Master Écophysiologie, Éthologie et Écologie

MASTER INTERNSHIP

January 2020 – August 2020
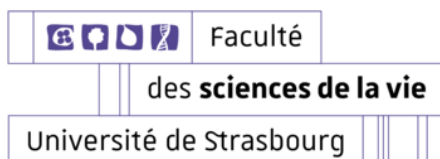
A measurement of optimism in kea
Nestor notabilis

By

Martin Dessart

Under the supervision of M. Taylor Alexander and Mrs. Bastos Amalia

Strasbourg, August 2020

Student contribution

DATA COLLECTION	
Training	100%
Testing	100%
EXPERIMENT DESIGN	
Apparatus design	100%
Session planning	100%
DATA TREATMENT	
Video analysis	0%
Statistics	100%
Report writing	100%

DESSART MARTIN   Master: Ecophysiology,   Ethology,
Ecology

Host Laboratory: Animal Mind Lab

Supervised by Alexander Taylor

Title: A measurement of optimism in kea (*nestor notabilis*)

Keywords: judgment bias, emotional state, nestor notabilis, animal welfare, parrot, vocalisations, emotion

How important is the link between vocalisations and internal emotional states? Emotional state influences cognition by three types of cognitive biases: attention, memory and judgment biases. The judgment bias test (JBT) has been introduced to measure animals' own judgment as an indicator of their emotional state and welfare.

Here, we created a paradigm using a visual JBT where we tested the effect of two vocalisations (warble call as a playful call, hence "positive-valenced" and trill call as an alarm call, hence "negative valenced") on 8 kea (N*estor notabilis*). For this experiment, kea were trained to distinguish the presentation of a positive (rewarding) or a negative (non-rewarding) cue, and then tested on novel ambiguous to probe their judgment of them. Simultaneously with the introduction of each ambiguous cue, kea heard a playback of a vocalisation.

We measured the latency to approach the cue for one positive, one negative and three types of ambiguous cues, using colored flap doors (0%, 20%, 40%, 60% and 80% grey) and a specific apparatus. Using Bayesian multilevel analysis, we did not find strong differences in latency between the two vocalisations conditions. Yet, kea showed a remarkable learning capacity and good consistency throughout the experiment.

Despites the unpredictable behavioural inter-individual differences of kea, this study provides robust support to set up JBT test on Psittacidae and calls for the analysis of animal welfare through the lens of animal emotional state as a support of other most commonly used tests.

Abbreviations

BRMS package: Bayesian Regression Models using 'Stan'

CCC: Concordance correlation coefficient

CIs : Credible intervals

Cond: Condition factor

CS +: Positive cue

CS -: Negative cue

ELPD: Expected Log Pointwise Density

FP: Flap Door

Gp: Group factor

ID: Subject factor

JBT: Judgment Bias Test

Lat: response variable latency

LOO: Leave-one-out cross validation method

MCMC: Markov chain Monte Carlo

OB: Opaque Barrier

PH1/2: Phase 1/2

Pos: Position factor

SE: Standard Error

SP: Starting Position

TB: Transparent Barrier

W1/2/3: Warble call played

T1/2/3: Trill call played

INTRODUCTION

1. A brief history

"Are we open-minded enough to assume that other species have a mental life? Are we creative enough to investigate it? Can we tease apart the roles of attention, motivation, and cognition?" Frans de Wall, 1982.

In the 80s, researchers from different fields began interested on the question of animal intelligence. In 1982, Frans de Waal introduced the concept of "Machiavellian intelligence" in his first book, "Chimpanzee Politics" (De Waal, 1982). His work was followed by Peter Marler, a pioneer in the study of animal communication. He collaborated in a study by Seyfarth et al., (1980) that showed a semantic communication in vervet monkeys (*Cercopithecus aethiops*) and was one of the first to suggest that birds show similar social interactions to primates (Marler, 1982).

This work inspired psychologists Richard Byrne and Andrew Whiten, as well as anthropologist Robin Dunbar. For them, individuals that live in societies and have to deal with multiple and complex relationships should show superior cognitive abilities (Barton and Dunbar, 1997; Byrne RW, Whiten A, 1988). Many scientists from comparative psychology, ethology and physiology, then documented birds' communicative behaviours and showed that their vocalisations can exhibit high levels of complexity, especially in highly social species, as is often the case with corvids and parrots (Emery, 2006).

In social species, complex communication enables sharing of information (e.g. "there is a predator on the ground or in the sky"; Seyfarth et al., 1980) as well as information about the emitter (e.g. age, sex, group membership; Wein et al., 2019). In particular, internal emotional states can be transmitted by vocalisations and have been largely studied (Briefer, 2018).

Hatfield, Cacioppo, and Rapson, (1993) defined emotional contagion as "the tendency to automatically mimic and synchronize facial expressions, vocalisations, postures and movements with those of another person, and consequently to converge emotionally". This definition was primarily defined for humans. Later, researchers found evidence that animals living in groups are also capable of emotional contagion (Briefer, 2018; Panksepp and Burgdorf, 2003; Rozin and Royzman, 2001; Schwing et al., 2017; Špinka, 2012). Emotional components may be recognized by others in social species and could represent a powerful mechanism to share information. For example, it has been demonstrated in situations that require defence against predation or coordination of group members (Špinka, 2012).

As a consequence, emotional contagion leads to state-matching and is useful towards understanding the consistency of social communication. It is well-studied in humans, but how can we consider emotion in non-human animals? (Briefer, 2018).

2. What is an emotion?

Animal emotions can be defined from different perspectives. In his book "Affective Neuroscience: The Foundations of Human and Animal Emotions", Panksepp defines emotion as a small number of discrete states that include rage, panic, fear, seeking, care, play and lust. Thanks to the advances in neurosciences, he assembled a synthetic framework that links these seven basic emotional states with subcortical neural circuits (Panksepp, 1998).

One can also consider emotions as continuous states in two main dimensions: valence and arousal (Russell, 1980) (Figure 1). Many possible associations between these two dimensions are supported in the literature, but the main model proposes emotional arousal as the intensity (e.g bodily changes like activation or excitation) and bipolar valence that comprises positive or negative experience (Rozin and Royzman, 2001) (Figure 1). To study these two dimensions, a framework has been proposed by (Mendl et al., 2010) which assesses emotions in four ways: behavioural, physiological, neural, and cognitive components (Bethell, 2015) (Figure 2).

The first two components are relatively common measures (e.g. locomotion behaviour and cortisol levels, see Neave et al., 2013) and give an indication of emotional intensity, or arousal. The third one extents to link emotion with neural substrate (Paul et al., 2020). Finally, the cognitive component would define the bigger part of the subjective experience, the valence (e.g. a pleasant or unpleasant experience, see Mendl et al., 2009) and is a promising component to evaluate (Roelofs et al., 2016). (Figure 2).

3. How can we measure an emotion?

Emotion can influence cognitive responses and impact three main cognitive biases as follows: perceptual memory, early attention, and judgment (Roelofs et al., 2016) (Figure 2).

For humans, studies present individuals with a paradigm and measure the bias inferred by this paradigm. In terms of judgement bias for example, people in a positive mood make more optimistic judgments during the evaluation of an ambiguous stimulus, while anxious people make more pessimistic judgments (Neave et al., 2013). For example, in a study by Pincus et al., (1996), people were trained to write down a list of ambiguous homophones that they heard, followed by a delayed free recall. Chronic pain participants presented a bias toward pain interpretation as well as a memory
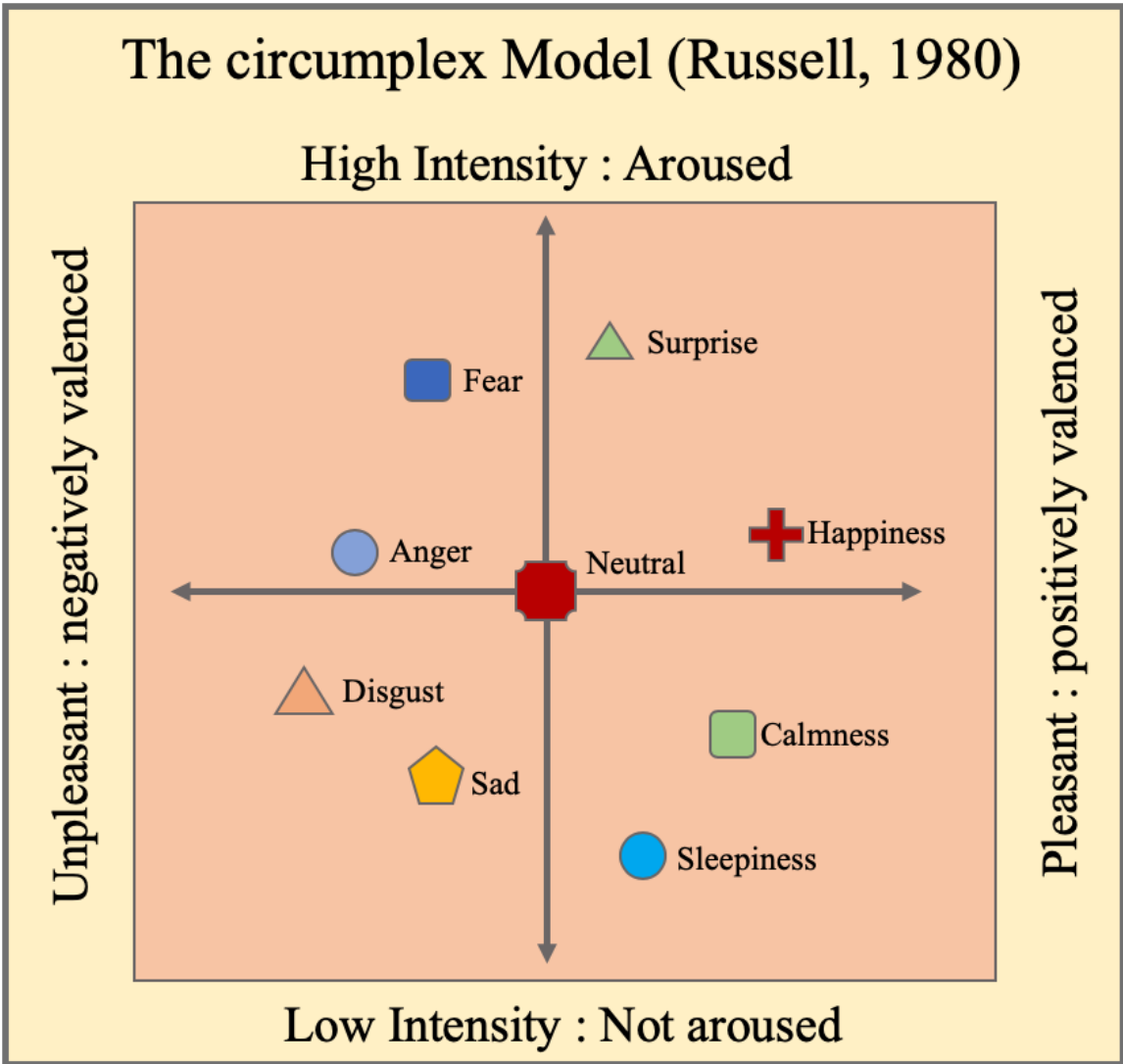
Figure 1: *Model of Emotional state*. Adapted from Russell, 1980.

bias for pain-related stimulus. This study showed that chronic pain is associated with bias in the processing of information.

Non-human animals are non-verbal; thus the method is different. To measure the valence of emotion, researchers focus on the cognitive component and modify one of the three biases we can manipulate (judgement, attention or memory; Neave et al., 2013).

Classical tests consist of testing anxiety by placing an individual in an unfamiliar situation where they will perceive some elements as threatening or adverse. This unconditional stimulus is used to compare a normal situation (without the unfamiliar stimulus) with a negative situation (in presence of the stimulus). The difference of behaviour between the two conditions is then interpreted as a measure of anxiety, but these responses do not represent the entire subjective part of emotion and are subject to flaws and lack of consistency (Ennaceur, 2014).

In contrast, Harding et al., (2004) designed a Judgment Bias Test to test the effect of an "enriched" environment on a group of rats. Although the method requires training, it represents now the most robust way of assessing the emotional state of an animal and is used mainly in psychopharmacology, neuroscience, animal welfare and evolutionary zoology (Roelofs et al., 2016) (Figure 2).

4. Judgement Bias Test (JBT)

A judgement bias test (JBT) provides a cognitive measure of the mood or of the emotional state of an individual by recording behavioural responses to a stimulus presented as ambiguous. For animal cognition, one can define "mood" as a lasting state that influences learning, decision-making and motivation (not explored in this study). It reflects the accumulation of repeated emotional events and is a more diffuse state than an emotion. On the opposite, emotions are intense and short-lived reactions (see Mendl and Paul, 2020; Raoult and Gygax, 2019 for other definitions of mood and emotion).

A JBT consists of training an individual to associate one cue with a positive outcome and another cue with a negative outcome. The positive cue is reinforced (e.g. with a food reward), increasing individual motivation for this cue. Next, an intermediate novel cue is presented in front of the individual and their behavioural reaction is recorded. The results are then interpreted as being either positively or negatively biased in comparison with the behaviour observed during training (Raoult and Gygax, 2019).

