UNIVERSITY OF CAMBRIDGE

Public Engagement Technology for Bioacoustic Citizen Science

Isak Herman

Darwin College

This dissertation is submitted for the Degree of Doctor of Philosophy, May, 2020

Abstract

Public Engagement Technology for Bioacoustic Citizen Science *Isak Herman*

Inexpensive mobile devices offer new capabilities for non-specialist use in the field for the purpose of conservation. This thesis explores the potential for such devices to be used by citizen scientists interacting with bioacoustic data such as birdsong. This thesis describes design research and field evaluation, in collaboration with conservationists and educators, and technological artefacts implemented as mobile applications for interactive educational gaming and creative composition.

This thesis considers, from a participant-centric collaborative design approach, conservationists' demand for interactive artefacts to motivate engagement in citizen science through gameful and playful interactions. Drawing on theories of motivation, frequently applied to the study of human computer interaction (HCI), and on approaches to designing for motivational engagement, this thesis introduces a novel pair of frameworks for the analysis of technological artefacts and for assessing participant engagement with bioacoustic citizen science from both game interaction design and citizen science project participation perspectives. This thesis reviews current theories of playful and gameful interaction developed for collaborative learning, data analysis, and ground-truth development, describes a process for design and analysis of motivational mobile games and toys, and explores the affordances of various game elements and mechanics for engaging participation in bioacoustic citizen science.

This thesis proposes research into progressions for scaffolding engagement with citizen science projects where participants interact with data collection and analysis artefacts. The research process includes the development of multiple designs, analyses of which explore the efficacy of game interactions to motivate engagement through interaction progressions, given proposed analysis frameworks. This thesis presents analysed results of experiments examining the usability of, and data-quality from, several prototypes and software artefacts, in both laboratory conditions and the field. This thesis culminates with an assessment of the efficacy of proposed design analysis frameworks, an analysis of designed artefacts, and a discussion of how these designs increase intrinsic and extrinsic motivation for participant engagement and affect resultant bioacoustic citizen science data quantity and quality.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Isak Herman
May, 2020

Acknowledgements

While all errors herein are necessarily mine, this work would not have happened without the guidance, assistance, and support of many people. I cannot thank enough my supervisor, Professor Alan Blackwell, who has nurtured, guided, coaxed, and cajoled me through the development of this project; without him, this research would not have seen the light of day. Many thanks also to my second supervisor, Dr. Chris Sandbrook, for guiding my exploration of the field of conservation, with which I had only avocational prior familiarity. For discussions and advice on how to approach research in schools, invaluable advice was offered by James de Winter, Dr. Alistair Stead, and Dr. Sam Aaron. Thanks as well to Dr. John Fanshawe for his insights into the needs of the British Trust for Ornithology (BTO) as well as the broader conservation community.

Thanks go as well to my collaborators in the field. In particular to Peter Cowdrey from Planet Birdsong (PBS) for his contributions to field deployment and introductions to a community of avocational ornithologists. Thanks to Paul Burgess and Sarah Kettlewell with the Wild Watch project, for organizing events throughout the Nidderdale Area of Outstanding Natural Beauty (AONB) and for providing me with the means to engage with populations I would not otherwise know, specifically, the students and teachers of Kirkby Malzear, Birstwith, Bishop Thornton, Summerbridge, Darley, Kettlesing, Lofthouse, St. Cuthberts, Burnt Yates, Grewelthorpe, and Glasshouses primary schools. Thanks due to Ben Elliott for his contribution of an initial live visualization prototype. Further thanks are due to Chris Ford at Royal Society for the Protection of Birds (RSPB) Minsmere for his role in organizing field work opportunities, advice on application deployment, and feedback on the potential for my designs within the context of the broader field of avian citizen science.

Finally, thanks to my family and friends, while the latter are too numerous to count, you know who you are, and thank you. Utmost thanks to my mother, Elaine Herman, who, whilst claiming to understand little of this topic, provided invaluable editing and unparalleled support. Thanks to Joana Guedes, who motivated me to focus on writing when the time came, and to Jason Allen, for his conversations and loan of a puppy when needed most. And finally, thanks to Stefania Radopoulou for the emotional inspiration I needed to complete this thesis.

Contents

Glossary of Abbreviations and Acronyms

ACR automatic call recognition.

AHQI acoustic habitat quality index.

ANOVA analysis of variance.

AONB Area of Outstanding Natural Beauty.

API application programming interface.

AR acoustic richness.

ARG alternate reality game.

ARU autonomous recording unit.

ASI automatic species identification.

ASR automated speech recognition.

BGBW Big Garden Birdwatch.

BLED band-limited energy detector.

BNT basic needs theory.

BTO British Trust for Ornithology.

CASA computational auditory scene analysis.

CBC Christmas Bird Count.

CET cognitive evaluation theory.

CLO Cornell Lab of Ornithology.

COT causality orientation theory.

DCT discrete cosine transform.

DFT discrete Fourier transform.

DT decision tree.

DTW dynamic time warping.

EML ecological metadata language.

FFT fast Fourier transform.

FT Fourier transform.

GBIF Global Biodiversity Information Facility.

GBW Garden BirdWatch.

GIS geographical information system.

GOMS Goals, Operators, Methods, and Selection rules.

GWAP game with a purpose.

GWCT Game and Wildlife Conservation Trust.

HCI human computer interaction.

HIT human intelligence task.

HitL human-in-the-loop.

HLF Heritage Lottery Fund.

HMM hidden Markov model.

HSM habitat suitability modelling.

MAP mean average precision.

MFCC Mel-frequency cepstral coefficient.

MLP multi-layer perceptron.

MP matching pursuit.

NAS National Audubon Society.

NGO non-governmental organisation.

ODE occupation-detection-expertise.

OIT organismic integration theory.

PBS Planet Birdsong.

PCA principal component analysis.

PPSR public participation in scientific research.

QDA quadratic discriminant analysis.

RAP rapid assessment program.

RBA rapid biodiversity assessment.

RBF-SVM radial basis function support vector machine.

RDB Rwanda Development Board.

ROI region of interest.

RSPB Royal Society for the Protection of Birds.

RTU recognisable taxonomic unit.

SAI stabilized auditory image.

SDT self-determination theory.

SNR signal-to-noise ratio.

SRNFN singleton-type recurrent neural fuzzy network.

STBF spectro-temporal box filter.

STFT short-time Fourier transform.

SVM support vector machine.

TAM technological acceptance model.

TLM-GOMS Touch-Level Model Goals, Operators, Methods, and Selection rules.

TRA theory of reasoned action.

TUI tangible user interface.

UI user interface.

UK United Kingdom.

VGI volunteered geographical information.

WMFCC wavelet Mel-frequency cepstral coefficient.

CHAPTER 1

Introduction

1.1 Research motivation

Research into the changing distribution and prevalence of avian populations is important to conservation initiatives, however, data collection by scientists is expensive and limited when they must collect data in the field. This thesis investigates tools and processes that can be used to collect avian presence data by training members of the public to contribute by identifying birds from bioacoustic data — the sounds of songs and calls. There is a significant tradition of avocational[1] interest in birds and their songs. Avocational contributions through *citizen science* initiatives offer a valuable mechanism to significantly increase the quantity of bioacoustic survey data, but introduce inevitable trade-offs in data accuracy, consistency, and coverage. Citizen science supports increased public engagement as avocational researchers become motivated to contribute to conservation initiatives, funding, and public discourse on biodiversity.

Data are shaped by the motivations of various stakeholders, including residents where data are collected, participants in data collection, scientists whose models depend on collected data, and policymakers. My research explores the collaborative design of tools to enhance learning and motivate engagement with avian conservation through citizen science. I explore mechanisms for motivating avocational data collection through play and discuss the need to validate data prior to incorporation into datasets underpinning biodiversity models. I explore questions of data validity, whether avocational collectors' motivations lead only to creation of high data volume or actual high quality data, and data ownership. I propose, in §2.7.2, a novel framework for characterising citizen scientists' motivation to engage with data they collect. My research encompasses experiments examining design of games for quantifying the value of future bioacoustic data collection given participants' performance during play.

[1] In contrast with professional ornithologists avocational researchers lack formal scientific training.

1.2 Bioacoustic monitoring: why birds?

This thesis explores birdsong because, of all bioacoustic signals in nature, these are most familiar to the most people. In general, bioacoustic signals can help to assess biodiversity and identify cryptic (unseen) species. Previous applications of technology to birdsong, have tended to focus on machine listening, including automatic call recognition (ACR) and automatic species identification (ASI), while broader bioacoustics research has proposed population density metrics based on bioacoustic energy within constrained acoustically-surveyable regions. Leveraging citizen science for bioacoustic data-collection, given the popularity of avocational ornithology, is necessary to build the data-sets prerequisite for future success of statistical models

1.2.1 Birding in Britain

In the United Kingdom (UK) some of the most evident biological shifts in recent decades have been amongst avian populations. This is likely due to the greater volume of survey data, when compared to data on other fauna, provided by avocational ornithologists, not necessarily to avian shifts being more significant or prevalent. There is significant participation in avocational bird-watching in the UK, although participation varies depending upon engagement requirements. Collection methods for behavioural, range, and population data currently combine expert research with censuses by motivated citizen scientists; maintaining and increasing the number of amateur participants is necessary.

