can provide a powerful tool for examining the mechanisms of observational learning with finer granularity. Future research could exploit these methods to further elucidate how and the extent to which the dynamics of attentional coordination may influence social learning by looking, for example, at the stability of the attentional coordination, and the relation between patterns of attentional coordination and learning trajectories, during iterative observational learning. Novel extensions of recurrence quantification analysis to multi-dimensional data might be successfully used to investigate patterns of learning involving larger groups of individuals interacting in real time (see Knight, Kennedy, & McComb, 2016; Wallot, Roepstorff, & Mønster, 2016 for recent developments in this direction).

We conclude that viewing social learning from the perspective of moment-to-moment attentional coordination might provide novel theoretical insights to the field, and we hope the present study will motivate further work that embraces the technological and analytical advances deployed here.

Funding: This work was supported by the University of St Andrews; the Konrad Lorenz Institute for Evolution and Cognition Research [Writing-up Fellowship awarded to MP]; the John Templeton Foundation [grant number 40128 awarded to KNL]; the Leverhulme Trust [grant number ECF-014-205 awarded to MIC]; and the Fundação para a Ciência e Tecnologia [grant number PTDC/PSI-ESP/30958/2017 awarded to MIC]. The funding sources were not involved in the study design; the collection, analysis and interpretation of data; the writing of the report; and the decision to submit the article for publication.

Declarations of interest: none

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419-439. https://doi.org/10.1006/jmla.1997.2558

Ashford, D., Bennett, S. J., & Davids, K. (2006). Observational Modeling Effects for Movement Dynamics and Movement Outcome Measures Across Differing Task Constraints: A Meta-Analysis. *Journal of Motor Behavior*, 38(3), 185-205. https://doi.org/10.3200/JMBR.38.3.185-205

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, 18(2), 151–164. https://doi.org/10.1037/a0030642

Bird, G., & Heyes, C. (2005). Effector-Dependent Learning by Observation of a Finger Movement Sequence. *Journal of Experimental Psychology: Human Perception and Performance*, 31(2), 262-275. https://doi.org/10.1037/0096-1523.31.2.262

Breslin, G., Hodges, N. J., & Williams, M. A. (2009). Effect of Information Load and Time on Observational Learning. *Research Quarterly for Exercise and Sport*, 80(3), 480-490. https://doi.org/10.1080/02701367.2009.10599586

Carcea, I., & Froemke, R. C. (2019). Biological mechanisms for observational learning. *Current Opinion in Neurobiology*, 54, 178-185. https://doi.org/10.1016/j.conb.2018.11.008

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A Probabilistic Programming Language. *2017, 76*(1), 32. https://doi.org/10.18637/jss.v076.i01

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs for the Society of Research in Child Development*, 63, 1-143. https://doi.org/10.2307/1166214

Carpenter, M., & Tomasello, M. (1995). Joint Attention and Imitative Learning in Children, Chimpanzees, and Enculturated Chimpanzees. *Social Development*, 4(3), 217-237. https://doi.org/10.1111/j.1467-9507.1995.tb00063.x

Casile, A., & Giese, M. A. (2006). Nonvisual Motor Training Influences Biological Motion Perception. *Current Biology*, 16(1), 69-74. https://doi.org/10.1016/j.cub.2005.10.071

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, Mass.: MIT Press.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*(1), 62-81. https://doi.org/10.1016/j.jml.2003.08.004

Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00510

Coco, M. I., Dale, R., & Keller, F. (2018). Performance in a Collaborative Search Task: The Role of Feedback and Alignment. *Topics in Cognitive Science, 10*(1), 55-79. https://doi.org/10.1111/tops.12300

Coco, M. I., & Keller, F. (2012). Scan Patterns Predict Sentence Production in the Cross-Modal Processing of Visual Scenes. *Cognitive Science, 36*(7), 1204-1223. https://doi.org/10.1111/j.1551-6709.2012.01246.x

Coco, M. I., & Keller, F. (2015). Integrating mechanisms of visual guidance in naturalistic language production. *Cognitive Processing, 16*(2), 131-150. https://doi.org/10.1007/s10339-014-0642-0

Coco, M. I., Keller, F., & Malcolm, G. L. (2016). Anticipation in Real-World Scenes: The Role of Visual Context and Visual Memory. *Cognitive Science, 40*(8), 1995-2024. https://doi.org/10.1111/cogs.12313

Coco, M. I., Badino, L., Cipresso, P., Chirico, A., Ferrari, E., Riva, G., Gaggioli, A. & D'Ausilio, A., 2017. Multilevel behavioral synchronization in a joint tower-building task. *IEEE Transactions on Cognitive and Developmental Systems*, 9(3), pp.223-233.