At present, JBT methods are varied and can be used to explore a myriad of research questions for many taxon including invertebrates (Perry and Baciadonna, 2017). Different meta-analyses

Emotional State measurement

Components

Valence proportion

Arousal proportion

Behavior
Scan observation
Video recording
Vocalisations playback
Radio tracking

Physiology
Hormone monitoring
Sampling methods
In situ hibridization
Muscle acitivity

Neural
Electrophysiology
Calcium imaging
Single neuron staining
Optophysiology

Cognition

Cognitive mechanism

Attention
Changes in perceptual vigilance and increase focus on negative signals

Memory
Changes in storage, consolidation and retrieval of memories

Judgment
Changes in interpretation, expectation and risk -taking

Cognitive experiments

Classical Test
Open Field Test - Elevated Plus Maze - Novel Object Test
Mesure of unconditional answer to a threatening stimulus

Judgment Bias Test
Training – Cues - Operant task – Reinforcer
Measure of conditional answer to an ambiguous stimulus

Applications

Biomedical Research
Drugs effects
Secondary effects
Housing conditions
Mood/emotion measurement

Animal Welfare
Enrichments effects
Housing conditions
Social conditions
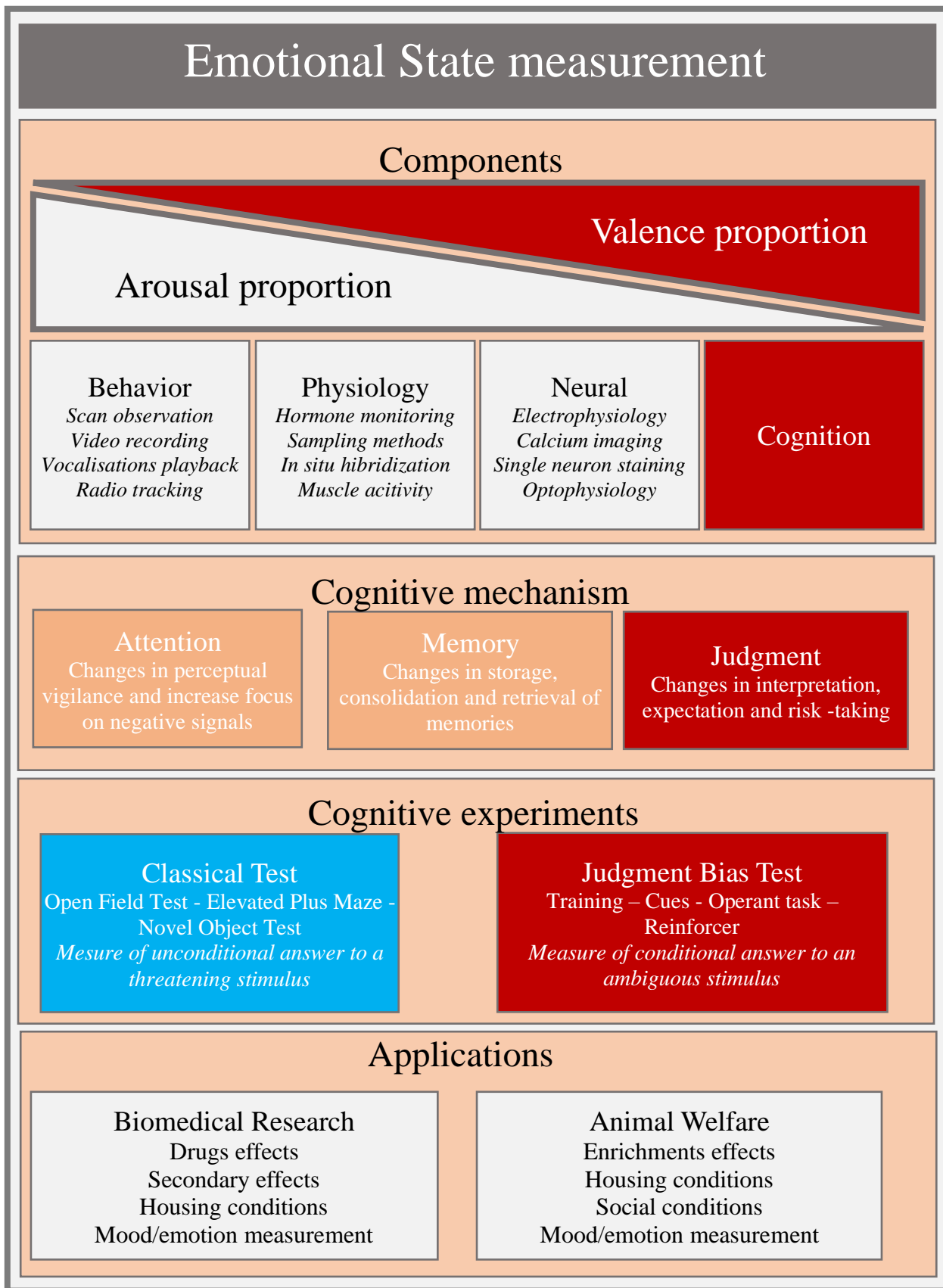Mood/emotion measurement

Figure 2: ***Overview of Emotional state measurement***. From the theory to the application.

highlight the main features and protocols that should be followed, compile data across multiple species and provide an overview of bias management, suitable statistical analyses, and applications (Bethell, 2015; Gygax, 2014; Perry and Baciadonna, 2017; Roelofs et al., 2016).

A large number of JBT studies have been done on mammals, specifically on rodents and ungulates and a few across different birds species (Bethell, 2015). To date, starlings (Brilot et al., 2010; Matheson et al., 2008), hens (Hernandez et al., 2015; Salmeto et al., 2011; Seehuus et al., 2013), canaries (Lalot et al., 2017) and corvids (Adriaense et al., 2019; McCoy et al., 2019) have been tested. Most of the subjects were tested using visual cues and using a Go/No Go design. The conditions were housing enrichment, effect of drugs, effect of sociality, and use of tools.

5. Conditions of applications

The design of a JBT leads to limitations (see the four meta-analyses for more precision; Gygax, 2014; Bethell, 2015; Roelofs et al., 2016; Perry and Baciadonna, 2017). In brief, at least three types of ambiguous cues (one neutral, one between neutral and positive, one between neutral and negative) are required to control for the alternative explanation that animals might associate the ambiguous cues as novelty and not as related as the training cues (Roelofs et al., 2016).

Moreover, two forms of JBT can be designed. Firstly, in Go/No-go methods, individuals are trained to respond to a positive cue and inhibit a response to a negative cue. Secondly, in Active-Choice methods, individuals have to perform an action in response to both cues, each one in a different location. Each option has its own advantages and difficulties, but a Go/no-go method was chosen here and is debated in the discussion.

In addition, the main issue with JBT is the loss of ambiguity as animals can learn to associate ambiguous cues with no outcome (e.g. Brilot et al., 2010). Solutions consist in using a secondary reinforcer, in reducing the reward rate during testing or in reducing the number of ambiguous trials (Brilot et al., 2010).

Finally, other effects have to be taken on board. For social species, isolation can lead to pessimism and a decrease of motivation (Salmeto et al., 2011), two consecutive trials should be clearly separated to avoid inter-influence (Hernandez et al., 2015), or training itself can represent an enrichment that lead to an increase of motivation (Raoult and Gygax, 2019).

6. Kea study

As social parrots that live in large groups, kea (*Nestor notabilis*) could provide an interesting model species for assessing emotional states. Kea are endemic species to New Zealand and known as

the world's only alpine parrots. Parrots differ from songbirds in that they use their tongue to articulate calls and their beak gape to modulate frequency (Raoult and Gygax, 2019).

Kea are known to make several vocalisations which seem to have different roles, and although there is substantial documentation of their call types (Bond and Diamond, 2005; Diamond, J., Bond, A.B., 1999; Wein et al., 2019), little is known about the subjective component of emotion involved and shared during these vocalisations. In a 2012 study, Schwing and colleagues categorized seven structurally distinct call types as follow: screech, chatter, trill, screech-trill, warble, mew and whistle.

In 2017, the same researchers showed that one of these calls, the warble, appears to increase the amount of play among both juveniles and adults, by acting as positive emotional contagion (Schwing et al., 2012, 2017). As suggested for calls in other species as well (Panksepp and Burgdorf, 2003; Vettin and Todt, 2005), this unique vocalisation could be similar to human laughter. Thus, working on the warble call could improve the knowledge on kea behaviour and provide information about kea internal states.

Starting with this paradigm, the goal of this study is to investigate the specific vocalisation "warble" influence on kea emotional state using a Judgement Bias Test. To date, this study will provide the first JBT on the order Pisttacidae and give a first measurement of the subjective emotional component for kea. We designed an experiment with the following expectations. Kea would learn to : i) approach the positive cue (go); ii) avoid the negative cue (no-go), and iii) show a monotonic graded response in regard to the ambiguous cues. Moreover, we expected that: iv) kea would stay steady in their behavioural reaction across the time, and v) the slope of their reaction would differ (it would be shifted) towards the extreme cues, between the two conditions.

The experiment consists of setting up a JBT within the kea aviary at Willowbank Reserve, Christchurch, New Zealand, while the design of the JBT responds to constraints mentioned above.

First, we trained the kea to discriminate one positive and one negative visual cue. Then, a playback of the warble call or a control call was played while an ambiguous cue was presented to the individual.

The vocalisation was used as a vector of emotional contagion and could have an effect on the kea behaviour. We predicted that the warble call would affect kea behaviour and skew their judgment, making subjects more 'optimistic'. Thus, a more 'optimistic' kea would approach the flap door faster than with a control call.

The originally planned six weeks of data collection were reduced to three weeks because of the 2020 Covid-19 pandemic. This meant that subjects received only a single opportunity to participate in each of eighteen test sessions. Hence, we present in this report a partial analysis of an incomplete dataset.

MATERIAL AND METHODS

1. Subjects

The subjects were 8 kea, housed in a large outdoor aviary at Willowbank Wildlife Reserve (see Table 1 for subject details). Subjects were free to come and leave the platform at any time. Food and water were available *ad libitum*. Subjects were tested on individual training platforms within the aviary and rewarded with Hill's Science Diet pellets during trials.

2. Apparatus

Subjects were tested on platforms measuring 125cm x 90cm x 30cm, designed specifically for this study (Figure 3). Subjects were trained to land on the platform and stay at the Starting Position (SP). The Transparent Barrier (TB) prevented subjects from reaching the remainder part of the platform and the Opaque Barrier (OB) prevented subjects from seeing the new Flap Door (FD). The FD is the visual cue, fixed only at the top to allow subjects to put their head through the door to reach the reward placed in the Food Bowl.

3. Cue

Subject learned that the background colour of the cue cards indicated the type of reward (one pellet vs nothing). For the positive and negative cue, 80% black background (hex code #333333) was randomly assigned as CS+ or CS- and 0% black background (hex code #FFFFFF) was randomly assigned as CS+ or CS- (see Table 1). The three intermediate cues used as probes were 60% black background (hex code #666666), 40% black background (hex code #999999) and 20% black background (hex code #CCCCCC). Cues shades are shown in Figure 4.

4. General procedure

A preliminary study conducted in 2018 (Bastos and Taylor, unpublished data) used a spatial discrimination with the kea, but subjects failed to pass criterion for training. Thus, to avoid any learning effects, this time we chose to use visual cues as it is the most common form of judgment bias task used on birds, and more ecologically relevant.

Moreover, we selected a Go/No-Go design with an absence of reward for the negative condition, because of training time constraints and in anticipation of fast learning by the kea. To avoid biases in motivation, ½ pellet was placed just beyond the start line of each trial on the platform. If the bird was not motivated enough, it would not cross the line and we would stop the session (Figure 5).

Due to the small sample size, we chose a within-subject study and birds experienced two conditions. To control for order effects, birds were randomly assigned to two groups (Table 1). Each of the two groups experienced a different condition first: either the warble call condition, or the trill

Subject	Age	Sex	Origin	Positive stimulus	Group	Number of sessions in Phase 1	Number of sessions in Phase 2
Blofeld	7	M	Captive bred	0% grey	2	9	3
Bruce	8	M	Wild	80% grey	2	16	3
Harley Quinn	6	F	Captive bred	0% grey	1	29	2
Loki	6	M	Captive bred	80% grey	1	9	2
Moriarty	6	M	Captive bred	80% grey	1	9	5
Neo	8	M	Captive bred	80% grey	2	6	3
Plankton	6	M	Captive bred	0% grey	2	11	15
Taz	8	M	Captive bred	0% grey	1	19	5

Table 1. Detail of individuals tested in this study. All subjects were parent reared. Ages (as of 2020) are estimated from hatch dates (captive bred subjects) or veterinarian evaluations (wild subjects). Subject were randomly assigned to one of two test groups. All subjects were randomly assigned to a flap colour and a group. The number of sessions columns show the amount of training (where 1 session = 10 trials) received by each bird until they reach the criterion for the next phase.
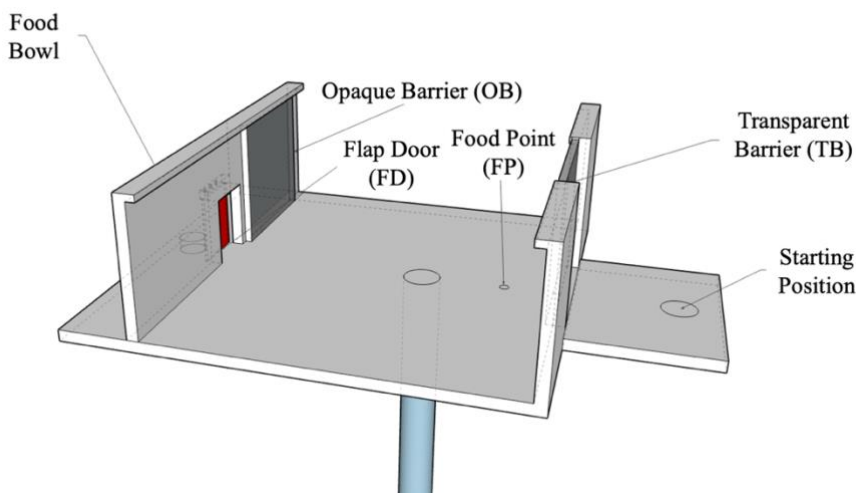

Figure 3: *Apparatus.* Legend: Food Point (FP), Opaque Barrier (OB), Transparent Barrier (TB), Flap door (FD).