The scientific community has long benefited from a population of avocational bird-watchers. Numerous non-governmental organisations (NGOs) exist to further public engagement with avian species and provide institutional support for data-collection and dissemination, model creation, and policy formulation. The annual Big Garden Birdwatch (BGBW) of the Royal Society for the Protection of Birds (RSPB), started in 1979, represents a lower bound on survey complexity as it requires a single hour's observation and no transects. It provides an upper bound on citizen scientist data-collection in the UK, with 472,758 participants in 2019[2]. For this survey, identification is made when a target species is perceived visually.

The Garden BirdWatch (GBW), organised by the British Trust for Ornithology (BTO), requires greater participant time commitment, as surveying guidance requests weekly data-collection throughout the year; participants are only instructed to report visual counts. Engagement is limited to participants self-selecting as likely to remain active, as a fee is required to fund management of collected data. Nonetheless, annually 15,000 citizen scientists subscribe[3]; this has proved a workable model for adding temporal continuity

[2] https://ww2.rspb.org.uk/get-involved/activities/birdwatch/results/

[3] However, not all participate. https://www.bto.org/volunteer-surveys/gbw/taking-part

to citizen science surveys. These surveys have contributed to long-term data-collection efforts by the BTO which have yielded "range change and abundance [data] for over 500 bird species that have been recorded in Britain and Ireland during the atlases spanning 1968 to 2011"[13]. As survey guidance protocols do not support acoustic data-collection, a gap exists.

1.2.2 Technology for birding

Advances in source separation and noise-reduction algorithms, which may run on mobile devices, provide tools for both passive monitoring and non-specialist participants' recording of geo-located avian audio. For such approaches — which provide high data volume — to underpin data-driven models, given current automated recognition system limitations, database entries must include human-validated annotated species metadata. I designed interactive approaches to citizen science for data-driven bioacoustic surveying and modelling. My research artefacts support Mason et. al's requirements of amalgamating bioacoustic data and acoustic event annotation and visualisation, enabling interactive analysis[150]. My research centres on identifying methods for increasing engagement and data-quality provided by citizen scientists in avian bioacoustic surveys.

Existent citizen science surveys focus on visual rather than acoustic species identification; trust — already contentious for scientists incorporating such data into biodiversity models — associated with visual identification has ranked higher. In visually-occluded habitats such as woodland, scrub, or, to a lesser degree, wetland, it is more common to hear than to see birds; furthermore birds of similar appearance frequently have distinguishable utterances. My games provide participant training and knowledge validation which increase potential trust in audio data. If games motivate participants to remain involved after collecting data, their implicit intrinsic motivation, and the quality and trust associated with their subsequent work, increases. Leveraging amateur naturalists' intrinsic motivation to collect avian data and children's interest in games, I investigate whether mobile game performance can validate submissions, supporting the incorporation of amateur acoustic survey results into databases underpinning conservation models.

1.3 Bioacoustic data collection: why mobile?

Market penetration of smartphone technology in the UK[4] now exceeds 90% in the under-34 demographic, and 80% population-wide. Recent widespread adoption of such devices brings previously unavailable capabilities for recording and analysing bioacoustic signals in the field. Therefore my research explores the use of mobile technology for enhancing knowledge

[4] https://www.statista.com/statistics/271851/smartphone-owners-in-the-united-kingdom-uk-by-age/

and increasing bioacoustic data-collection in a citizen science project. Software on a mobile device can be used "to select the best subset of privately owned sensors ... to estimate a complex spatial phenomenon"[125] through real-time recording and transmitting of bioacoustic data. This contrasts with traditional avian survey methods, such as the RSPB and BTO surveys previously described, where data are documented and later uploaded to a database. Treating collected information from a data-driven perspective has proven useful when studying global effects and macro-ecology[68]; collecting spatio-temporally diverse data such as avian utterances requires either vast static sensor networks, or mobile citizen sensor networks. My research involves interface designs on familiar mobile devices motivating citizen science project engagement and data-collection by both avocational ornithologists and those previously uninterested in avian bioacoustics.

1.3.1 Technological use among bird-watchers

Bird-watchers fall broadly into two categories: some define their pursuit as bird-watching, while for others bird-watching is ancillary to other outdoor pursuits. The former frequently have invested in the technological accoutrements of bird-watching, including visual and acoustic augmentation and recording devices; the latter are less likely to invest in hobby-specific technology or carry bulky tools whilst outdoors. The majority of avian citizen science contributions come from participants in the former category, yet they represent a minority of potentially involved citizen scientists. Both groups are likely to carry mobile devices which can enhance user engagement with the surrounding environment.

Standard mobile phones now have several capabilities necessary for bioacoustic data-collection and sufficient power for use in the field. Audio recordings require significant device memory, or network bandwidth if data are to be immediately uploaded, both of which are now available. Smartphone geo-location capabilities eliminate the need for citizen scientists to manually annotate location data, removing a common error source. However, participant-centric sensing can increase species disturbance and ecological degradation, while remote locations can become data deserts. As a result, static sensor networks for passive acoustic recording are the conventional technological alternative for tracking location of and development in nests. My goal is to create mobile interactions that enhance engagement and learning without causing ecological disruption.

1.3.2 Learning with mobile technology

Mobile devices can be pre-loaded with digitised versions of field guides[5], the primary historic tool of avocational bird-watchers for learning to identify unfamiliar species. Interactive guidebooks allow users to work through the identification process in ways not possible

[5]See guides for Android and iOS: http://www.birderslibrary.com/features/bird-apps-of-the-world.htm

with conventional books; for instance, a geo-tagged time-stamped database of prior bird observations seeding a front-end field guide on devices aware of location, time, and season can sort likely matches to avoid proposing improbable species. The interactive nature of mobile devices allows for engagement modalities beyond those available within conventional field guides. Existing real-time automated birdsong recognition applications[6] are constrained in geographic scope and number of identifiable species. Mobile devices can also be leveraged as bi-directional teaching tools, not only helping users identify unfamiliar birds and reinforcing prior knowledge, but also allowing users to add to datasets from which other users can learn. My research involves the design of interfaces for familiar mobile devices to augment avian knowledge and motivate citizen science project engagement and data-collection through games.

1.4 Motivating citizen scientists: why games?

Most current avian citizen science projects assume that intrinsic rewards will motivate data-collection. Mobile games potentially provide a familiar interface from which to explore motivating engagement with citizen scientists. My research explores whether engagement and data-quality increase when participation and learning are mediated through play. Engagement metrics can measure project success in terms of retention of prior participants, engagement with current participants, and outreach to potential participants. An issue evident from engagement surveys from prior citizen science projects is the perceived unidirectional data-transfer between those responsible for data-collection and those responsible for model construction and policy formulation (see §3.3.2.3). My research incorporates local stakeholders in data-acquisition processes, building the knowledge-base necessary for both game development and grassroots action.

1.4.1 Measuring engagement & data-quality

Motivating engagement through play is integral to my research into designing an interaction model which encourages citizen scientists to contribute quality data. Conventional metrics for participant activity include data-contribution counts, time spent collecting data, and results of surveys querying participants' perceptions of engagement. Many citizen scientists trained in the field fail to contribute due to a mismatch between their personal goals, and training which does not correspond to their motivation for participation. Conversely, projects with minimal engagement requirements, and simplified data-collection protocols, such as the BGBW, result in data of questionable quality. The current interaction model for user submissions to such databases is passive; there is little motivation beyond the

[6] Examples include https://www.warblr.co.uk and http://www.chirpomatic.com

intrinsic reward of *contributing to science*, and database uploads occur with some delay from the original record.

While bias and error in citizen scientist-generated datasets diminish with clearly designed data-entry protocols, open databases accept participants' data without prior expectation regarding their submission accuracy. Models generally assign low trust values to citizen science data; gamification provides tools for quantifying trust in citizen science data while concurrently providing interactions which may increase motivation and user retention. The data results from my primary fieldwork eventually reached Simon Pickles[7], Director of the North and East Yorkshire Ecological Data Centre, who bemoaned haphazard application of metadata collection protocols across regional projects; my designs have incorporated ecological metadata language (EML)[79] to constrain data-input and ensure its scientific value. Concurrent utilisation of geographical information system (GIS) application programming interfaces (APIs) and collaborative gaming on mobile devices can increase public engagement with citizen science projects. These also increase biodiversity model designers' expectations of the validity of data collected through such interactions. My research explores whether motivating involvement with citizen science through gamification or open-ended play increases user engagement and enhances data-quality.