D'Innocenzo, G., Gonzalez, C. C., Williams, A. M., & Bishop, D. T. (2016). Looking to Learn: The Effects of Visual Guidance on Observational Learning of the Golf Swing. *PLoS ONE*, 11(5), e0155442. https://doi.org/10.1371/journal.pone.0155442

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The Self-Organization of Human Interaction. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 59, pp. 43-95): Academic Press.

Dale, R., Warlaumont, A. S., & Richardson, D. C. (2011). Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos, 21*(4), 1153-1161. https://doi.org/10.1142/s0218127411028970

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences, 6*(4), 485-507. https://doi.org/10.1007/s11097-007-9076-9

Degotardi, S. (2017). Joint attention in infant-toddler early childhood programs: Its dynamics and potential for collaborative learning. *Contemporary Issues in Early Childhood*, 18(4), 409-421. https://doi.org/10.1177/1463949117742786

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, 22(2), 240–261. https://doi.org/10.1037/met0000065

Fogel, A. (1993). *Developing through relationships : origins of communication, self, and culture*. Chicago: University of Chicago Press.

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing Social Interactions: The Promises and Challenges of Using Cross Recurrence Quantification Analysis. *Translational Recurrences: From Mathematical Theory to Real-World Applications, 103*, 137-155. https://doi.org/10.1007/978-3-319-09531-8_9

Galef, B. G. (1988). Imitation in animals: History, definition, and interpretation of data from the psychological laboratory. In Z. T. R. & G. B. G. (Eds.), *Social Learning: Psychological and Biological Perspectives* (pp. 3-28). Hillsdale, NJ: Erlbaum.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third edition. ed.). Boca Raton: CRC Press.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science, 14*(5), 462-466. https://doi.org/10.1111/1467-9280.02454

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274-279. https://doi.org/10.1111/1467-9280.00255

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*(5), 347-356. https://doi.org/10.1007/bf00336922

Heyes, C. (1994). Social-Learning in Animals - Categories and Mechanisms. *Biological Reviews of the Cambridge Philosophical Society, 69*(2), 207-231. https://doi.org/10.1111/j.1469-185X.1994.tb01506.x

Hodges, N. J., Williams, A. M., Hayes, S. J., & Breslin, G. (2007). What is modelled during observational learning? *Journal of Sports Sciences*, 25(5), 531-545. https://doi.org/10.1080/02640410600946860

Hoppitt, W. J. E., & Laland, K. N. (2013). *Social learning : an introduction to mechanisms, methods, and models*. Princeton: Princeton University Press.

Horn, R. R., Williams, A. M., Hayes, S. J., Hodges, N. J., & Scott, M. A. (2007). Demonstration as a rate enhancer to changes in coordination during early skill acquisition. *Journal of Sports Sciences*, 25(5), 599-614. https://doi.org/10.1080/02640410600947165

Horn, R. R., Williams, A. M., Scott, M. A., & Hodges, N. J. (2005). Visual search and coordination changes in response to video and point-light demonstrations without KR. *Journal of Motor Behavior*, 37(4), 265-274.

Ingold, T. (2001). From the transmission of representations to the education of attention. In H. Whitehouse (Ed.), *The Debated Mind: Evolutionary psychology versus ethnography* (pp. 113-153). Oxford: Berg.

Johansson, R. S., Westling, G. R., Backstrom, A., & Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. *Journal of Neuroscience, 21*(17), 6917-6932.

Kantz, H. (1994). Quantifying the Closeness of Fractal Measures. *Physical Review E, 49*(6), 5091-5097. https://doi.org/10.1103/PhysRevE.49.5091

Kelso, J. A. S. (1995). *Dynamic patterns : the self-organization of brain and behavior*. Cambridge, Mass.: MIT Press.

Kelso, J. A. S. (2016). On the Self-Organizing Origins of Agency. *Trends Cogn Sci, 20*(7), 490-499. https://doi.org/10.1016/j.tics.2016.04.004

Knight, A. P., Kennedy, D. M., & McComb, S. A. (2016). Using recurrence analysis to examine group dynamics. *Group Dynamics: Theory, Research, and Practice*, 20(3), 223–241. https://doi.org/10.1037/gdn0000046

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science, 30*(3), 481-529. https://doi.org/10.1207/s15516709cog0000_65

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*(25-26), 3559-3565. https://doi.org/10.1016/s0042-6989(01)00102-x

Marwan, N., & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A, 302*(5-6), 299-307.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports-Review Section of Physics Letters, 438*(5-6), 237-329. https://doi.org/10.1016/j.physrep.2006.11.001