Figure 4: *Five types of cues used in this study.* One positive cue and one negative cue (80% and 0% dark) and three ambiguous cues (60%, 40% and 20% dark).

call condition. To ensure that birds didn't learn that ambiguous cues were not rewarded, testing days were interspersed with non-testing days. This allowed us to keep a track of each individual's performance as well as space out testing days and ensure that birds' motivations remained constant.

Testing and training used the same general procedure. At the start of each trial, the experimenter invited the kea onto the testing platform. Once kea stood behind the plexiglass across the starting line, the experimenter positioned the Flap Door of the correct colour, removed the opaque barrier and allowed 15 seconds for kea to look at the door. (Figure 5).

Then, ½ pellet was placed at the "food point" (see Figure 3) to ensure that subjects were motivated to cross the starting line even in negative training trials. The transparent barrier was then removed. Once it was removed, the subject was allowed 30 seconds to approach the flap door. A second experimenter started a timer from the moment they stepped past the starting line. They stopped the timer either when the bird touched the flap or once 30 seconds had elapsed (Figure 5).

The birds needed to reach the positive cue (e.g. "beak touch" the flap door) in under five seconds or the negative cue in over 20 seconds for the trial to be considered successful. If the bird flew off the platform, or did not eat the ½ pellet, the whole session was stopped. Time to approach was coded from video data and recorded for each test trial (Figure 5).

Apparatus habituation

Kea were positively reinforced for landing and remaining on the testing platforms for thirty seconds. They were next encouraged to touch the flap door with their beak and ate the pellet placed behind the flap.

Playback set up

A speaker (Amplifier AAO420 with 2x 80 watts RMS units) was set up next to a high-ranking individual's (Taz) platform. His platform was chosen as well as his calls because he was the highest of the hierarchy with accessible calls (Walker C, Demler C, Nelson X., 2018, Appendix). On testing days, the 2 vocalisations played (trill and warble) were randomly selected from a database of this individual's calls. One call was then cut and repeatedly played for 15 seconds total, where each of five call repetitions was 1 second long, with a 2 second gap between repetitions.

Playback habituation

For the habituation, a single screech call recording was played over 15 seconds. As with test cue, the call was 1 second long and we left a 2 second gap between each of the five repetitions. All

- Change Flap Door (FP)
- Cue S+: put a pellet
- Cue = S- : fake put a pellet
- Ambiguous cue : tape the flap to the door and fake put a pellet

- Remove Opaque Barrier (OB)
- Wait 15 seconds or
- Play the 15-seconds call for ambigous
- Put ½ pellet at the Food Point (FP)
- Remove Transparent Barrier (TB)

- Wait until kea reach the FD or 30 sec
- Stop the time and close OB
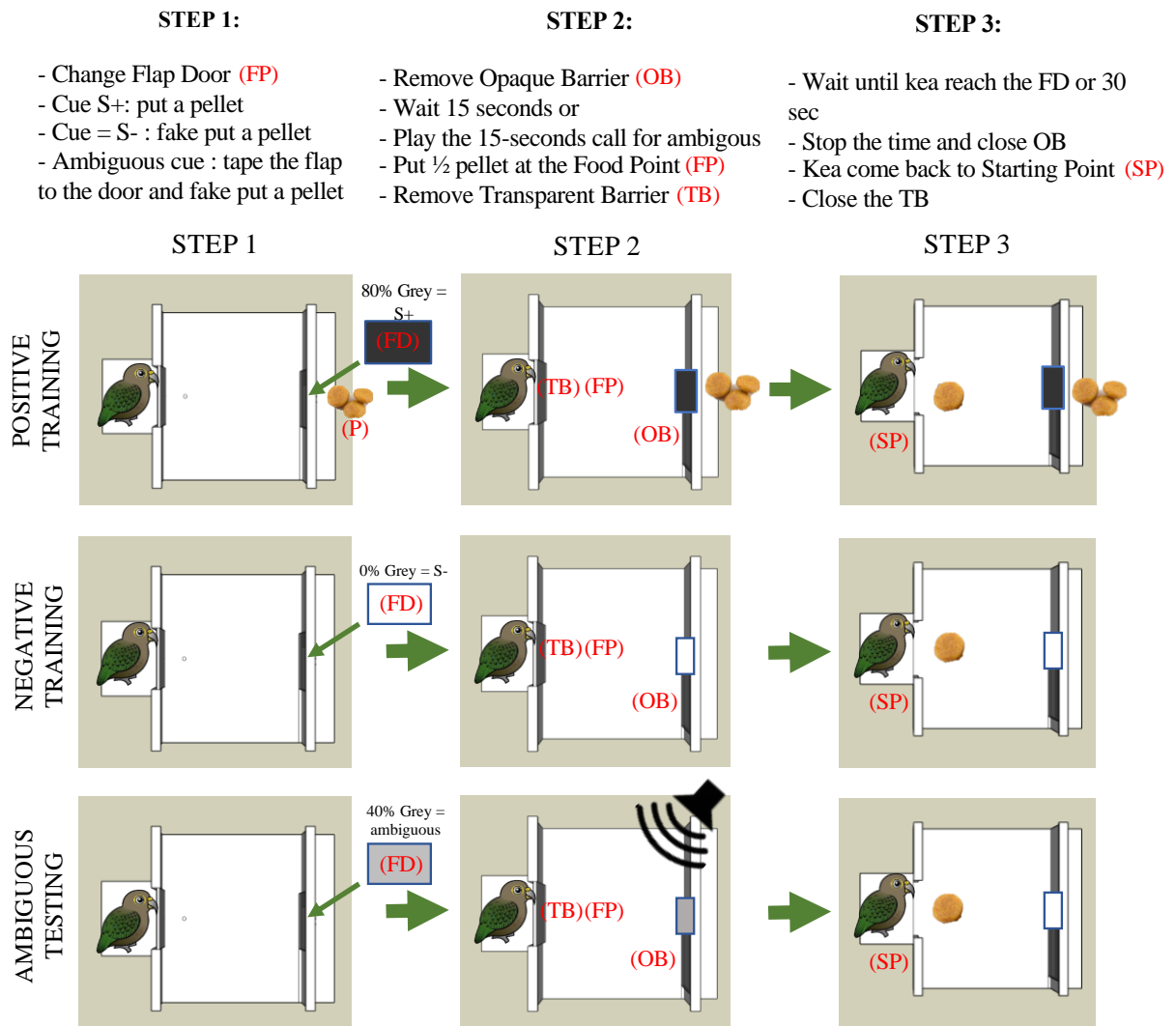- Kea come back to Starting Point (SP)
- Close the TB

Figure 5: ***Trial process for training and testing.*** Three steps are required to run one trial. The situation is exposed following for the positive cue (S+), then the negative cue (S-) and finally the ambiguous cue.

birds remained on the platform, showing they were comfortable to remain on the platform whilst a call was being played from the speaker.

5. Training

Phase 1

Using the procedure described above, trials were presented in two blocks of five successive negative and positive trials. Birds had to reach a criterion of 10/10 during one session to proceed to training phase 2. A planning summary is showed in Figure 6.

Phase 2

Trials were presented in 10 pseudorandomized negative and positive trials. No more than 2 successive same trials were presented in a row. Birds had to reach a criterion of 17/20 during two consecutive sessions to proceed to testing phase. A planning summary is showed in Figure 6.

6. Testing

Testing schedule

Birds were separated into two randomly assigned groups. The trial order was the same within the group. To avoid cross-over effects in mood between different calls, the entire aviary could only experience one type of call per day (trill or warble).

Each subject experienced each individual call three times. For each call, three ambiguous cues were presented for each individual. To reduce the number of calls played across the aviary, birds were tested in pairs using two platforms simultaneously. A planning summary is showed in Figure 7.

Non-Testing days

To ensure that birds did not learn any specific order, two types of training were presented. The first type, 10-block training, was similar to training phase 2. Each individual experienced one session of 10 pseudorandomised trials. Criterion to succeed was 9/10 for a maximum of 3 consecutive sessions. If the bird failed during these 3 sessions, training was over for that bird on that day.

The second type, 4-block training, was similar to testing days. Each individual experienced one session of 4 interspersed trials but without the 5_{th} ambiguous trial. Criterion to succeed was 4/4 for a maximum of 3 consecutive sessions. If the bird failed during these sessions, training was over.

Training procedures slightly varied for birds that showed inconsistent performance on test days: during a testing day, if a bird failed 3 times and had not experienced the 3 ambiguous cues, it was labelled as an "inconsistent bird".

Training phase

	Apparatus and playback habituation	Phase 1 (PH1): two blocks of 5 trials			Phase 2 (PH2): 10 trials of pseudorandomised trials						
Blofeld		PH1	PH1	PH2	PH2						
Harley					PH1	PH1	PH1	PH1	PH1	PH1	PH2
Bruce			PH1	PH1	PH1	PH1	PH1	PH1	PH2		
Loki		PH1	PH1	PH2	PH2	PH2					
Moriarty			PH1	PH1	PH1	PH2	PH2	PH2			
Neo		PH1	PH1	PH2	PH2	PH2	PH2				
Plankton		PH1	PH1	PH1	PH2	PH2	PH2	PH2	PH2	PH2	PH2
Taz			PH1	PH1	PH1	PH1	PH1	PH1	PH2	PH2	
Dates	27/01	29/01	30/01	31/01	03/02	04/02	05/02	06/02	07/02	10/02	14/02

Figure 6: ***Planning of the training procedure***. Each bird succeeded in the two training phases. Some birds could experience the two phases on the same day (e.g. Blofeld on the 31/01/2020: he first reached the criterion of the Phase 1 before to experience Phase 2).

Testing phase

	18/02	20/02	21/02	24/02	25/02	27/02	28/02	02/03	03/03	04/03	05/03	06/03	09/03	11/03	12/03	13/03	16/03	Total	
Blofeld	T1 T2 T3	10	W1 W2 W3	10	10	10	W1 W2 W3	10	T1 T2 T3	4		10	T1 T2 T3	10	W1 W2 W3	4	4	S1 / S2 / S3	6 / 6 / 6
Harley	10	4		10	T1 T2 T3	10	4	4		10	10		10	W1 W2 W3	4	10	T1	S1 / S2 / S3	3 / 2 / 2
Bruce		T1 T2 T3	W1	10	10	10		10		4		10	T1 T2 T3	10	W1	4	4	S1 / S2 / S3	4 / 2 / 2
Loki	10	4	W1 W2 W3	10	T1	10	4	4	T1 T2 T3	T10	T10	W1 W2 W3	10	W1 W2 W3	4	10	T1 T2 T3	S1 / S2 / S3	6 / 5 / 5
Moriarty	10	4	W1 W2 W3	10	T1 T2 T3	10	4	4		T10	T10	W1 W2 W3	10	W1 W2 W3	4	10	T1 T3	S1 / S2 / S3	5 / 4 / 5
Neo	T1 T2 T3	10	W1 W2 W3	10	10	10		10		4		10	T1 T2 T3	10	W1 W2 W3	4	4	S1 / S2 / S3	4 / 4 / 4
Plankton		T1 T3		10	10	10	W1 W2 W3	10	T1 T2 T3	4		10		10	W1 W2 W3	4	4	S1 / S2 / S3	4 / 2 / 4
Taz	10	4		10		10	4	4		10	10	W3	10	W1 W2 W3	4	10	T3	S1 / S2 / S3	1 / 1 / 3
Dates:	18/02	20/02	21/02	24/02	25/02	27/02	28/02	02/03	03/03	04/03	05/03	06/03	09/03	11/03	12/03	13/03	16/03		

Figure 7: ***Planning of the testing procedure***. Each bird has experienced T10 (pale grey), T4 (dark) and testing days. W is for warble condition and T for trill condition. Bird should have experienced 3 ambiguous (W or T 1, 2 and 3 per testing day, except if they failed during the previous for Positive or Negative trials. Blofeld, Loki, Moriarty, Neo and Plankton (in green) were consistent as they experienced a total of more than 10 ambiguous cues across the Testing session. Other birds are were more inconsistent.

For subjects classified as inconsistent, if they approached the flap door in under 20 seconds during a negative trial, the experimenter closed the platform and waited 30 seconds before the next trial.

Testing days

Kea were first presented with a single block of four interspersed negative and positive trials. All birds had to succeed in all the four trials to be tested. If a bird failed to pass this criterion of 4/4, the platform was removed and at least 10 min were spent before we attempted to test that bird again. After 3 consecutive failures, testing was ended for that bird that day.

If the bird succeeded, we proceeded to a fifth trial where a new flap door was introduced (60%, 40% or 20% grey, Figure 4, Figure 5). Unlike in negative and positive trials, the flap door was now taped shut to ensure the bird could not access the food bowl even if they attempted to open it. If the bird did not approach, the trial was stopped at 30 seconds. If the bird did approach, the experimenter quickly interrupted and repositioned the opaque barrier right after the bird beaked touch the door. That way, birds never gained any information about the contents of the food bowl in any ambiguous trials.

After changing the flap colour and fixing it to the door (Figure 5), the experimenter on the platform waited for the experimenter in charge of the call playbacks.

7. Experimenters

Five experimenters were present for all testing sessions (two experimenters were working on the two platforms; two experimenters were filming the tests and one experimenter was in charge of the call playbacks). The two experimenters working on platforms were blind to the experimental conditions and hypotheses.