1.4.2 Motivational rewards from gamification

Gamification is the incorporation of game design features to reward actions or processes that are not inherently game-like. In the context of avian surveying, an exemplar gamified element is the leaderboard model in the BirdTrack web application from the BTO[8]. Gamification has been identified as a tool for encouraging participation in human-in-the-loop (HitL) computations[9] which contain human intelligence tasks (HITs)[10]. However, gamified citizen science projects frequently suffer from a mismatch between game success mechanics and the scientific needs of those designing the games. Gamified interactions frequently provide rewards for high submission quantity in the absence of concurrent submission quality; undesirable results flourish when game mechanics reinforce participant success without mandating actions beneficial to the creator's goals. I explore, through interaction designs incorporating multiple stakeholders' needs, gamified interactions which motivate participants who would not otherwise be engaged with citizen science and compare their performance with that of participants who interact with similar data through more complex games and open-ended play.

[7] http://www.nfbr.org.uk/?q=user/68
[8] https://app.bto.org/birdtrack/main/data-home.jsp
[9] HitL computations benefit from computational complexity reduction by humans.
[10] HITs are inference tasks which exceed the capabilities of digital computers.

1.4.3 Motivating engagement through purposeful games

My research into avian bioacoustic games extends the model for participant interaction with avian bioacoustic data beyond rewards for gamified interaction into purposeful gaming. Although not previously applied to citizen science, purposeful games engage and motivate participants while simultaneously providing validation of their collected data. This is useful for those building biodiversity models. Luis Von Ahn identifies benefits of human computation in games with a purpose (GWAPs), given interfaces well integrated into computational problems characterised by input-output behaviour[4, 5]. GWAPs describe models where citizen scientists can contribute directly to databases, although contributions may not be given much credence, or to processes of data-validation and trust enhancement. Standard game design heuristics such as "goals that are both well-specified and challenging lead to higher levels of effort and task performance than goals that are too easy or vague"[ibid] must be kept in mind when designing ludic[11] interactions. I designed citizen science avian surveying games where trust associated with future data are validated by performance success in games designed to query participants' underlying knowledge, allowing hypothetical contribution trust metrics and user credibility scores. While not contributing to the collection of increased data-volume, my research designs have involved the development and analysis of games which provide motivational reward mechanisms for interacting with data. Games and open-ended play can be designed to reward complex interactions with data and enhance participant motivation with rewards linked to learning or data-exploration. My research explores how game feedback which provides participants with motivational rewards for learning can shift engagement motivation from quantified success — rewards for producing new data — to motivation to pursue continued interaction through open-ended play.

1.4.4 Motivation from open-ended play

Open-ended play has been explored in the context of video games but not heretofore as a mechanism for citizen science engagement. Within the game design community, success measures of open-ended play are frequently posited using heuristics about flow[222]. Flow theory proposes a foundation for identifying elements of an experience which provide users intrinsic motivation[53]. Participants in citizen science projects, if extrinsically motivated by initial gamified interactions whilst intrinsically motivated to pursue further knowledge development through purposeful games, will participate in projects longer than if they are only provided one motivation source. Intrinsic motivation from interactions associated with learning and exploring data can lead to increased user retention in a citizen science project through the development of engagement opportunities associated with open-ended

[11] Showing spontaneous and undirected playfulness.

play.

My research combines game design with citizen science interaction models which leverage mobile devices as tools for training games and for collecting sensor data in the field. I developed mobile applications which enable novice users to accurately select a visual representation of birdsong in a noisy environment. I investigate the role of different forms of intrinsic and extrinsic motivation driving interaction with my games; these range from competitive gamification of knowledge acquisition to aesthetic goals associated with birdsong composition as open-ended play.

1.5 Thesis outline

In this thesis I explore enhancing citizen scientist interaction with and motivation for participating in a project through a set of mobile interfaces, packaged as games, co-created with local avocational ornithologists, for data-collection, knowledge development, and open-ended play.

In Chapter 2 I introduce theories of motivation, discuss how play motivates engagement, necessary for user retention, and offer caveats regarding contribution trust limitations when assessing data-quality. I survey the game psychology literature, identifying a spectrum from gamification to open-ended play. I discuss the role of flow in designing motivational games and summarise similarities inherent to, and my differentiation between, gamification and play. I summarise approaches to audio interaction design and identify affordances of various interfaces for the representation of bioacoustic signals. I conclude by proposing two novel frameworks for discussing motivation, engagement, and play: the first provides dimensions for discussing game design characteristics for engaging users and enabling flow; the second proposes dimensions for discussing user motivation and ancillary engagement and retention in the context of citizen science.

In Chapter 3 I reflect on prior work in biological monitoring and identify how my research supports and enhances preexisting approaches. I present a survey of conservation monitoring targeting the opportunities presented when designing models and metrics from citizen scientist data. I consider the value of focussing on acoustic data for avian monitoring and review bioacoustic data-analysis approaches to avian conservation, critiquing the limits of purely statistical approaches. I review citizen science in principle and practice, focussing on design theories, ethical considerations, and prior implementations of bioacoustic ecological monitoring projects. I conclude with a summary of case studies of interaction designs of prior conservation projects, identifying conservation objectives and treatment of participant contributions.

In Chapter 4 I outline interaction design theory and practice as applied to citizen science. I describe preliminary experiments comparing familiar and novel interfaces for representation of and interactions with bioacoustic data to identify baseline visual audio data-representation familiarity and potential mobile touch interactions for region of interest (ROI) selection. Smartphone depictions of spectral representations of avian biophonies[12] in a noisy environment are explored in detail. This guided an initial interface design (constructed as a final year undergraduate thesis under my supervision) for exploring the potential for recording a library of geo-tagged bioacoustic data. I describe initial fieldwork performed with this application in RSPB Minsmere. I conclude with a summary of preliminary results and discuss how results regarding representation, interaction, and goal-state preference, *apropos* the frameworks in §2.7, guide design of the games described in chapter 5.

In Chapter 5 I describe development of a software framework for building games to explore the progression from gamified interaction to open-ended play and describe how interactions with each ludic mode motivate engagement and learning; I propose to explore qualitatively and quantitatively how different datasets and game combinations enable learning and enhance motivation. I introduce the context of my primary fieldwork, pursued in the Nidderdale Area of Outstanding Natural Beauty (AONB) in collaboration with the Wild Watch project[13], with two cohorts of elementary students. I provide details of my game implementations and discuss how each interface fits the model proposed in §2.7.1. I present my experimental design and data-collection procedures for investigating the efficacy of my interaction models in motivating and engaging citizen scientists, *apropos* described experiments and surveys. I discuss how I anticipate analysing each cohort's results to determine the suitability of my interfaces and the validity of my underlying designs. I conclude by presenting categories of questions regarding learning, motivation, and collaborative design, and enumerate research exploration questions.

In Chapter 6 I present qualitative and quantitative results from both cohorts, combining subjective survey results with objective results from user performance during play. After providing a summary of the experimental participants, I consider each research category introduced in §5.5.x. I identify interface and user behaviour characteristics which lead to strongest engagement and most effective learning. I evaluate learning and motivation across cohorts and interfaces and discuss how my interfaces work in the context of the framework introduced in §2.7.2. I conclude with results and discussion regarding my collaborative design process involving focus groups of local adult stakeholders.

In Chapter 7 I summarise conclusions from my research in each exploration category. I discuss my research contributions and include an assessment of the efficacy of my frameworks introduced in §2.7 as a foundation upon which to build the tools for future interactive citizen science projects. I discuss my game implementation framework as a basis

[12] The sounds vocalising animals create in their environment.
[13] https://nidderdaleaonb.org.uk/about-us/nidderdale-aonb-projects/the-wild-watch/

for rapid generation of targeted interactions providing motivation and enhancing learning. I identify research implications and provide a summary of research limitations and biases inherent to my experimental process. I conclude with a presentation of extensions to my work, including practical applications currently in development in light of my research, and identify steps that can be taken to overcome existing design issues and identify scope for potential future research.

CHAPTER 2

Designing for engagement through play

INTERFACES incorporating game elements are frequently proposed for engaging citizen scientists. Despite limited human computer interaction (HCI) focus in citizen science, a design-thinking approach to project interfaces is valuable[174]. I review theories of motivation and engagement and discuss mechanisms for quantifying trust in data-quality. I introduce theories of play and describe theoretical approaches applicable to designing citizen science project interactions. I survey audio interaction design as applicable to bioacoustic data and metadata. At the end of this chapter I propose two novel frameworks characterising (a) relationships amongst game mechanic complexity, data-representation and goal-state, and (b) game design complexity, participant motivation sources, and scope of biodiversity engagement.