Mattar, A. A. G., & Gribble, P. L. (2005). Motor Learning by Observing. *Neuron*, 46(1), 153-160. https://doi.org/10.1016/j.neuron.2005.02.009

McElreath, R. (2016). *Statistical rethinking : a Bayesian course with examples in R and Stan*. Boca Raton: CRC Press/Taylor & Francis Group.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition, 66*(2), B25-B33. https://doi.org/10.1016/s0010-0277(98)00009-2

Mundy, P. & Newell, L. (2009) Attention, joint attention, and social cognition. *Curr Dir Psychol Sci* 16(5): 269-74

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19(1), 1-32. https://doi.org/10.1016/0010-0285(87)90002-8

Péter, A. (2016). Solomon coder. Retrieved from http://solomoncoder.com/

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045-1060. https://doi.org/10.1207/s15516709cog0000_29

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination - Common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*(5), 407-413. https://doi.org/10.1111/j.1467-9280.2007.01914.x

Richardson, M. J., Dale, R., & Marsh, K. L. (2014). *Complex Dynamical Systems in Social and Personality Psychology Theory, Modeling, and Analysis*.

Schertz, H. H., Odom, S. L., Baggett, K. M., & Sideris, J. H. (2013). Effects of Joint Attention Mediated Learning for toddlers with autism spectrum disorders: An initial randomized controlled study. *Early Childhood Research Quarterly*, 28(2), 249-258. https://doi.org/10.1016/j.ecresq.2012.06.006

Schoner, G., & Kelso, J. A. S. (1988). Dynamic pattern generation in behavioral and neural systems. *Science, 239*(4847), 1513-1520. https://doi.org/10.1126/science.3281253

Schoner, G., Zanone, P. G., & Kelso, J. A. S. (1992). Learning as change in coordination dynamics - theory and experiment. *Journal of Motor Behavior, 24*(1), 29-48. https://doi.org/10.1080/00222895.1992.9941599

Shockley, K., Butwill, M., Zbilut, J. P., & Webber, C. L. (2002). Cross recurrence quantification of coupled oscillators. *Physics Letters A, 305*(1-2), 59-69.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, Mass.: Harvard University Press.

Tomasello, M. (2003). *Constructing a language : a usage-based theory of language acquisition*. Cambridge, Mass.: Harvard University Press.

Tomasello, M. (2009). *Why we cooperate*. Cambridge, Mass.: MIT Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675-691. https://doi.org/10.1017/S0140525x05000129

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, 16(3), 495-511. https://doi.org/10.1017/S0140525X0003123X

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. Child Development, 85, 842–860. https://doi.org/10.1111/cdev.12169

van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior, 25*(3), 785-791. https://doi.org/10.1016/j.chb.2009.02.007

Vogt, S. (1995). On relations between perceiving, imagining and performing in the learning of cyclical movement sequences. *British Journal of Psychology*, 86(2), 191-216. https://doi.org/10.1111/j.2044-8295.1995.tb02556.x

Wallot, S., Mitkidis, P., McGraw, J. J., & Roepstorff, A. (2016). Beyond Synchrony: Joint Action in a Complex Production Task Reveals Beneficial Effects of Decreased Interpersonal Synchrony. *PLoS ONE, 11*(12), e0168306. https://doi.org/10.1371/journal.pone.0168306

Wallot, S., Roepstorff, A., & Mønster, D. (2016). Multidimensional Recurrence Quantification Analysis (MdRQA) for the Analysis of Multidimensional Time series: A Software Implementation in MATLAB and Its Application to Group-Level Data in Joint Action. *Frontiers in Psychology, 7*(1835). https://doi.org/10.3389/fpsyg.2016.01835

Webber, C. L., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. In M. A. Riley & G. C. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 26-94).

Whiten, A., & Ham, R. (1992). On the nature and evolution of imitation in the animal kingdom: Reappraisal of a century of research. *Advances in the Study of Behaviour*(21), 239-283.

Whiten, A., Horner, V., Litchfield, C. A., & Marshall-Pescini, S. (2004). How do apes ape? *Animal Learning & Behavior, 32*(1), 36-52. https://doi.org/10.3758/bf03196005

Williams, A. M., & Hodges, N. J. (2005). Practice, instruction and skill acquisition in soccer: Challenging tradition. *Journal of Sports Sciences*, 23(6), 637-650. https://doi.org/10.1080/02640410400021328

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE, 8*(11), e79659. https://doi.org/10.1371/journal.pone.0079659

Zbilut, J. P., Giuliani, A., & Webber, C. L. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A, 246*(1-2), 122-128. https://doi.org/10.1016/S0375-9601(98)00457-5