8. Analyses

Data collection

A blind coder measured the approach time in all videos. 20% of his coding was checked for inter-observer reliability with another blind coder. Lin's concordance correlation coefficient (CCC) was calculated using DescTools package and the function CCC. Concordance correlation coefficient was excellent, ρ=99,18% $\pm$ 0.004.

Statistical Design

In order to analyse the subjects' latency to approach the door, we fitted multilevel Bayesian models, as advised by several authors for nested data (e.g. Gygax, 2014; Mendl and Paul, 2020). Moreover, model estimation using Bayesian statistics allows for a better understanding on the quantification of uncertainty and on the size of the desired effect and allows to incorporate explicitly prior knowledge about predictors into the model (Kruschke and Liddell, 2018).

Our analysis process included firstly the determination of priors based on assumptions about our experimental design and the determination of our model factors, secondly the comparison of these models and the choice of the optimal model, and finally the analysis of the posterior distribution, the result distribution predicted by the model.

We worked on R software (version 3.6.3) with the package brms (Bürkner, 2017) that uses the language Stan (Bürkner, 2017). Stan uses Hamiltonian Monte Carlo algorithms, a sampling method based on random probability distributions. The process uses a Markov chain, that is a sequence of successive probabilities, where each probability is weighted based on the previous probability. When the distribution is stationary, that is to say for N probabilities, the distribution of probability is $\pi = (\pi_1, \pi_2, ..., \pi_N)$, the Monte Carlo algorithm can estimate the distribution by iterating each step of sampling (Vehtari et al., 2017). The response variable "latency" is consequently modelled using a Weibull distribution rather than the standard normal. The Weibull distribution is frequently used in survival analysis to model positive real time-to-event data and it takes two main parameters: k parameter, known as the slope, and λ parameter known as the scale.

$$\text{Weibull survival distribution: } f(T; k, \lambda) = e^{-\left(\frac{T}{\lambda}\right)^k}$$

We also accounted for the fact that many of our observations were right-censored, as we stopped recording latency at 30 seconds by including in the dataset a Censored factor. Indeed, there is a relevant difference between a latency measure and a behaviour response measure that needs to get into our models in an explicit way (Gygax, 2014).

We used a step-up method and first created a null (intercept-only) model with the dependent response latency and a random intercept over individual birds to deal with non-independence in our data. We fitted our intercept parameter using a weakly regularising Student-t prior with 3 degrees of freedom, 1.8 as location parameter and 0.4 as scale parameter. We then added a Student-t prior for the standard deviation of the random intercept using 0 as location parameter and 0.4 as scale parameter. These priors were decided on by using prior predictive simulation.

We then constructed models of increasing complexity. First, we included the variable Cue (cue) that comprised 5 levels: "0" as negative cue, "0,25" as the near-negative ambiguous cue, "0,5" as the middle ambiguous cue, "0,75" as the near-positive cue and "1" as the positive cue. We changed our model syntax to include cue as a fixed effect and we added this factor into the random part of the model to add a random slope of Cue over individual birds. This allowed us to take into account the within-subject design of the study - each bird experienced every cue. In a separate model, we included the variable Condition (cond), that comprises two levels: "warble" and "trill". This model also included random slopes over birds. In a final model, we explored the interaction between cue and condition, which included further random slopes over birds.

We were also interested in how other experimental variables influenced the birds' latency and so we fitted further models. We constructed one model with the variable group (gp) which represents a between-subject condition, given that each bird was randomly assigned to one of these two categories. This variable represents two pseudo-randomized groups starting with different conditions (warble first or trill first).

We repeated this process with the variable position (pos) which was either: positive = 0% grey colour and positive = 80% grey colour. We also looked at the variables session (ses) and trial (trial) as fixed effects. A summary of our models figured in Table 2.

After constructing our models, we focused on model estimation and comparison rather than engaging in null-hypothesis significance testing (Kruschke and Liddell, 2018). Firstly, we used the Leave-One-Out Cross Validation (LOO) to compare our models. This method is reviewed in Bürkner, (2017) and consists of finding the model that is the best at prediciting out-of-sample data. This methods is preferred because it indicates a better fit as well as providing an estimate of the uncertainty (in the form of standard deviation and credible intervals) (Bürkner, 2017).

We also analysed the posterior distribution of our models and calculated differences between our factor levels. Posterior distribution comprises 4000 estimations made within 4 MCMC chains and allowed us to generated approach latencies as well as the uncertainty of each effect.

9. Ethics Declaration

This research was conducted under ethics approval from The University of Auckland Ethics Committee (reference number 001816). Our research was conducted in accordance with the New Zealand National Animals Ethics Advisory Committee guidelines.

RESULTS

1. Training

Phase 1 and 2:

Among the 9 birds, only Spike failed to pass the first criterion. As birds were free to come, stay and leave the platform as they liked, Spike's 3 visits over the course of 2 weeks led to his exclusion from the experiment. The 8 other birds reached the phase 1 with a criterion of 10/10 within an average of 5.25±0.88 (mean ± SEM) days. Then, they reached the criterion 17/20 of phase 2 within an average of 2.87±0.51 days.

Overall, Harley was the slowest bird to reach training criteria, with a total of 29 sessions of 10 trials, and Moriarty was the fastest bird with a total 14 sessions of 10 trials across the two training phases. During phase 1, The mean latency approach was 17.25±0.52 seconds for negative cue and 4.28±0.17 seconds for positive cue (Figure 8). During phase 2, The mean latency approach was 23.34±0.63 seconds for negative cue and 3.75±0.15 seconds for positive cue. (Figure 8).

Finally, for the negative cue, the birds failed 43.5% of the time to wait at least 20 seconds and succeeded in waiting for 56.5% of all negative trials. For the positive cue, the birds only failed to reach and touch the flap door within 5 seconds in 14.6% of trials (they were successful 85.4% of all positive trials).

We found a decrease of the average approach latency across time for the positive cue, meaning that kea improved their performance over multiple trials. On the other hand, the average approach latency increased for the negative cue during Phase 1, also suggesting improved inhibition over multiple trials. Overall, Phase 2 latencies were better than Phase 1, indicating they transferred their knowledge between training phases (Figure 8). Finally, average approach latency for the negative cue during Phase 2 is decreasing, indicating an overall fall on training performances, but this effect is due to two individuals, Taz and Plankton, that reached Phase 1 criterion lately in comparison with the rest of the group and were not consistent at the beginning of Phase 2. Figure 9 showed the average approach latency for the negative cue during Phase 2 with and without these two individuals.

In between Testing:

Trainings in between testing sessions were important to evaluate bird's performance. In total, 25 sessions of 4 or 10 trials were conducted for 18 days. For T10 sessions, the mean latency approach was 23.12±0.61 seconds for negative cue and 3.72±0.14 seconds for positive cue (Figure 10). For T4 sessions, the mean latency approach was 24.83±1.14 seconds for the negative cue and 3.87±0.38 seconds for the positive cue (Figure 10). On T10 sessions, birds were successful in 71.5% of trials
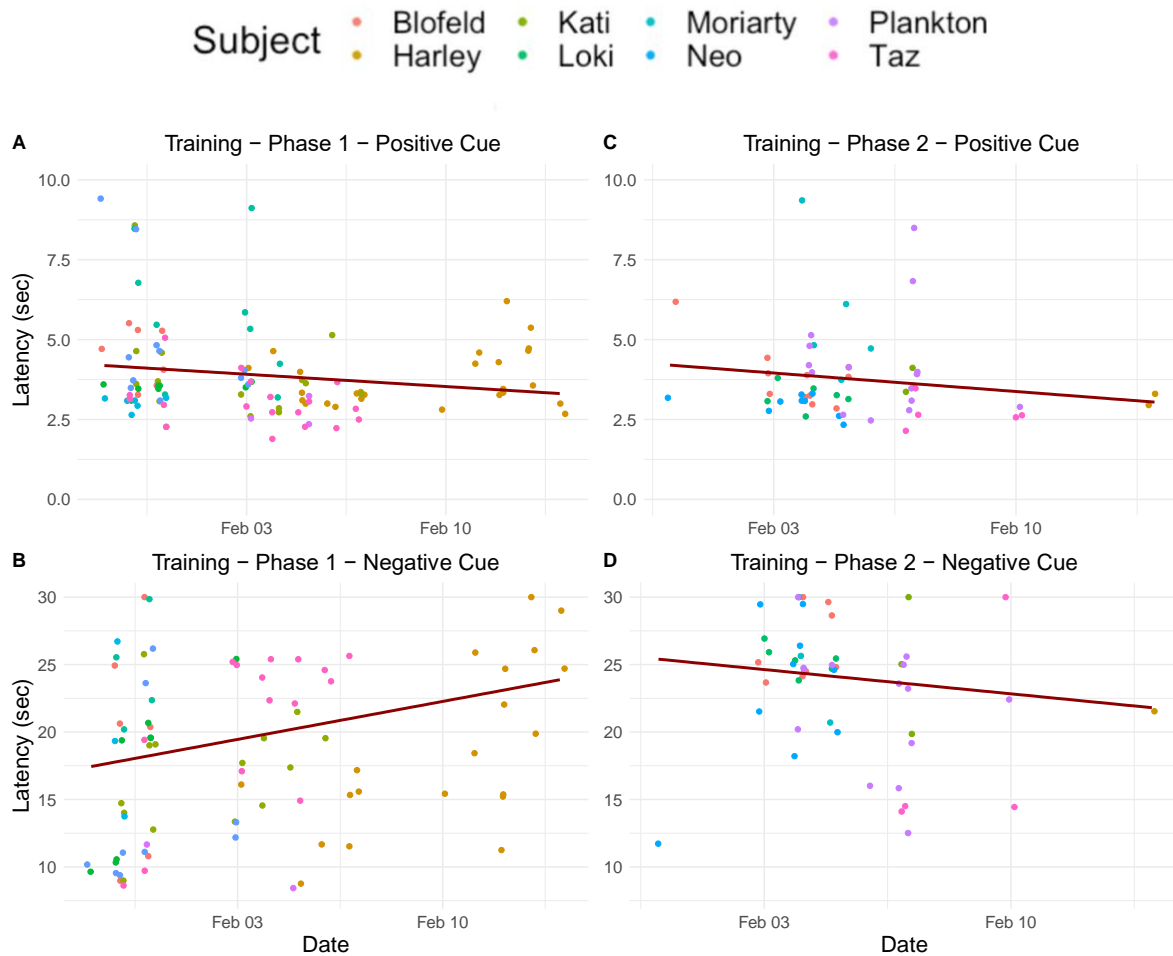
Figure 8: **Training results.** Kea learnt to discriminate positive and negative cues. (A) and (B) give the average approach latencies for each subject across the time for the Phase 1 of the training part, for the positive cue (A) and the negative cue (B). (C) and (D) give the average approach latencies for each subject across the time for the Phase 2 of the training part, for the positive cue (C) and the negative cue (D). In red: linear regression line for the average of latency for all subjects.

for the negative cue, and 89.5% of trials for the positive cue. On T4 birds were successful in 75.1% of trials for the negative cue, and 94.4% of trials for the positive cue.

Despite several failed trials, especially for the negative cue, birds were consistent across the time during the training in between testing sessions. There was a light increase of latency for T10 negative cue, that indicated a robust response of the birds (Figure 10).

2. Effect of Condition factor

To evaluate the effect of the factor Condition, we constructed and compared several models. Firstly, adding the factor Cue greatly improved the fit over the null model (ELPD-LOO difference ± 95% credible interval: 569.9 ± 28.2, Table 2). Secondly, adding the factor Condition also slightly outperformed the null model (1.9 ± 0.4, Table 2). Lastly, adding the interaction between the factor Condition and Cue outperformed all other models (see Table 2). However, the small difference between model 2 that comprises Cue factor only and model 4 that comprises the Cue and Condition interaction revealed that allowing Condition factor to interact with Cue did not seems to improve the fit further (Table 2).

Model 5 was analysed more in details by further analysing posterior's distributions. Appendix 3 depicts the estimations of this model for the intercept, the standard deviation of the Condition factor and the correlation between these two conditions. This plot is a first view at the MCMC method and shows that intercept and standard deviation for the two conditions seem to be very similar. On the right, the chain process is confirmed as we did not have divergent transitions (i.e. the Monte Carlo algorithm calculated a value very different from the initial value or all the successive values are similar). Indeed, chains were well mixed within a stable interval for all parameters (Appendix 3). Moreover, the Rhat value, indicator of information on the convergence of the MCMC algorithm, is 1.00 for each parameter (Bürkner, 2017).

To formally compare the effect of Condition factor, we calculated the posterior difference, that is the proportion of MCMC samples in the posterior distribution, reflecting the latency probability, between the Warble Condition and Trill Condition.

Overall, the interaction model predicted that the birds took 14.34 seconds [10.23 18.94] (median [95% Credible intervals]) to approach the ambiguous cues in the Warble condition, and 13.21 seconds (95% CIs [9.23 16.90]) in the Trill condition. These latencies did not differ (median posterior difference = 1.13, 95% CIs [-0.23 2.46]).