2.1 Theories of motivation

A critical question for citizen science is what motivates people to participate. Historic approaches informing modern theories of motivation include Maslow's hierarchy of needs[148]. While professional scientists are motivated by employment, avocational scientists have motivations that satisfy a different set of priorities in Maslow's hierarchy. This thesis explores models for avocational participants' motivation, and ways that technology can encourage public engagement with citizen science. While critics have taken issue with the theory's hierarchical nature, the needs defined (physiological, safety, community, esteem, and self-actualisation) remain foundational to current motivation theory. Operant theory[206] proposes that "reinforcement contingencies in the environment control behaviour . . . , preclud[ing] the existence of inherently satisfying activities performed for non-separable outcomes"[234] and eliminates the self-actualisation need. Interaction design analyses of human motivation retain much of Maslow, superseding Skinner. The primary theory guiding my research into motivation from games is self-determination theory (SDT), comprised of several component theories which guide analysis of interface designs and game mechanics.

2.1.1 Self-determination theory

SDT, "a macrotheory of human motivation that is principally concerned with the potential of social contexts to provide experiences that satisfy universal human needs"[180] posits three essential needs summarised in basic needs theory (BNT), autonomy, competence, and relatedness, through which motivation can be discussed (fig, 2.1, pg. 34)[193, 195]. Autonomy represents the need "to be causal agents of one's own life and [to] act in

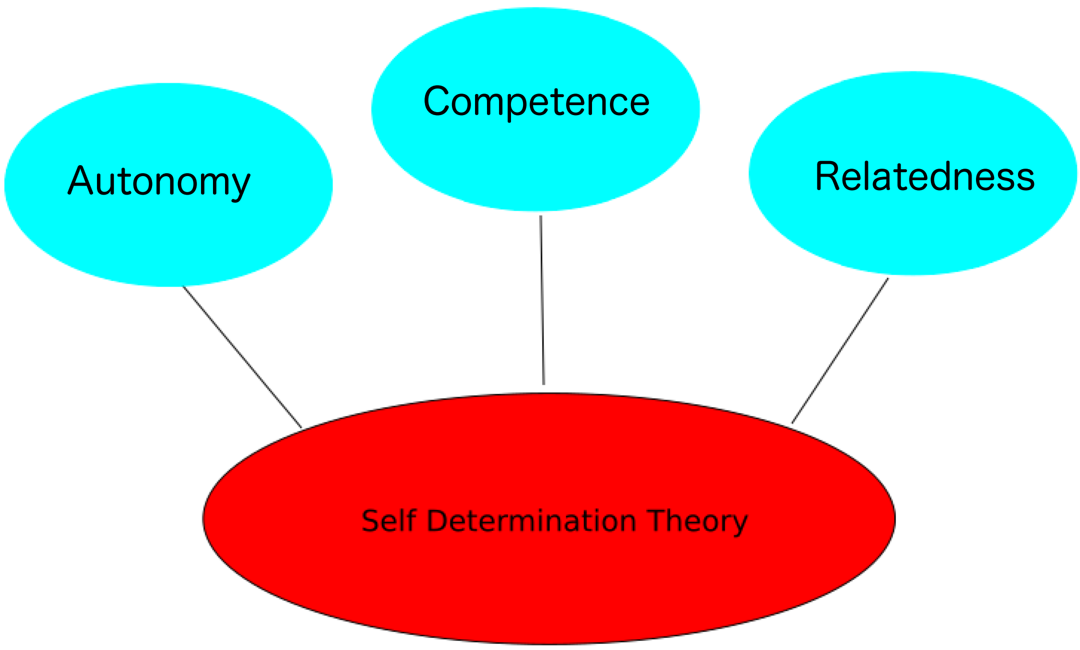

Figure 2.1: Component needs of BNT, contributing to SDT, through which motivation can be analysed, are autonomy, competence, and relatedness.

harmony with one's integrated self"[61]; competence, the need to control outcomes and experience mastery[242], prerequisite for flow; and relatedness, the need to interact with and feel positively connected to others[16]. SDT categorises motivations as intrinsic or extrinsic[59, 60] and forms the basis for numerous analyses of citizen science project participation[167]. Intrinsic motivations produce autotelic[1] rewards for participation, such as explicit pursuit of enjoyment from participation in un-gamified citizen science projects[159]; subsequent analyses[166] sometimes include the hedonic motive, excluded

[1]Having a purpose in and not apart from itself.

here. Conversely, extrinsic motivations drive engagement through expectation of separable outcomes[234]. While both have been shown to promote performance gains, only intrinsic motivation has been associated with improved psychological well-being, enhanced creativity, and learning[153, 193]. My research explores whether extrinsic ludic motivation drives learning, counter to this claimed limit on association. The following SDT sub-theories describe how game-mechanic feedback motivates citizen scientists engaged through play.

2.1.1.1 Intrinsic motivation theory

Cognitive evaluation theory (CET) posits that people by nature possess intrinsic motivation which encourages engagement in curiosity-based behaviours, sometimes called interest[234]. Interest refers to the initial attraction individuals associate with activities; game-mechanic feedback influences interest positively or negatively. Positive feedback enhances intrinsic motivation by supporting competence, the perceived extent to which actions cause desired consequences[153]. Controlling feedback, which forces specific behaviours, undermines intrinsic motivation by limiting autonomy[ibid](fig. 2.2, pg. 36). Interface designs can support users who either prioritise competence — which benefits from informational feedback — or autonomy, while enforcing data-collection protocols. Competence alone fails to enhance intrinsic motivation when positive feedback perceived as controlling diminishes autonomy[ibid]. My gamified feedback designs consider that intrinsic motivation increases when participants experience their behaviour as self-determined.

Causality orientation theory (COT) classifies feedback as controlling or informational, depending upon whether users experience their actions as self-determined[153](fig. 2.2, pg. 36). COT postulates that predicting motivation from feedback is contingent upon classifying users, as in CET, as control- or autonomy-oriented[64]. Feedback perceived as controlling *vs.* informational influences autonomy- and control-oriented users differently; control-oriented users perceive feedback as diminishing autonomy, which decreases intrinsic motivation, while autonomy-oriented users perceive identical feedback as informational, which increases competence and intrinsic motivation. The difficulty of predicting participant motivation requires my game interface designs to satisfy both user archetypes.

2.1.1.2 Extrinsic motivation & organismic integration theory

While CET and COT describe intrinsic motivational feedback effects, organismic integration theory (OIT) considers how extrinsic motivation influences intrinsic motivation; again, external motivation affects autonomy- and control-oriented users differently. The most controlling form of regulation of extrinsic motivation is external; introjected regulation is less controlling; identified regulation allows a greater degree of autonomy; and integrated regulation fully supports autonomy (table §2.1, pg. 36). Internalisation involves endorsing the value of extrinsically motivated behaviour[234]; I explore whether data-quality is

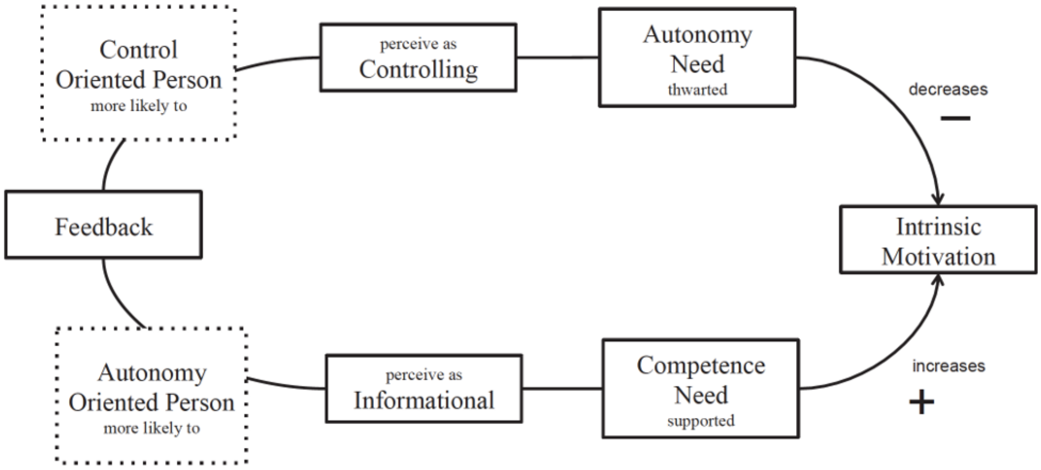

Figure 2.2: How feedback affects autonomy- and control-oriented participants' intrinsic motivation: control-oriented participants, given autonomy, are negatively motivated, while autonomy-oriented participants, given competence, are positively motivated[153]

Taxonomy of extrinsic motivation				
Regulatory Style	External	Introjected	Identified	Integrated
Locus of Causality	External	Somewhat External	Somewhat Internal	Internal
Associated Processes	Salience of extrinsic reward	Ego Involvement	Conscious valuing of activity	Hierarchical goal synthesis
	Compliance/ Reactance	Focus on approval (self or other)	Goal Self-endorsement	Congruence

Table 2.1: Taxonomy of extrinsic motivation, of relevance to OIT, adapted from[193]. External and introjected regulation contribute to controlled motivation, identified and integrated regulation support autonomous motivation.

enhanced when extrinsic ludic rewards create intrinsic value from learning to identify avian species. Perceived self-worth increases with task completion if participants retain sense-of-self when performing extrinsically-motivated tasks. This continuum of motivation intentionality, mediated by internal and external control, suggests that game elements may be intrinsically motivating regardless of external reward[194, 200]. This guides my exploration of whether open-ended play provides more motivation than gamification.