Going further, we calculated this percentage of posterior differences for each cue. The model predicted that birds took 3.48 seconds (95% CIs [3.26 3.69]) to approach the positive cue (0) in the
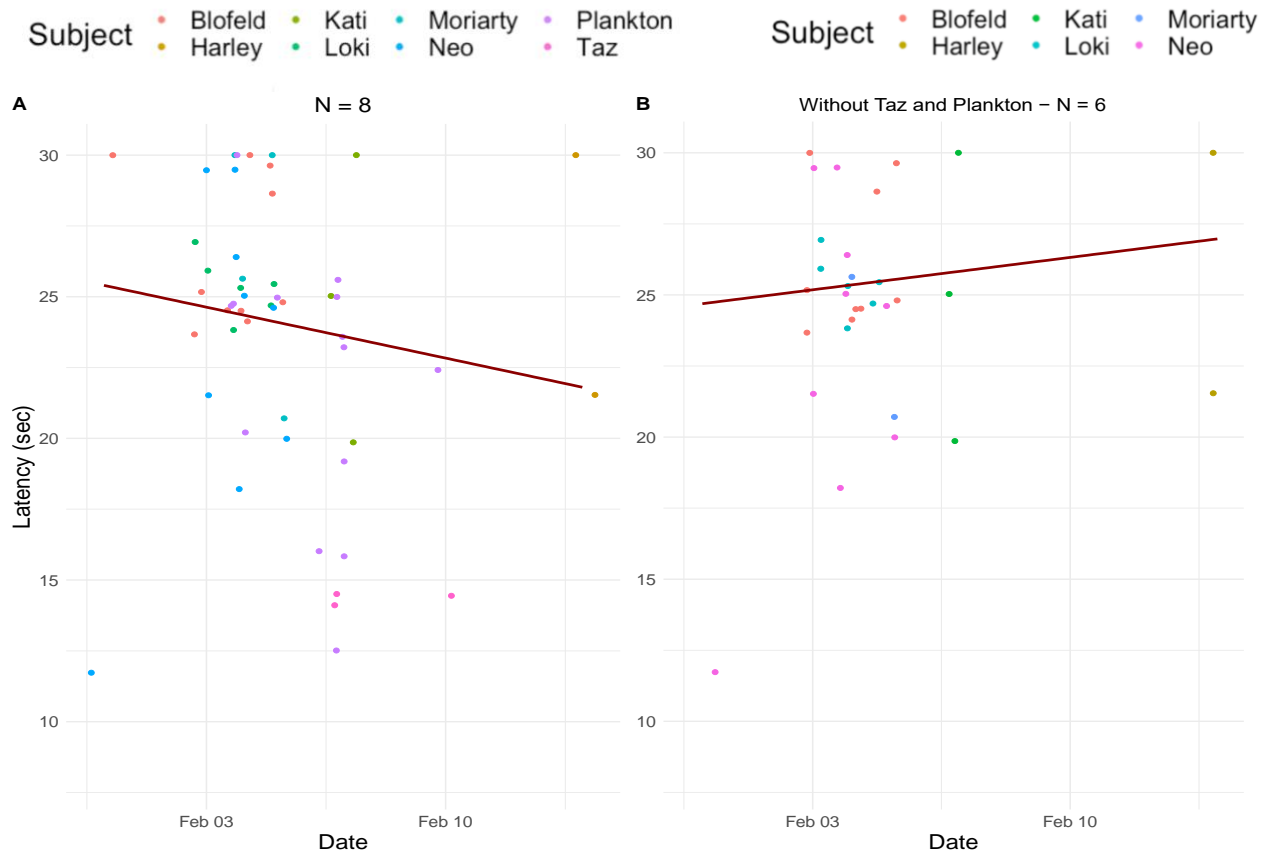
Figure 9: **Phase 2 Training results.** Two individuals (Taz and Plankton) were not consistent at the beginning of the Phase 2. (A) and (B) give the average approach latencies for each subject across the time for the Phase 2 of the training part, for all individudals (A) and without Taz and Plankton (B). In red: linear regression line for the average of latency across all subjects.

Model	ELPD	SE	ΔELPD	Formula
m1	-1030,6	28,0	-575,1	Latency | cens (censored) ~ 1 + (1 | ID) , family = Weibull
m2	-460,8	33,4	-5,3	Latency | cens (censored) ~ 0 + cue + (0 + cue | ID) , family = Weibull
m3	-1032,6	27,9	-577,1	Latency | cens (censored) ~ 0 + cond + (0 + cond | ID) , family = Weibull
m4	**-455,5**	**33,2**	**0,0**	Latency | cens (censored) ~ 0 + cue:cond + (0 + cue:cond | ID) , family = Weibull
m5	-1031,5	28,1	-576,0	Latency | cens (censored) ~ 1 + gp + (1 | ID) , family = Weibull
m6	-1031,1	28,1	-575,6	Latency | cens (censored) ~ 1 + pos + (1 | ID) , family = Weibull
m7	-1031,6	28,1	-576,1	Latency | cens (censored) ~ 1 + ses + (1 | ID) , family = Weibull
m8	-1033,6	28,5	-578,1	Latency | cens (censored) ~ 1 + trial + (1 | ID) , family = Weibull

Table 2: summary of models run with the package BRMS in R (version 3.6.3). "ELPD" is the theoretical expected log pointwise predictive density for a new dataset which is estimated using cross-validation. It represents the sum of log-transformed probabilities densities estimated by the model. The higher a model's ELPD score, the better the model will fit the data. SE is the standard error, a score that describe the uncertainty about the predictive performance. ΔELPD and ΔSE are the differences between the best model (XX) and the actual model.

Warble condition, and 3.77 seconds (95% Cis [3.24 4.45]) in the Trill condition. Latencies did not differ (median posterior difference = 0.29, 95% CIs [-0.26 1.02).

For the near-positive cue (0.25), model predicted that birds took 4.68 seconds (95% CIs [3.41 6.42]) in the Warble condition, and 4.31 seconds (95% CIs [3.64 5.25]) in the Trill condition. Latencies did not differ (median posterior difference = 0.36, 95% CIs [-1.20 2.26]).

For the middle-ambiguous cue (0.5), model predicted that birds took 8.17 seconds (95% CIs [4.32 17.06]) in the Warble condition, and 7.82 seconds (95% Cis [4.31 15.43]) in the Trill condition. Latencies did not differ, even if the distribution was much larger (median posterior difference = 0.38, 95% CIs [-7.69 9.21]).

For the near-negative and the negative cue, the model begins to predict latencies above 30 seconds. This is because the model correctly takes into account the right-censored nature of the variable latency, and therefore makes predictions for latencies greater than 30 seconds (i.e as if the experimenter had not stopped recording the latency.

For the near-negative cue (0.75), model predicted that birds took 36.98 seconds (95% CIs [27.54 58.09]) in the Warble condition, and 28.07 seconds (95% Cis [18.69 39.26]) in the Trill condition.

Finally, for the negative cue (1), we chose to not report the predicted latencies. Indeed, we could not trust the predictions as 96,8% of the latency measurements from the Warble condition and 100% of the latency measurements from the Trill conditions were censored at 30 seconds.

In addition, we took advantage of the random effects approach by comparing the predicted latencies for the middle-ambiguous cue (0.5) by individual (Table 3). We found large inter-individual differences with three individuals (Harley, Moriarty and Blofeld) that are estimating judging the middle-ambiguous cue as a negative cue. Blofeld is particularly constant with 30 seconds latency to all middle-ambiguous cues.

Beyond the non-reliable estimated latencies for the near-negative and the negative cue, Figure 11 shows the average of latency predicted by the model 5 in function of the cue. The latencies are shaped as a slope from the positive (0) cue to the negative (1), which confirmed the expectation that iii) kea exhibited constant behavioural reactions to the cues presented.

3. Effect of other factors

To calculate the effect of other factors, we ran several models (Table 2). The model m5 provides useful support to analyse the effect of Group factor. This factor comprises two pseudorandomised groups that define the type of vocalisation played first. On a log scale, the
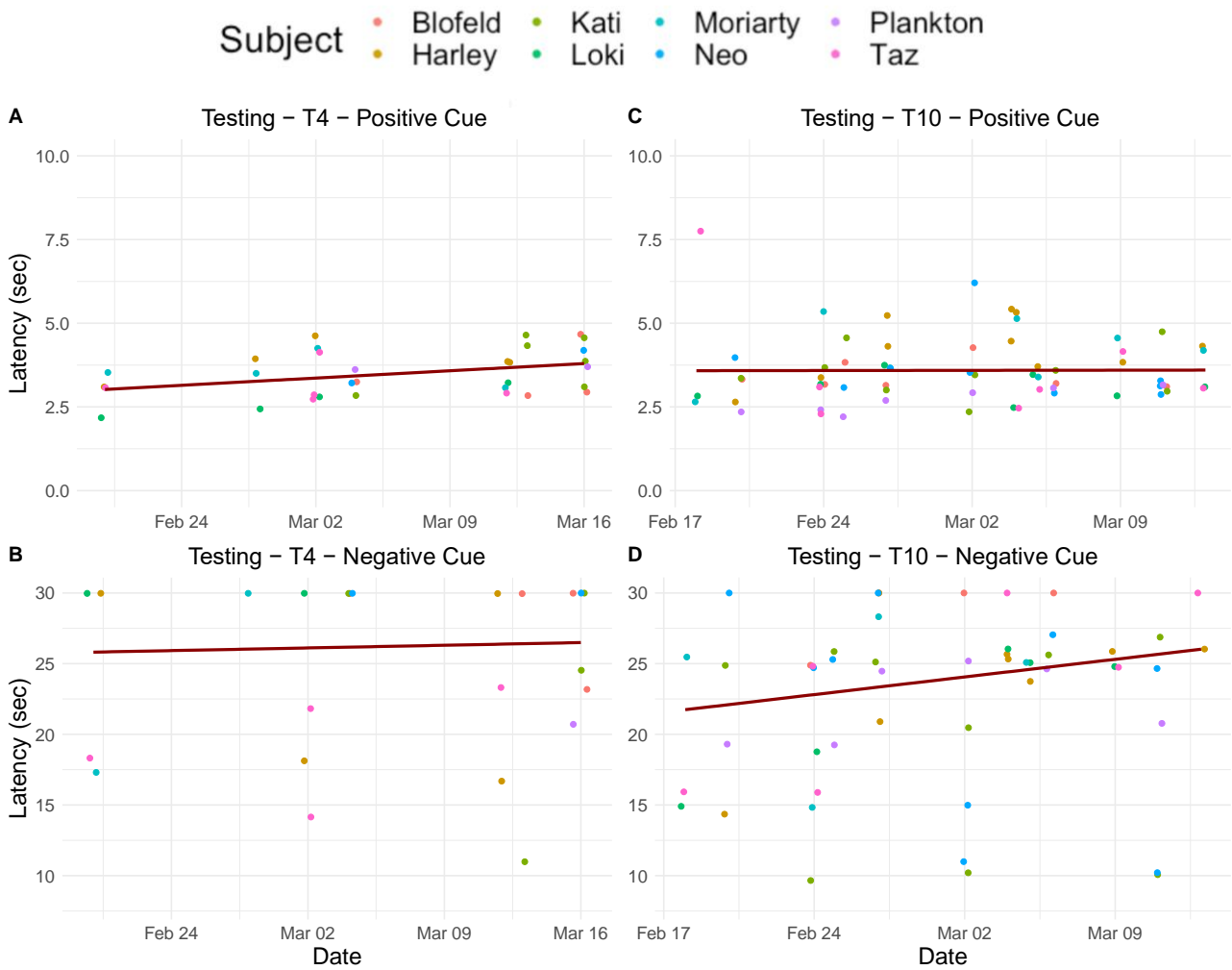
Figure 10: **In between testing results.** Kea were consistent across the testing part. (A) and (B) give the average approach latencies for each subject across the time for T4 trainings, for the positive cue (A) and the negative cue (B). (C) and (D) give the average approach latencies for each subject across the time for T10 trainings, for the positive cue (C) and the negative cue (D). In red: linear regression line for the average of latency for all subjects.

posteriors differences between the two groups were distributed around 0 (median posterior difference = 0.07, 95% CIs [-0.51 0.47]), hence we did not find difference between groups. Similarly, model m6 provides support to analyse the effect of Position factor. This factor defines which colour is associated with the rewarding cue and the non-rewarding cue. On a log scale, posteriors differences between the two groups did not differ (median posterior difference = 0.18, 95% CIs [-0.65 0.29]), hence we assumed no effect of Position factor.

We then ran two models to see if session or trial influenced kea behavior. Model m7 included Session factor only. The posteriors predictions figured on Table 4. Similarly, m8 included Trial factor and the posteriors predictions figured on Table 5. No consistent differences were found either on the session or on the trial numbers.

Finally, in order to investigate trials consistency, we show in Figure 12 the average approach latency recorded during all training trials, that comprises both Phase 1 and Phase 2 training, plus between-test T10 and T4 trainings. The smoothing line was created using "LOESS" local method regression. For each training phase and only on Negative cues, the graph points out Trial 1 and Trial 2 smaller latencies.

Subject	Estimate	Est.Error	Q2.5	Q.97.5
Loki	6.08	0.801	4.79	7.92
Harley	31.6	11.0	19.2	60.6
Moriarty	38.8	12.3	24.5	69.2
Taz	6.32	2.79	3.49	13.8
Neo	4.90	0.786	3.74	6.75
Plankton	3.51	0.848	2.38	5.54
Kati	4.64	1.10	3.16	7.33
Blofeld	97.6	189.	34.4	339.