2.1.2 Applying theories of motivation to citizen science goals

Achievement theory differentiates between performance and mastery goals. My gamified designs described in §5.3.3 provide intra-play performance rewards, while the complex game designs introduced in §5.3.4 guerdon primarily mastery upon completion. Mekler et al. in a "[m]eta-analysis on the effects of performance and mastery goals on intrinsic motivation found that informational feedback increased intrinsic motivation for mastery goals, whereas performance goals were unaffected"; only intrinsic motivation yielded higher data-quality[153]. Nov et al. posit a motivational model for engaging citizen scientists to continue and increase participation based on a classification of motivations for engagement in citizen science projects which encompasses collective motives, norm-oriented motives, reward motives, and collective identification[166]. An additional class proposed to increase scientific output is the hedonic or intrinsic motive (used here synonymously, unlike within SDT), identified as enjoyment associated with participation[ibid]. The authors combine these motives into a framework based on Davis et. al.'s technological acceptance model (TAM)[58], adapted from the theory of reasoned action (TRA) posited by Venkatash[235], who incorporated intrinsic motivation into the TAM.

A positive external effect on intrinsic motivation is called crowding-in, a negative effect is called crowding-out[168]. Nov et al. propose an extension involving crowding effects to the relationship between intrinsic and extrinsic motivation characterised in OIT. An example of a crowding-in effect is when extrinsic motivation from reputation scores reinforces intrinsic motivation, contingent upon participant perception of self-determination[158, 167]. Coleman et al. propose a motivational model derived from Rotman et al.'s motivational dimensions of egoism, altruism, collectivism, and principalism[191]; they find altruism to supersede interest, intellectual stimulation, protection or enhancement of prior investment or reputation, social reward, and creative self-expression as a motivator[44]. Nov et al., in contrast, claim that egoism has the strongest influence[167], in support of Rotman's citizen science participation model (fig. 2.3, pg. 37). My analysis of citizen scientists'

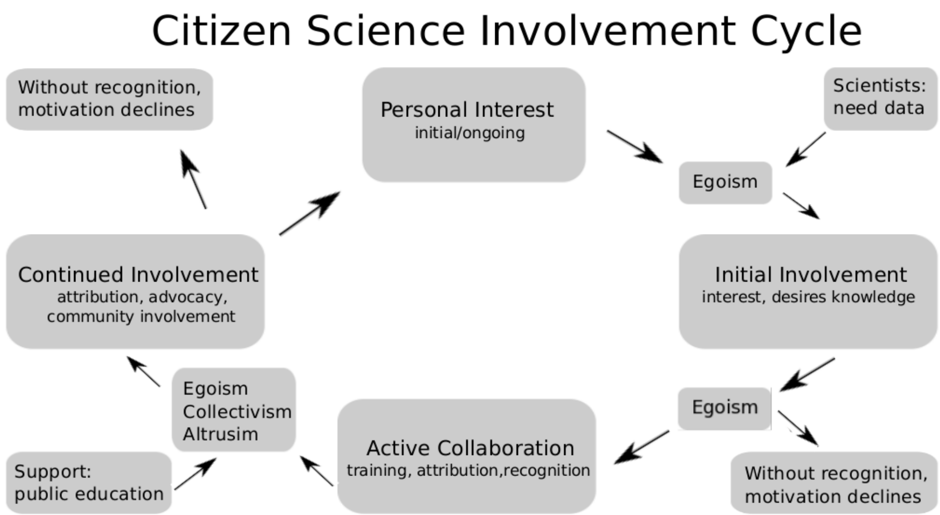

Figure 2.3: Citizen science involvement cycle, describing movement from initial involvement, through active collaboration, to continued involvement, adapted from [191]; the predominant role of egoism supports Nov's claims[167].

engagement considers combinations of these motivational dimensions while exploring how extrinsic ludic factors can influence intrinsic and extrinsic motivation. My engagement framework (§2.7.2) introduces a dimension to discuss how internal and external influences affect participant motivation.

2.2 Motivation & engagement

In general, engaging citizen scientists requires firstly that participants be intrinsically motivated for recruitment, after which extrinsic motivators maintain continued involvement. Incorporating my games into school curricula overcomes potential lack of intrinsic motivation, but introduces biases.

2.2.1 The motivational arc

Crowston and Fagnot propose a motivational arc for engaging participants in massive collaborative projects[51]. The arc trajectory traces from initial participation to continued participation, to meta-contribution (fig. 2.4, pg. 39). Initial project contributions result from a combination of intrinsic motivators, including desire and perceived contribution capability, and extrinsic motivators, including external gains from contributions. Sustained contribution is intrinsically motivated when participants associate contributions with altruism and identify with project ideology, and extrinsically motivated by assimilation into project culture and continued positive cost benefit analyses. Motivating meta-contribution occurs when participants' trust in project communities supports development of community knowledge and standards. My design research explores collaborative designs which motivate movement along this arc and encourage engagement, learning, and data-production.

2.2.1.1 Quantifying & qualifying engagement

Citizen science projects frequently quantify participation as volume of data produced. Quantity without quality is insufficient for predicting project success and is a poor engagement metric if data-quality is not quantified. Lukyanenko et al. propose data-quality dimensions of accuracy, precision, sample size, and sampling procedure standards, with inevitable trade-offs[141].

Motivational factors which affect participation rates and increase data-quality include participant enjoyment, perception of reciprocity, identification with community, and perception of group membership[167]. My qualitative survey research measures participant enjoyment, correlating this with quantitative game-play success, while my collaborative design approach builds community and perception of group membership. Nov et al. found that contribution quantity correlates with collective motives, norm-oriented motives,

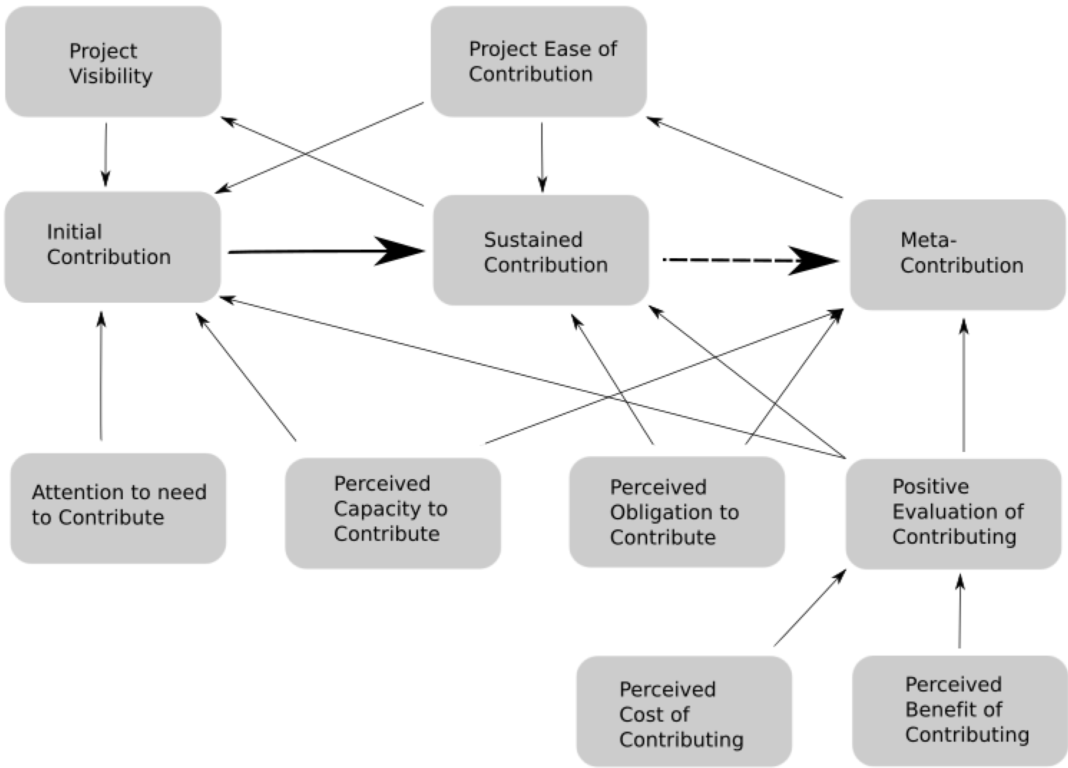

Figure 2.4: The motivational arc of citizen science contribution, adapted from[51], identifies sources of motivation as participants advance from initial, via sustained, to meta-contribution.