Table 3: Summary of predictive latencies (sec) for each subject with 95% credible interval. When the estimation error is high, we could not trust the predictions (e.g. Blofeld) because most of the latencies were censored at 30 seconds.
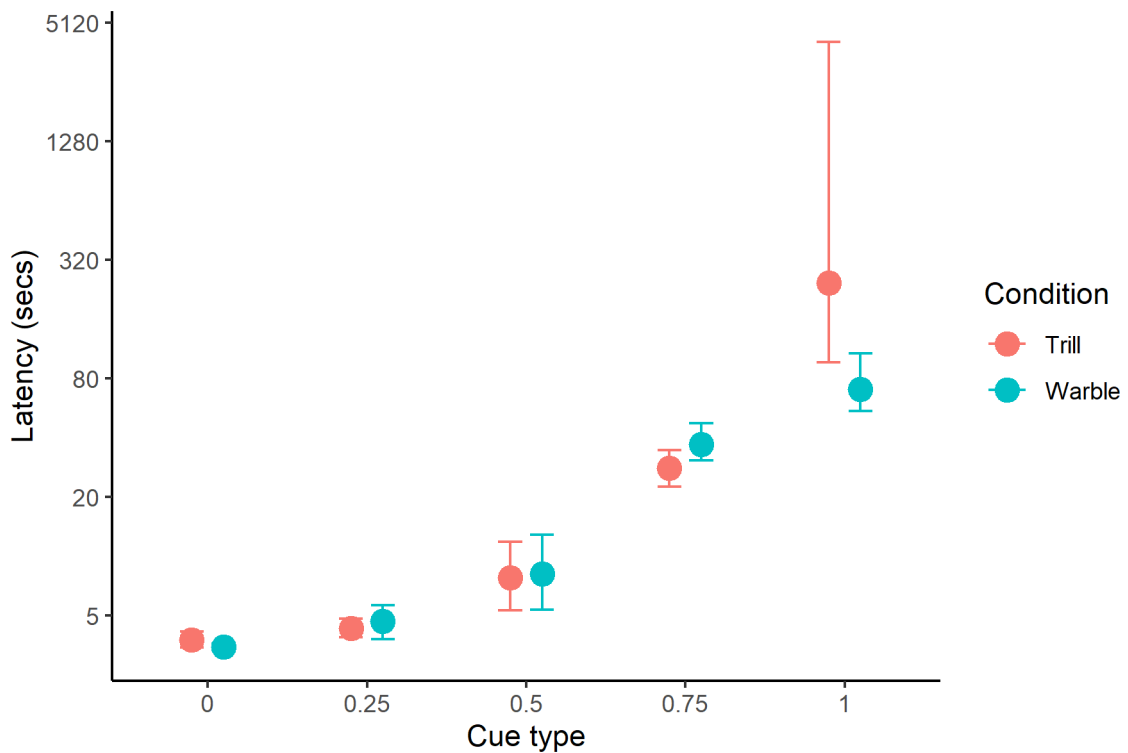
Figure 11: **Average of latency estimated by model 5 in function of the cue type (0: Positive, 0,25: near-Positive, 0,5: neutral, 0,75: near-Negative, 1: Negative).** The latency is above our 30 seconds-limit for Condition 0,75 and 1 because of the right censored nature of our data (when latency=30sec). 98,33% of our data is censored when Cue = 1, 96,80% when Condition = 1 and 100% when Condition = 0.
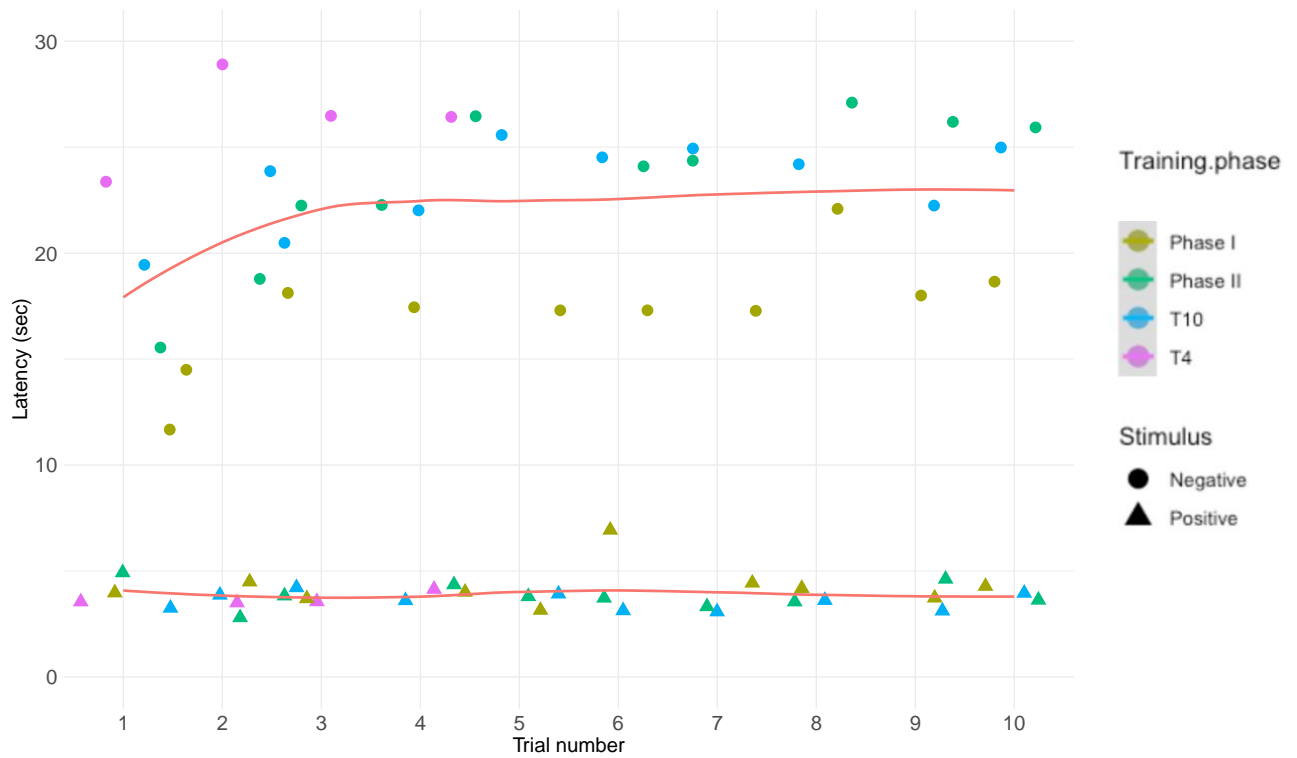
Figure 12: **Average approach latencies for all the birds, for the negative and the positive cue, during all training phases, from trial 1 to trial 10.** In red: linear regression line using the function "LOOEM" from ggplot package, for the average of latency for the negative cue and the positive cue.

Session	Estimate	Est.Error	Q.2.5	Q.97.5	Rhat
Intercept	3.96	0.21	3.57	4.38	1.00
ses2	0.04	0.41	-0.77	0.87	1.00
ses3	-0.06	0.27	-0.60	0.49	1.00
ses4	0.16	0.31	-0.43	0.77	1.00
ses5	-0.08	0.33	-0.72	0.57	1.00
ses6	-0.09	0.29	-0.68	0.47	1.00
ses7	-0.00	0.32	-0.63	0.64	1.00
ses8	-0.06	0.29	-0.64	0.52	1.00
ses9	0.14	0.28	-0.40	0.72	1.00
ses10	0.05	0.30	-0.53	0.65	1.00
ses11	-0.10	0.31	-0.70	0.51	1.00

Trial	Estimate	Est.Error	Q.2.5	Q.97.5	Rhat
Intercept	3.94	0.20	3.57	4.35	1.00
trial2	0.32	0.31	-0.29	0.92	1.00
trial3	-0.10	0.31	-0.70	0.51	1.00
trial4	0.27	0.32	-0.32	0.90	1.00
trial5	-0.39	0.30	-0.99	0.22	1.00
trial6	0.37	0.34	-0.28	1.05	1.00
trial7	0.06	0.32	-0.54	0.68	1.00
trial8	0.15	0.33	-0.51	0.81	1.00
trial9	-0.13	0.33	-0.75	0.54	1.00
trial10	-0.09	0.32	-0.72	0.56	1.00
trial11	-0.21	0.33	-0.85	0.46	1.00
trial12	0.43	0.35	-0.24	1.11	1.00
trial13	0.12	0.34	-0.53	0.79	1.00
trial14	0.10	0.34	-0.55	0.78	1.00
trial15	-0.27	0.33	-0.90	0.39	1.00

Table 4 and Table 5: **Posterior predictions of the latency from model m7 (Session) and m8 (Trial).** Each line corresponds to the intercept difference from the first intercept (Intercept) and the predicted intercept (sesX, trialX). Estimate represents the mean of the factor, Est.Error is the standard deviation from the predicted posterior and Q2.5 and Q.97.5 are the two-sided 95% credible intervals.

DISCUSSION

Our study's primary goal was to link a hypothesised laughter-like vocalisation in kea to a change in emotion that could translate into a measurable cognitive effect. To measure this potential effect on an emotional state, we used a visual Judgment Bias Test, as recommended by several authors (e.g. Gygax, 2014; Roelofs et al., 2016).

Given our results, we cannot conclude that the warble call causes an improvement in kea, given that approach latencies for ambiguous cues were not significantly different in the trill and warble conditions. This is despite the fact that kea correctly learned to distinguish between positive and negative cues and responded significantly different to them. The failure to elicit a significant change in kea emotional state also cannot be ascribed to methodological issues such as subjects' misunderstanding of the ambiguous cue, given that training results showed that kea have a robust capacity to learn and that their learned approach latencies are consistent over time.

As outlined in the introduction, our study aimed to establish five expectations as follow. Over the course of our study, we could validate four of them: (i) kea learnt to approach the positive cue in less than five seconds, (ii) they avoided the negative cue, (iii) they showed a monotonic response in regard to new introduced ambiguous cues and (iv) their behavioural response stayed consistent across the time.

Despite the fact that warble calls did not significantly affect kea emotion in a measurable way, or at least not in a way captured by the JBT, our study nonetheless shows the appropriateness of the JBT and colour cues for studying emotional variations in birds. Our results show that kea quickly learnt to approach the flap door to get a reward in positive but not negative trials, and our randomised testing sessions greatly reduced the risk of subjects learning that ambiguous cues were not rewarded, through the randomisation of testing days and the addition of two types of training days (T10 and T4) in addition to the pre-test "warm-up" trials that were always experienced before the ambiguous cue.

These two types of behaviours (approach or no approach to the flap) can be seen as a flaw in terms of JBT design. Indeed, "Active Choice" studies are preferred because they incorporate a difference between high reward and low reward. Ambiguous are then introduced as middle rewarded. This design leads to a more robust cognitive measurement because of the same behaviour in all trials and the measure of a difference in latency. However, with this form of discrimination, training is often more difficult, and the difference in latencies can be smaller, especially for complex test set-ups as we had with kea. Furthermore, the only study to date that uses active choice task on birds behaviour was run by Matheson et al. (2008) on European starlings, using delayed reward. All other

studies were done with Go/No-Go designs except Hernandez et al., (2015) that designed a spatial JBT on hens but did not find an effect of environment stress.

Confronted with an ambiguous cue, kea showed hesitation to approach or not approach the flap door. Our results show that kea were more hesitant for the neutral ambiguous than near-positive or near-negative cues. Figure 12 leans on the same behavioural response of the kea in front of the ambiguous cue, as ambiguous latencies were comprised in a relevant range between positive and negative cues. That step is required to validate our design study as it ensures that kea perceived ambiguous cues in the same manner as the extreme cues and not as a novelty cue that could increase their emotional state. For the purpose of preventing any order effect and any colour-association effect, we pseudo-randomised our trials and sessions order and ran models that comprise these factors. As expected, we did not find influence of these two parameters that could explain our data.

Nevertheless, the lack of a clear effect of the warble call could potentially be ascribed to one of five factors.

First, despite the strong consistencies in approach latencies within training sessions, kea showed clear inter- and intra-variability in their approach latencies. The causes for these behavioural differences are difficult to define because of kea's complex behaviour, which may vary as a function of daily conditions or events. Likewise, in situ punctual events (e.g. one keeper enters the enclosure) could also be perceive and modify kea attention to the experiment. During trials, each bird is alone on their platform, but other individuals are still visible and obviously capable of vocalising. One way to control for this in (Wein et al., 2018) consists of isolating the bird into an experimental compartment, but this option could introduce intrinsic changes, cause a large increase in stress or negative mood due to social isolation (therefore invalidating the experimental manipulation), and would require extensive pre-training and habituation (Roelofs et al., 2016). This may have resulted in a low signal-to-noise ratio in our dataset, which may have potentially masked a small effect, deeming it undetectable.

Second, it is also possible that Taz's call was less effective than we had anticipated. Kea have a non-linear hierarchy that can be estimated, for example by looking at antagonistic interactions, displacements and physical attacks (Walker C, Demler C, Nelson X., 2018, Appendix). Even if Taz was the highest-ranking individual in previous study, changes to the social group since then may have significantly affected the group's hierarchy and therefore call responsiveness. For example, a juvenile kea born in 2019 was introduced to the group since that study and may have caused considerable changes to the groups' social dynamics and hierarchy. Moreover, the establishment of the hierarchy

was done using antagonistic interactions only. Yet, some authors advise researchers to differentiate rank in hierarchy and power (e.g. De Waal, 1982). Indeed, in fission-fusion groups like kea (Diamond, J., Bond, A.B., 1999), it is important to notice that even if an individual is clearly dominant to another, they won't necessarily win every antagonistic interaction. It is also unclear from existing literature whether the more dominant individuals' call might have the greatest emotional effect on the group. It is certainly possible that, as with vervet monkeys (Seyfarth et al., 1980) for example, certain kea act as sentinels or more reliably produce alarm (trill) or positive (warble) calls (Bond, 2005; Schwing et al., 2012). If that is the case, then these individuals might be attended to more closely by the group than even the more dominant individuals. Further research in wild populations is required to establish how social position or division of labour might affect kea's reactions to certain individual's calls.

The third factor concerns missing data. During the testing phase, we set up a criterion of 4/4, to ensure that the bird was equally attentive in all test trials and that it was using his memory to its best ability (Figure 5, Figure 10). While we designed a test that kept both these cognitive mechanisms stable, in practice we found that kea did not always perform to such a high level. We designed the training criterion (17/20) because birds had a tendency to not inhibit their approach behaviour for the first negative test in a block. This was the case for most birds, but there were exceptions. Blofeld, for example, never failed at the first negative trial in any pre-testing block. As the main problem in JBT studies is to prevent the extinction of a response to the ambiguous cue, we choose to pseudo-randomize trial order even this might cause errors during testing.

Fourth, the call transmission method or the warble call itself could also have no effect on individual behaviour. During the preliminary study (Bastos and Taylor, unpublished data), experimenters set up the speaker behind the platform, very close the kea. Kea reaction was stronger than predicted as most of the birds flew off the platform. Having a fixed position (near Taz's platform) for the speaker was a more biologically appropriate use of his vocalisations, but both the call intensity and played time (1 second call with a 2 second gap for 15 seconds, 5 calls total) may not have been optimal in triggering a change in kea's emotional state. Moreover, the warble could have no effect on kea emotional state, despite this type of vocalisation has been showed by several authors to affect external behaviour by triggering play (Panksepp and Burgdorf, 2003; Schwing et al., 2017; Vettin and Todt, 2005). As the alarm trill call was used as control, we could also see an opposite effect of this call, which might trigger pessimistic internal states. Warble calls specifically could also distract the kea and reduce their attention, but as approach latencies to ambiguous cues were approximately the same for both trill and warble calls, this hypothesis is unlikely to be true.

Finally, Bond and Diamond, (2005) illustrated regional and ontogenetic variation in the "screech" contact calls of kea. The captive population living at the Willowbank reserve could have

independently developed a distinct function for the warble call in comparison to their wild counterparts. Studies on kea calls have thus far not fully examined the differences in calls and their meanings across populations, and research on captive populations is lacking.

Over the course of the design and implementation of this study, it became clear to us that the JBT method requires a real compromise between learning and the number of ambiguous trials (Raoult and Gygax, 2019).

During a test session, we performed 4 trials before to introduce an ambiguous cue. With 3 types of ambiguous for 8 birds for one session, we decided to test 2 birds in parallel using 2 platforms and 2 experimenters per platform. This improvement allowed us to divide by 2 the total number of calls played, but as the call were played into the whole aviary, a perfect synchronization was mandatory to run both ambiguous tests accordingly. This synchronization could necessitate one bird to wait few seconds (<30seconds) before to get exposed to an ambiguous cue, and could introduce a potential attention bias even though kea were not exposed yet to the ambiguous cue and did not show eagerness behaviour.

One other compromise in the design is the notion of treatment. Indeed, some studies used a poor versus enriched environment (e.g. Brilot et al., 2010; Harding et al., 2004) but the introduction of stressful event or stressor can lead to other emotional state modification. For example, Hernandez et al.,(2015) relate that isolation before an ambiguous state can lead to an increase of optimism, and Roelofs et al., (2016) argue that stress can enhance attention to a cue, hence increasing the learning performance for the ambiguous trials. Hence, we decided to introduce the negative cue with no reward to exclude these difficulties as well as having in mind the welfare requirement of kea.

Here we showed that birds did not learn that ambiguous was not rewarded over up to eighteen trials. To date, this is the first time a judgment bias test has been set up on Psittacidae. Training was fast in comparison with other species, which might reflect the complex cognitive capabilities of kea (Bastos and Taylor, 2020; Diamond, J., Bond, A.B., 1999; Wein et al., 2018, 2019).

The JBT method may serve a better function as a complement to other techniques to assess animals' internal emotional states, as it represents an ambitious gap to fill in the understanding of animal welfare.

The JBT method can also be linked with animal personality with replication and long-term measurements (Adriaense et al., 2019). The accumulation of data using JBT and classic personality tests could help building a full picture of one individual's emotional state, and to date only Lalot et

al. (2017) included a personality analysis with a JBT study for birds. Social influence, hierarchy and emotional propagation are another driving that could help to understand the role of communication.

Our pilot study establishes a methodology to assess emotional states in parrots and these tests could be applied more widely as robust measures of individual and group-level welfare and husbandry practices for a wide range of species.

Adriaense JEC, Martin JS, Schiestl M, et al (2019) Negative emotional contagion and cognitive bias in common ravens (*Corvus corax*). Proc Natl Acad Sci 116:11547–11552. https://doi.org/10.1073/pnas.1817066116

Barton RA, Dunbar RIM (1997) Evolution of the social brain. In: Whiten A, Byrne RW (eds) Machiavellian Intelligence II, 1st edn. Cambridge University Press, pp 240–263

Bastos APM, Taylor AH (2020) Kea show three signatures of domain-general statistical inference. Nat Commun 11:828. https://doi.org/10.1038/s41467-020-14695-1

Bethell EJ (2015) A "How-To" Guide for Designing Judgment Bias Studies to Assess Captive Animal Welfare. J Appl Anim Welf Sci 18:S18–S42. https://doi.org/10.1080/10888705.2015.1075833

Bond AB, Diamond J (2005) Geographic and ontogenetic variation in the contact calls of the kea (Nestor notabilis). Behaviour 142:1–20. https://doi.org/10.1163/1568539053627721

Briefer EF (2018) Vocal contagion of emotions in non-human animals. Proc R Soc B Biol Sci 285:20172783. https://doi.org/10.1098/rspb.2017.2783

Brilot BO, Asher L, Bateson M (2010) Stereotyping starlings are more 'pessimistic.' Anim Cogn 13:721–731. https://doi.org/10.1007/s10071-010-0323-z

Bürkner P-C (2017) **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. J Stat Softw 80:. https://doi.org/10.18637/jss.v080.i01

Byrne RW, Whiten A (1988) Machiavellian intelligence. Social expertise and the evolution of the intellect in monkeys, apes, and humans. Clarendon Press Oxf

De Waal (1982) Chimpanzee Politics – Power and Sex Among Apes. Johns Hopkins Univ Press

Diamond, J., Bond, A.B., (1999) Kea, Bird of Paradox. Univ Calif Press Berkley

Emery NJ (2006) Cognitive ornithology: the evolution of avian intelligence. Philos Trans R Soc B Biol Sci 361:23–43. https://doi.org/10.1098/rstb.2005.1736

Ennaceur A (2014) Tests of unconditioned anxiety — Pitfalls and disappointments. Physiol Behav 135:55–71. https://doi.org/10.1016/j.physbeh.2014.05.032

Gygax L (2014) The A to Z of statistics for testing cognitive judgement bias. Anim Behav 95:59–69. https://doi.org/10.1016/j.anbehav.2014.06.013

Harding EJ, Paul ES, Mendl M (2004) Cognitive bias and affective state. Nature 427:312–312. https://doi.org/10.1038/427312a

Hatfield E, Cacioppo J, Rapson R (1993) Emotional Contagion. Curr Dir Psychol Sci Vol. 2, No. 3 (Jun., 1993), pp. 96–99:

Hernandez CE, Hinch G, Lea J, et al (2015) Acute stress enhances sensitivity to a highly attractive food reward without affecting judgement bias in laying hens. Appl Anim Behav Sci 163:135–143. https://doi.org/10.1016/j.applanim.2014.12.002

Kruschke JK, Liddell TM (2018) The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev 25:178–206. https://doi.org/10.3758/s13423-016-1221-4

Lalot M, Ung D, Péron F, et al (2017) You know what? I'm happy. Cognitive bias is not related to personality but is induced by pair-housing in canaries (Serinus canaria). Behav Processes 134:70–77. https://doi.org/10.1016/j.beproc.2016.09.012

Marler PR (1982) Avian and primate communication: The problem of natural categories. Neurosci Biobehav Rev 6:87–94. https://doi.org/10.1016/0149-7634(82)90010-0

Matheson SM, Asher L, Bateson M (2008) Larger, enriched cages are associated with 'optimistic' response biases in captive European starlings (Sturnus vulgaris). Appl Anim Behav Sci 109:374–383. https://doi.org/10.1016/j.applanim.2007.03.007

McCoy DE, Schiestl M, Neilands P, et al (2019) New Caledonian Crows Behave Optimistically after Using Tools. Curr Biol 29:2737-2742.e3. https://doi.org/10.1016/j.cub.2019.06.080

Mendl M, Burman OHP, Parker RMA, Paul ES (2009) Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. Appl Anim Behav Sci 21

Mendl M, Burman OHP, Paul ES (2010) An integrative and functional framework for the study of animal emotion and mood. Proc R Soc B Biol Sci 277:2895–2904. https://doi.org/10.1098/rspb.2010.0303

Mendl M, Paul ES (2020) Animal affect and decision-making. Neurosci Biobehav Rev 112:144–163. https://doi.org/10.1016/j.neubiorev.2020.01.025

Neave HW, Daros RR, Costa JHC, et al (2013) Pain and Pessimism: Dairy Calves Exhibit Negative Judgement Bias following Hot-Iron Disbudding. PLoS ONE 8:e80556. https://doi.org/10.1371/journal.pone.0080556

Panksepp (1998) Affective Neuroscience: The Foundations of Human and Animal Emotions. New York: Oxford University Press.:

Panksepp J, Burgdorf J (2003) "Laughing" rats and the evolutionary antecedents of human joy? Physiol Behav 79:533–547. https://doi.org/10.1016/S0031-9384(03)00159-8

Paul ES, Sher S, Tamietto M, et al (2020) Towards a comparative science of emotion: Affect and consciousness in humans and animals. Neurosci Biobehav Rev 108:749–770. https://doi.org/10.1016/j.neubiorev.2019.11.014

Perry CJ, Baciadonna L (2017) Studying emotion in invertebrates: what has been done, what can be measured and what they can provide. J Exp Biol 220:3856–3868. https://doi.org/10.1242/jeb.151308

Pincus T, Pearce S, Perrott A (1996) Pain patients' bias in the interpretation of ambiguous homophones. Br J Med Psychol 69:259–266. https://doi.org/10.1111/j.2044-8341.1996.tb01868.x

Raoult CMC, Gygax L (2019) Mood induction alters attention toward negative-positive stimulus pairs in sheep. Sci Rep 9:7759. https://doi.org/10.1038/s41598-019-44330-z

Roelofs S, Boleij H, Nordquist RE, van der Staay FJ (2016) Making Decisions under Ambiguity: Judgment Bias Tasks for Assessing Emotional State in Animals. Front Behav Neurosci 10:. https://doi.org/10.3389/fnbeh.2016.00119

Rozin P, Royzman EB (2001) Negativity Bias, Negativity Dominance, and Contagion. Personal Soc Psychol Rev 5:296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178. https://doi.org/10.1037/h0077714

Salmeto AL, Hymel KA, Carpenter EC, et al (2011) Cognitive bias in the chick anxiety–depression model. Brain Res 1373:124–130. https://doi.org/10.1016/j.brainres.2010.12.007

Schwing R, Nelson XJ, Wein A, Parsons S (2017) Positive emotional contagion in a New Zealand parrot. Curr Biol 27:R213–R214. https://doi.org/10.1016/j.cub.2017.02.020

Schwing R, Parsons S, Nelson XJ (2012) Vocal repertoire of the New Zealand kea parrot *Nestor notabilis*. Curr Zool 58:727–740. https://doi.org/10.1093/czoolo/58.5.727

Seehuus B, Mendl M, Keeling LJ, Blokhuis H (2013) Disrupting motivational sequences in chicks: Are there affective consequences? Appl Anim Behav Sci 148:85–92. https://doi.org/10.1016/j.applanim.2013.07.008

Seyfarth RM, Cheney DL, Marler P (1980) Vervet monkey alarm calls: Semantic communication in a free-ranging primate. Anim Behav 28:1070–1094. https://doi.org/10.1016/S0003-3472(80)80097-2

Špinka M (2012) Social dimension of emotions and its implication for animal welfare. Appl Anim Behav Sci 12

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27:1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vettin J, Todt D (2005) Human laughter, social play, and play vocalizations of non-human primates: an evolutionary approach. 24

Wein A, Schwing R, Hausberger M, et al (2018) Vocal conditioning in kea parrots (Nestor notabilis). J Comp Psychol 132:97–105. https://doi.org/10.1037/com0000098

Wein A, Schwing R, Huber L (2019) Kea *Nestor notabilis* mothers produce nest-specific calls with low amplitude and high entropy. Ibis ibi.12753. https://doi.org/10.1111/ibi.12753

Appendix 1: Walker C, Demler C, Nelson X. 2018: **Dominance hierarchy and its effect on levels of interactions**