reputation and intrinsic motives such as social affiliation, utilitarian motives, self-expression, and commitment to community[ibid]. Collective and intrinsic motivations are powerful, but project attrition rates are high, due to perceived "distinction[s] between the volunteers making the contributions, and those benefiting from the aggregate effort"; however, "expectations of reciprocity had a significant effect only on contribution quantity, whereas altruism influenced the quality of contributions"[166, 167]. Projects benefit from supporting collective motives by incorporating reputation management internally — recognising users' contributions within the community — and externally — citing participants in project publications. Gamified levels, which allow participants to move up a contribution hierarchy, motivate engagement when designs incorporate citation and other attribution rewards. My collaborative design approach rewards participants by supporting reputation and building community. Nov et al. note that "[e]ffects of reputation are determined to a large extent by what indicator of performance is made visible publicly" and that intrinsic and extrinsic motivations are not additive as economic theories assume[ibid]. They conclude that "the fact that intrinsic motivation was not found to enhance quality stresses the need to develop more enjoyable, game-like, participation mechanisms . . . [and] mechanisms such as social network features should be put in place to create and emphasise social influences, linking them to the quality of one's contribution, so that norm-oriented motives would be positively linked to contribution quality"[ibid]. I contrast intrinsic motivation from open-ended play with extrinsic motivation from gamified interactions as means for increasing participant engagement, data-quantity, and quality. My research explores the space of prior conflicting results where only intrinsic motivation was found to enhance data-quality[153], and where it was not[167].

2.2.1.2 Designing for motivation

Tinati et al. contend that "most citizen science projects rely on contributions primarily driven by intrinsic motivations" and categorise projects as either *crowdsourcing* — treating projects as large-scale volunteer-driven human computation systems, or *online communities* — communication systems which support citizen science work[228]. They propose interaction design objectives and identify factors influencing project success, including task granularity, task completion speed, and the frequency and form of rewards[228]. Interestingly, motivation increases when tasks encourage participants to make guesses. This apparently avoids the *don't know* effort trap that arises when participants fail to provide data due to perceived lack of competence[126, 228, 229]. Given findings that unmotivated users quit projects within 90 seconds of exposure, Tinati et al. propose that designers avoid tutorials in lieu of in-task user interface (UI) guidance, concluding that performance feedback engages users and that adding task context improves community engagement[228]. My designs avoid tutorials and provide a spectrum from familiar to abstract games which provide motivation through performance feedback.

Rotman et al. identify motivations common to committed participants and discuss why most projects fail to maintain volunteer participation[190]. While initial participation frequently stems from self-directed motivation, long-term engagement is predicated upon collaborative motivations. In projects with poor scientist-volunteer communication trust levels suffers and participation flags. In cases where volunteers form communities, there is less user attrition; my design process explores enhancing engagement through community development.

Avoiding motivational failures from UIs can increase data production by long-tail participants. Eveleigh et al., exploring the role of weak contributors in citizen science, describe the solitary experience of 'dabbling'[2]. They found that offering opportunities for engagement through brief interactions allows projects to tempt dabbling participants to increase their contributions gradually, albeit by small degrees[75]. My short-duration games support dabbling behaviours. Extrinsic gamified rewards combining competition and target-setting serve dabblers who, as an audience, encourage community growth. Projects typically ignore dabblers, but involving them with discrete time-limited tasks incorporating quality feedback mechanisms increases data-quantity[ibid]. Eveleigh et al. reported that 94% of their study participants created 15% of the data[75]. Participants who contributed

[2]Curiosity-driven behaviour done in free time, preferably involving highly granular tasks.

minimally cited task boredom and concerns about contribution usefulness, given lack of communication from project scientists[ibid]. However, their results suffered from sampling bias as participants in a survey about motivation are *a priori* likely to have stronger egoistic or intrinsic motives. Designing for dabbling is key to motivating engagement with long-tail participants[3]. In both data-collection and data-processing interactions, increased task granularity allows users to anticipate minimal time commitment. My research applies the following design criteria to enhance dabbling throughout the motivational arc: facilitate independent work and participant choice during initial contribution; provide short-duration tasks, encouraging sustained contribution; and publicise scientific outcomes referencing participants, stimulating meta-contribution[51, 75].

2.2.2 Stakeholder engagement: from scientists to communities

Projects engage different classes of new users by providing multiple entry-points for contribution at different commitment levels, given variable prior domain knowledge[20]. Regardless of entry-point, "citizen science can contribute positively to social well-being by influencing the questions that are being addressed and by giving people a voice in local environmental decision-making"[ibid]. My designs place amateur knowledge alongside scientific expertise to frame relevant questions, empowering participants to engage with the decision-making process. Bonney et al. classify citizen science projects by participant activities. These include categories of data-collection, data-processing, curriculum-based and community science[ibid]. My designs support interactions in multiple categories. My collaborative game designs involving multiple local stakeholders are examples of community science projects designed to achieve public understanding outcomes by involving participants at multiple stages of the scientific process. Curriculum-based citizen science can further engage participants, even those typically underrepresented, because participation is enforced by educational requirements. I pursue this approach for my fieldwork, introduced in chapter 5.

Irwin identifies a feature of citizen science projects, *public understanding of science*, which is significant for analysing engagement as deliberative, participatory, and essential for fostering science-in-society relationships[108]. Bonney et al. contend that:

> "both local and scientific communities ought learn new methods of discourse and deliberation while challenging scientific institutions to expand their notions of what expert knowledge is and whose knowledge counts within the realm of science . . . issues of trust, fairness, equity, and risk will need to be embedded into the dialogue as seamlessly as issues concerning volunteer recruitment, protocols, and data quality"[20].

[3]The long tail describes the majority of project participants whose contributions are minimal.

Engaging scientists and conservation practitioners with my designs and their output has been integral to my research design process.

Maslow's self-actualisation need motivates volunteer activities when participation supports social affiliation, personal achievement, and esteem[149, 205]. Silvertown et al. report that training most increases volunteer performance and engagement. Novices, lacking preconceived methodological notions, respond better to training and provide better consistency, although their work is generally slower and their boredom threshold lower than amongst professionals'. Given time and training by project scientists and through interactions with experienced participants, novices can reach expert levels of domain-specific knowledge[205]. My research explores motivating engagement through curriculum-based training games. Silvertown et al. found that users were most motivated to participate by intrinsic motivations to help the environment and increase learning, while project involvement, esteem, and social interaction were less motivating[ibid]. My research involves motivating students, community stakeholders, and scientists to become collaborative co-designers of a bioacoustic citizen science project. Given various sources of volunteer motivation, my recruitment messages are tailored to specific modes of interaction within the project and convey to stakeholders how their wants will be satisfied.

2.3 Assessing data-quality

When building models from variable quality data, application of computational trust and reputation metrics can strengthen model-based assumptions derived from crowdsourced data[7, 151, 196, 248, 249]. I design games to directly validate user knowledge. A data-quality framework used to analyse citizen science output encompasses dimensions of: intrinsic data-quality — is it believable?; contextual data-quality — is it complete and timely?; representational data-quality — can it be interpreted?; and accessibility — can it be accessed for use?[177, 240]. Chapter 4 describes experiments where users make and annotate avian recordings. The experimental UI requires syntax, format, and value validation on constrained input to compare contributed data against ground-truth, limiting erroneous data inclusion. Responsibility for maintaining data quality falls to interface designers, ensuring citizen scientists contribute within a project's protocol constraints. In my games, described in chapter 5, training ground-truth allows play-performance success to become a direct metric for trust in subsequent user-collected data.

2.3.1 Colloquial variation & incomplete knowledge

Crowdsourced metadata annotations are susceptible to noise due to *sentiment ambiguity* and *lexical uncertainty*; weather descriptions are partially subjective[4] and many languages enforce mismatching colloquial identification taxonomies[5]. However, even noisy, non-expert annotations are metadata with potential for modellers[104]. Referring to eBird[6], where frequent contributors are likely knowledgeable amateurs, Lukyanenko proposes that projects support inclusivity through "a flexible, instance-based approach to data-collection that allows a contributor to classify data at the level at which they feel competent"[141]; this approach enhances infrequent contributors' motivation, avoiding the *don't know* trap[228], at the expense of protocol consistency. Users' capacity to name bird species colloquially as opposed to scientifically varies with native language, prior knowledge, and training; a controlled vocabulary for structured entries reinforces scientific trust in data collected by networks of amateurs, as noted in relation to the Christmas Bird Count (CBC)[7][92].