Appendix 2: **Pictures of the apparatus**
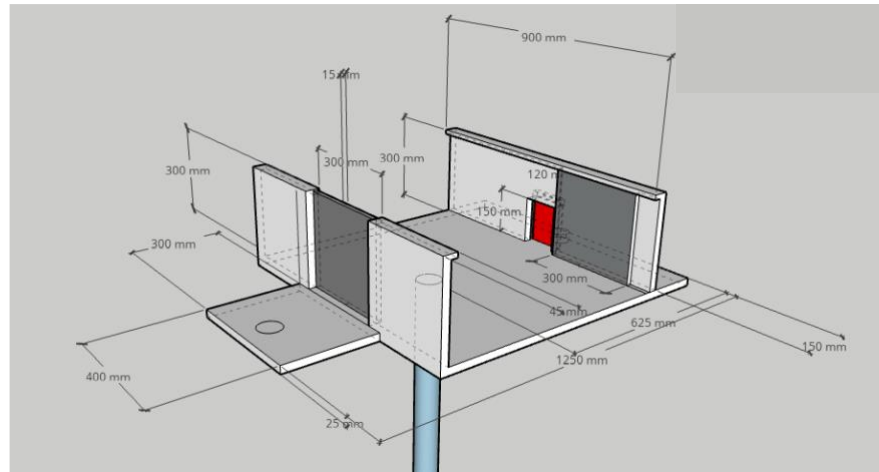
Appendix 3: **Output of the R code of the description of the model m4**

```
summary(m4)
##  Family: weibull
##   Links: mu = log; shape = identity
## Formula: lat | cens(censored) ~ 0 + cond:cue + (0 + cond:cue | ID)
##    Data: d (Number of observations: 450)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~ID (Number of levels: 8)
##                                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(cond0:cue0)                     0.18      0.08     0.07     0.37 1.01 1191 1758
## sd(cond1:cue0)                     0.04      0.03     0.00     0.13 1.00 2138 2106
## sd(cond0:cue0.25)                  0.13      0.10     0.01     0.39 1.00 1798 2036
## sd(cond1:cue0.25)                  0.37      0.16     0.13     0.74 1.00 2127 2089
## sd(cond0:cue0.5)                   1.13      0.38     0.64     2.09 1.00 2597 2788
## sd(cond1:cue0.5)                   1.23      0.44     0.68     2.37 1.00 2839 2648
## sd(cond0:cue0.75)                  0.34      0.23     0.04     0.92 1.00 1833 1954
## sd(cond1:cue0.75)                  0.27      0.21     0.01     0.80 1.00 2243 2234
## sd(cond0:cue1)                     0.40      0.41     0.01     1.37 1.00 5076 2528
## sd(cond1:cue1)                     0.37      0.30     0.02     1.17 1.00 2211 2486
## cor(cond0:cue0,cond1:cue0)         0.04      0.30    -0.55     0.59 1.00 6920 2744
## cor(cond0:cue0,cond0:cue0.25)     -0.02      0.30    -0.59     0.56 1.00 6583 2613
## cor(cond1:cue0,cond0:cue0.25)     -0.03      0.30    -0.60     0.55 1.00 4374 3268
## cor(cond0:cue0,cond1:cue0.25)      0.02      0.27    -0.49     0.55 1.00 3928 3322
## cor(cond1:cue0,cond1:cue0.25)     -0.00      0.30    -0.57     0.57 1.00 2326 3064
## cor(cond0:cue0.25,cond1:cue0.25)  -0.03      0.30    -0.58     0.54 1.00 3080 3334
## cor(cond0:cue0,cond0:cue0.5)       0.15      0.24    -0.33     0.60 1.00 3652 3089
## cor(cond1:cue0,cond0:cue0.5)       0.04      0.29    -0.52     0.58 1.00 2171 2629
## cor(cond0:cue0.25,cond0:cue0.5)   -0.10      0.28    -0.64     0.46 1.00 2277 2656
## cor(cond1:cue0.25,cond0:cue0.5)    0.16      0.25    -0.35     0.63 1.00 3803 3151
## cor(cond0:cue0,cond1:cue0.5)       0.05      0.25    -0.44     0.51 1.00 3480 3116
## cor(cond1:cue0,cond1:cue0.5)       0.07      0.29    -0.51     0.59 1.00 2128 2661
## cor(cond0:cue0.25,cond1:cue0.5)   -0.13      0.28    -0.65     0.44 1.00 1810 2406
## cor(cond1:cue0.25,cond1:cue0.5)    0.04      0.25    -0.44     0.52 1.00 3814 3134
## cor(cond0:cue0.5,cond1:cue0.5)     0.41      0.23    -0.08     0.79 1.00 3405 3254
## cor(cond0:cue0,cond0:cue0.75)      0.00      0.28    -0.53     0.54 1.00 5463 3318
## cor(cond1:cue0,cond0:cue0.75)      0.01      0.30    -0.55     0.59 1.00 3477 3105
## cor(cond0:cue0.25,cond0:cue0.75)  -0.07      0.30    -0.61     0.52 1.00 3337 3038
## cor(cond1:cue0.25,cond0:cue0.75)  -0.00      0.29    -0.55     0.57 1.00 4272 3033
## cor(cond0:cue0.5,cond0:cue0.75)    0.22      0.27    -0.34     0.69 1.00 3701 2986
## cor(cond1:cue0.5,cond0:cue0.75)    0.24      0.27    -0.31     0.71 1.00 3280 3447
## cor(cond0:cue0,cond1:cue0.75)      0.04      0.29    -0.52     0.59 1.00 6067 2840
## cor(cond1:cue0,cond1:cue0.75)      0.02      0.30    -0.56     0.58 1.00 4005 2832
## cor(cond0:cue0.25,cond1:cue0.75)  -0.07      0.30    -0.62     0.52 1.00 4048 3143
## cor(cond1:cue0.25,cond1:cue0.75)   0.11      0.29    -0.45     0.63 1.00 4336 3300
## cor(cond0:cue0.5,cond1:cue0.75)    0.21      0.28    -0.38     0.70 1.00 3552 3090
## cor(cond1:cue0.5,cond1:cue0.75)    0.19      0.28    -0.40     0.68 1.00 3509 2991
## cor(cond0:cue0.75,cond1:cue0.75)   0.11      0.30    -0.47     0.65 1.00 3002 2924
## cor(cond0:cue0,cond0:cue1)        -0.00      0.30    -0.58     0.57 1.00 8320 2536
```

```
## cor(cond1:cue0,cond0:cue1)         -0.01        0.31      -0.60       0.59 1.00 5833 2840
## cor(cond0:cue0.25,cond0:cue1)      -0.00        0.31      -0.58       0.58 1.00 3988 2879
## cor(cond1:cue0.25,cond0:cue1)      -0.00        0.30      -0.58       0.58 1.00 5018 2918
## cor(cond0:cue0.5,cond0:cue1)        0.01        0.31      -0.59       0.59 1.00 4785 2758
## cor(cond1:cue0.5,cond0:cue1)        0.01        0.30      -0.57       0.59 1.00 4215 3415
## cor(cond0:cue0.75,cond0:cue1)       0.00        0.30      -0.57       0.60 1.00 3537 3605
## cor(cond1:cue0.75,cond0:cue1)       0.01        0.30      -0.57       0.58 1.00 2987 3193
## cor(cond0:cue0,cond1:cue1)          0.14        0.30      -0.45       0.67 1.00 5794 2764
## cor(cond1:cue0,cond1:cue1)          0.01        0.30      -0.59       0.59 1.00 5228 3579
## cor(cond0:cue0.25,cond1:cue1)      -0.03        0.30      -0.60       0.56 1.00 3673 3180
## cor(cond1:cue0.25,cond1:cue1)       0.06        0.29      -0.48       0.61 1.00 4367 3673
## cor(cond0:cue0.5,cond1:cue1)        0.18        0.28      -0.39       0.69 1.00 4751 2996
## cor(cond1:cue0.5,cond1:cue1)        0.13        0.28      -0.44       0.63 1.00 3926 3372
## cor(cond0:cue0.75,cond1:cue1)       0.08        0.29      -0.49       0.62 1.00 3314 3338
## cor(cond1:cue0.75,cond1:cue1)       0.07        0.30      -0.51       0.63 1.00 2877 3413
## cor(cond0:cue1,cond1:cue1)         -0.01        0.31      -0.61       0.57 1.00 2555 2959
##
## Population-Level Effects:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## cond0:cue0        1.33      0.08     1.18     1.49 1.00     1699     1907
## cond1:cue0        1.25      0.03     1.18     1.31 1.00     4605     3179
## cond0:cue0.25     1.46      0.09     1.29     1.66 1.00     4210     2910
## cond1:cue0.25     1.54      0.16     1.23     1.86 1.00     2749     2436
## cond0:cue0.5      2.07      0.32     1.46     2.74 1.00     3166     2250
## cond1:cue0.5      2.11      0.35     1.46     2.84 1.00     3696     2714
## cond0:cue0.75     3.33      0.19     2.93     3.67 1.00     2342     1411
## cond1:cue0.75     3.63      0.18     3.32     4.06 1.00     3085     2438
## cond0:cue1        6.11      2.05     4.31    11.58 1.00     1375     1302
## cond1:cue1        4.31      0.31     3.87     5.10 1.00     3205     1971
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     4.03      0.24     3.58     4.52 1.00     4345     3189
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The output describes several parameters. On the top, general information is given such as the family, formula and number of iterations. Next, group-level effects are showed separately for each grouping factor in terms of standard deviation (sd) as well as correlations between group-level effects. Estimate is the mean of the parameter, Est. error is the standard deviation, l-95% CI and u-95% CI are the two-sided 95% credible intervals of the posterior distribution, Rhat is an indicator of the fit of the MCMC algorithm, while Bulk_ESS and Tail_ESS are two indicators of the number of samples (ESS: effective sample size) that contain an informative amount of data. Bulk_ESS should be superior to 100 times the number of chains. On the bottom, the output shows population-level effects (regression coefficients) and specific parameters. Here shape is showed as the Weibull shape (see Methods for formula).

Appendix 4: **Output of the R code of the MCMC process for model m4**
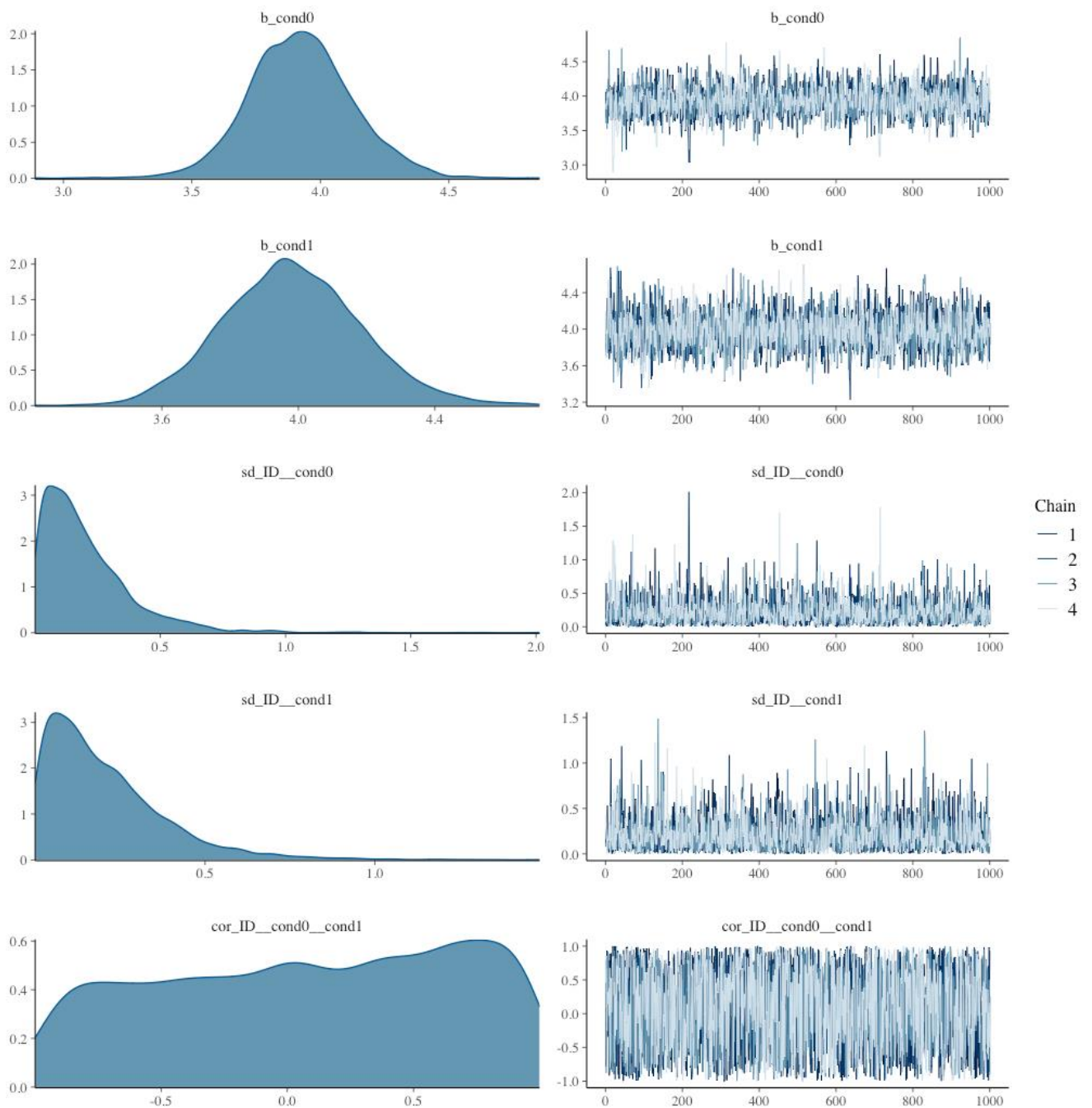
The left part show density plots taken from the posterior distribution for the intercept of the condition trill (b_cond0) and warble (b_cond1), and the standard deviation for these two conditions (sd_ID_cond0 and sd_ID_cond0). Finally, the correlation coefficient is showed (cor_ID_cond0_cond1). The right part is the trace plot and shows the trace plbehaviour of the four simulations (chains) where x-axis corresponds to the number of iterations (1000 per chain) and the y-axis represents the value of the parameter. Chains that constructed the model should be "mixed" i.e. exploring random values between a fixed mean value.