Wiersma et al. submit that the data-quality dimension of accuracy be given preeminence, although precision, timeliness, completeness, and believability are also relevant[243]. They posit that "if the beginning birder is frustrated by their lack of ability to properly log the species identification in a site like eBird, they may simply opt out of participating and sightings would go unreported, thus rendering the data of lower quality on the dimensionality of completeness"[ibid]. They find that accuracy improves when contributors contribute at their comfort level, "freeing citizen scientists . . . from the data entry constraints imposed by scientists/experts may increase the data quality dimensions of accuracy and completeness"[ibid]. Conventional mistrust of citizen scientists' results may be unwarranted, even though many datasets suffer from poor data-quality; while anonymous participants can sabotage data, risk is diminished when participants are in direct contact with scientists or when a trust metric is created for contributions. In my preliminary research, syntax, format and content constraints are enforced on data-entry; when incomplete knowledge blocks users from contributing to their collection library, they are guided to my games which support relevant learning.

2.3.2 User validation

Primary considerations for data quality are variable observer skill and non-uniform spatiotemporality. In eBird, observer skill is modelled based on ranking *species accumulation rates*

[4] *e.g.* distinctions between sleet, freezing rain, and mixed precipitation.

[5] *e.g.*English supports ape/monkey and turtle/tortoise distinctions while many languages do not. Conversely, colloquial 'batfish' in the Caribbean and the Pacific are entirely different orders.

[6] The Cornell Lab of Ornithology (CLO) online community for checklist data. https://ebird.org/home

[7] https://www.audubon.org/conservation/science/christmas-bird-count

and comparing species accumulation curves across users; expert feedback, requested when an improbable claim is flagged due to inconsistencies with historic data, improves data-quality before the data are integrated into the eBird database[117]. Kelling et al, propose a "sensor-calibration" approach to determining eBird contributors' abilities to detect and identify birds; users who produce more records sit higher on the species accumulation curve and are assumed to recognise more birds accurately. However, this assumption is unwarranted unless the records can subsequently be validated. When records include audio, validation is more likely, but most records are visual; my audio-based approach increases validation likelihood.

When crowdsourcing raw data, creating metrics helps distinguish individual's contribution quality as errors are not randomly distributed[95]. In data-quality analysis, crowdsourcing is controversial as there are risks when amateurs do not follow common practices in data-collection, verification, and use; relevant aspects of quality include "completeness, logical consistency, positional accuracy, temporal accuracy, thematic accuracy, purpose, usage, and lineage"[ibid]. For bioacoustic or ecological data-collection, social data-collection approaches — *transect parties* — could concurrently provide mechanisms for engagement and community validation of data-quality and solve problems of non-uniform spatio-temporal results, avoiding inconsistent geographic coverage, the data deserts identified in §1.3. Local citizen science training schemes can overcome risks of individual error through data redundancy if multiple participants concurrently collect data; my collaborators ran training schemes for small groups of participants. However, data-validation mechanisms, which define ground-truth through repeated data-collection in a specific spot by multiple participants, are problematic for avian bioacoustics, as transient data require participants to be contemporaneously co-located. My game designs support user-validation, through game-performance metrics, which is more applicable to my research than data-validation, through upload protocol constraints.

2.3.3 Guidance for designing interactions

Sprinks et al., applying task workflow design to citizen science projects, found that while volunteers wanted greater autonomy and variety in their interfaces, this yielded worse performance[211]. Limiting users' possible outcomes led to closest agreement with experts' results. Autonomy yielded neither better data-quality nor quantity, while limiting task types yielded greater quantity. Input protocols for my preliminary fieldwork designs are guided by these results and the supposition that constraining input supports control-oriented users; however, quantity without quality is insufficient. Although Mekler et. al., applying CET and COT, have theorised that autonomy provides intrinsic motivation, they found that: gamification elements increased data-quantity but not quality; quantity was a negative predictor of quality; game elements were not seen as informational; and the motivational appeal of many games lies in providing challenges to master, allowing users to experience competence[153]. My initial game design, introduced in chapter 5, provides motivation to increase quality through gamified play. My game interfaces constrain teaching to scientific ground-truth, limiting subsequent erroneous contributions. My final design provides open-ended interactions with constrained data, posited to increase quality for autonomy-oriented users.

2.4 Theories of play

A proposition of education research is that play simultaneously encourages motivation and increases learning[170]. Paras et al. contend that games "act as effective learning environments by integrating reflection into the process of play, producing an endogenous[8] learning experience that is intrinsically motivating" and propose that "games foster play, which produces a state of flow, which increases motivation, which supports the learning process"[ibid]. My research explores whether citizen science games, introduced in schools, motivate participant engagement while enabling bioacoustic learning.

2.4.1 Defining games & play

In order to understand the variety of ways that games can be used in citizen science, it is necessary to consider the scope of definition for games and play. Predating computers, Huizinga defined games as non-serious engaging activities structured by rules and social boundaries[105]. Avedon extends this definition of games to encompass voluntary activities bounded by rules, entailing conflict and unequal results; this represents a shift from play, since competition may stress participants[12]. Greenhill defines play as an activity conducted for its own sake and characterises play as: self-chosen and self-directed; having means more valued than ends; having structure or rules; imaginative, non-literal, and removed from serious life; and involving an active, alert, non-stressed, frame of mind[94]. Crawford, in an early specific analysis of computer games, proposed that they: require representations of reality; are predicated on interaction between system and user; and provide safe conflict through simulation[49]. Salen and Zimmerman define games as "a system in which players engage in artificial conflict, defined by rules, that results in a quantifiable outcome"[198]. Juul et al. describe 6 main features of games: rules, variables, quantifiable outcomes, value-laden outcomes, and player effort and investment[114]. Some of my designs involve competition against a clock — characteristic of games, while others allow open-ended exploration without quantifiable outcome or time limitations — characteristic of play. All my game designs contain interaction rules; however, quantified,

[8] Having an internal origin.

value-laden outcomes — scores — are only elements in my constrained goal-state designs.

2.4.1.1 From gamification to play

I categorise citizen science interfaces by whether activities allow playful behaviours or constrain platform-participant interactions. Prior gamified projects implement some game features — points, levels, badges, and leaderboards — to motivate participation, but constrain goal-state complexity. Gamified design elements are theorised to: only provide extrinsic motivation; often be weakly tied to underlying project goals; and, OIT predicts, be insufficient to motivate extending participation. Some researchers limit the term *gamification* to describe systems which implement extrinsic motivators, and *gameful design* to describe systems in which intrinsic motivators exist. I use the term gamification without limiting motivation source, *apropos* modes of regulation identified for OIT, and consider instead game elements implemented and possible types of interaction.

Play is distinguishable from games in that games require both structure and goals. Gamification, a subset of *playification*, combines non-game actions with motivational ludic elements[63, 144, 200]. Other terms used synonymously or overlapping with gamification include *applied gaming*, *productivity games*, *behavioural games*, and *surveillance entertainment*[200]. Seaborn et al. propose that *gamefulness* refers to the lived experience, *gameful interaction* refers to the objects, tools, and contexts that bring about the experience of gamefulness, and *gameful design* refers to the practice of crafting a gameful experience[ibid]. Bowser et al. define *serious games* as stand-alone games with a primary purpose other than entertainment, giving as an example FoldIt[9], a protein folding puzzle game[25]. They contend that stand-alone citizen science games are more prevalent than gamified projects but this is unsupported in current literature. My designs incorporate gameful interactions for knowledge development through applied gaming.

Games with a purpose (GWAPs), alternate reality games (ARGs), and pervasive or augmented reality games incorporate reality into the representational game world and may provide more complex motivation for citizen scientists[4, 5, 200]. My software designs support users collecting — necessarily in nature — and playing — *in situ* — with data. Prestopnik et al. explored the role of diegesis[10] in citizen science game design, where participant contributions to science are embedded within a game world narrative, but found no increased engagement or scientific output[179]. My research explores the relationship between participant motivation, given gamified and gamised interfaces, and outputs useful to the project, data selected or participants trained.

[9] https://fold.it/portal/

[10] Having a narrative or plot.

2.4.1.2 Game mechanics & elements

Game mechanics describe the interactions supported when playing a game. Game mechanics for citizen science projects should support pattern recognition, data-collection, and data-analysis[200]. My gamised designs provide a spectrum of mechanics, from card matching, to puzzle reconstruction, to open-ended composition creation, motivating diverse player archetypes. Greenhill et al. define gamised behaviour as user-generated play in a platform, in contrast to gamification, where designers purposely embed game mechanics. They propose that:

> "when a dichotomy is established between the process of data categorisation and the science as end product, could an understanding of play as categorisation and work as science emerge? ... [for if that] definition is carried through in terms of understanding citizen science participation in an online crowdsourcing platform, a lowering of enjoyment must ensue"[94].

Game mechanics that turn work into play enhance citizen science projects if they merge work and play so that participants experience flow. My individual games incorporate familiar play mechanics while supporting users' desire for variation in playful interactions through designs which avoid presenting actions as work.

Game elements refer to features present in games. Game element selection considerations include identifying the player archetypes each game will support, methods of on-boarding, and how mastery will be achieved[200]; an experience can be designed, but not guaranteed, for a given user. Points, levels, badges and leaderboards are game elements supporting gamified success metrics which provide information performance feedback, according to COT[153]. Prestopnik et al. create a taxonomy of player archetypes, including killers, socialisers, achievers, and explorers and identify game elements which motivate each[175]. In chapter 6 I present and discuss performance results from subsets of my participants who represent achievers and explorers.

2.4.2 User-centred game design

Seaborn et al. suggest that "user-centred design methodology may help elucidate intrinsic motivators for a given user population"[200]. They present *universal design for learning* which considers how best to provide for users through diverse presentation of content, mastery through multiple activities, and multi-linear learning paths[ibid]. They state that *situated relevance* means that users make decisions about what is meaningful, while *situated motivational affordances* entail a match between user archetypes and game elements[ibid]. Nicholson states that *user-centred design* connects these disparate theories, presenting "meaningful gamification through a user-centred exploration of theories behind OIT, situat[ed] relevance, situated motivational affordances, universal design for learning, and player-generated content", and proposes *meaningful gamification* built on intrinsic rather than extrinsic motivation[164]. My simple designs presume that gamified elements provide only extrinsic motivation. In contrast, my advanced game mechanics, including support for open-ended exploration, provide intrinsic motivation. My incorporation of end-users into the design process supports construction of a game-design framework which supports universal design for learning, allowing diverse presentation of content, mastery through multiple activities, and multi-linear learning paths.

2.5 Designing motivational games

Games can support intrinsic motivation by satisfying the BNT needs of autonomy, competence, and relatedness[180]. Researchers have found that autonomy-supportive and competence-supportive game features were viewed positively by users[193, 195]. Early game design research identified a taxonomy of heuristics to support intrinsic motivation, encompassing challenge, fantasy, and curiosity[144]. Challenge and curiosity provide motivations for citizen scientists in generalised GWAPs and diegetic fantasy has been incorporated in subsequent work by Prestopnik et al.[4, 5, 175, 179]. Additional game design heuristics include variable difficulty, multiple goal levels, performance feedback and informative feedback; an example of the latter being that "[t]o engage a learner's curiosity, feedback should be surprising. . . . [t]o be educational, feedback should be constructive"[144].

Mekler et al. posit that gamification elements (*e.g.* points, levels, or badges) "afford feelings of competence and hence enhance intrinsic motivation and promote performance gains" but counterproductive over-reliance on these elements diminishes intrinsic motivation[153]. Citing Deterding et al.[63, 64], Zhang et. al. suggest that "the effects of *individual* game design elements on user motivation should be studied, referring to the concept of *motivational affordance*, that is, *the properties of an object that determine whether and how it [. . .] supports one's motivational needs*"[252]. While extrinsic rewards may reduce, they do not invariably undermine, intrinsic motivation[153], as noted in table §2.1.

2.5.1 Gamification: success & limitations

Incorporation of games into life, building towards ARGs, is leading to the *ludification* of culture[200]. Gamification is gaining popularity in the game studies movement at the intersection of inexpensive personal data-tracking technology and increased user familiarity with simple mobile games[65, 200]. Gamified designs in non-game contexts incorporate game design elements, but such systems pursue divergent goals as they are neither purely functional software nor entirely games[66]. Critics of gamification disparagingly call it pointsification, lamenting that implementations lack a hook between game elements and tasks to which they are applied[17, 18, 70]. Prestopnik contrasts *game taskification* — incorporating non-entertainment tasks into a game world, thus forcing tasks to become part of the gaming experience, with *task gamification* — implementing *pointsification* for citizen science[176].

2.5.1.1 Designing games which motivate citizen science participation

While extrinsic motivators such as gamification elements and achievement recognition may extend participants' engagement, intrinsic motivation is paramount for initial participation. Iacovides et al., studying the motivations of citizen science volunteers, report that game elements supporting achievement recognition sustain engagement more than gamified designs which provide only points and levels[106]. This supports Nov's findings that collective and intrinsic motivations are salient[166]. Bowser et al., comparing gamers and nature-oriented participants in citizen science games, found game interfaces alienate traditional citizen science volunteers[26]. Von Ahn proposes three classes of GWAPs yet to be applied to citizen science: in output-agreement games users share an unclassified input and gain credit for sharing the same output classification; in inversion problem games one participant describes a known input and the other gains credit for classifying it; and in input-agreement games two players are given either the same or a different input and credit is gained when they determine which[4, 5]. Such game mechanisms motivate engagement and increase data-quality through social interaction.

My design research explores the claim that while contribution to scientific research is a primary motivator for citizen science participation, the opportunity to play games within projects supports continued involvement[54]. Curtis et al. distinguish between passive involvement in citizen science[11], and volunteer or distributed-thinking endeavours requiring active involvement[54, 96][12]. They report that motivation increases proportionally to task granularity, supporting my designs for dabblers. They consider how educational contexts affect intrinsic and extrinsic motivations for learning from tasks, relevant to my design of games for schools. They conclude by formulating a framework for categorising motivation in citizen science that extends SDT to projects which may contain games (fig. 2.5, pg. 50). This motivational framework influences both my artefact design decisions and my engagement framework introduced in §2.7.2.

2.5.1.2 Motivating the creation of quality data

Prestopnik et al. designed a platform for testing games as tools to motivate participants in citizen science[52, 175, 177, 179]. Using a design-science methodology encompassing tech-

[11] *e.g.* Distributed computing endeavours like SETI@Home. https://setiathome.berkeley.edu

[12] [96] downloaded from: https://povesham.wordpress.com/2011/07/20/classification-of-citizen-science-activities/

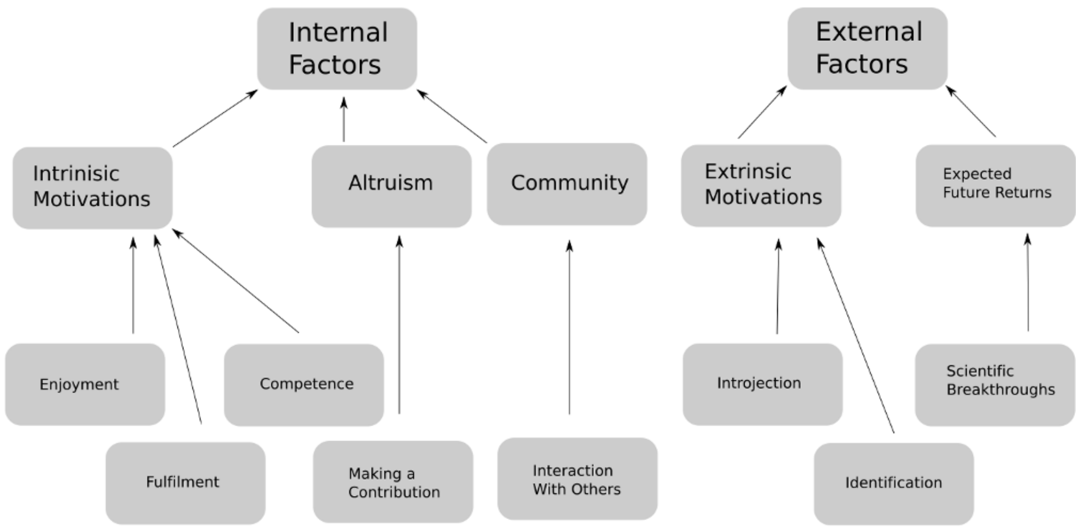

Figure 2.5: Motivational framework for citizen science project participants, adapted from a table in[54], broadly distinguishes between internal and external factors.

nological artefact design and evaluation of artefact efficacy, they explored how motivations affect data-quality, concluding that altruism motivates participation in data-collection but rarely in data-analysis. They note that games can dilute the information or experience that provides altruistic motivation and that when game designs interfere with sharing expert scientific knowledge, avocational experts' engagement is inhibited.

As an example of a game designed to enforce data quality, agreement-games are a class of purposeful games which are designed to derive ground-truth, such as a taxonomic classification, from user agreement about shared input data[5, 52]. In collaboration with domain experts, Crowston and Prestopnik crafted a pair of games[13] which explored motivations that engaged participants with citizen science projects and yielded quality data, allowing amateur replication of scientific behaviours and confirming the validity of participants' data[52, 177]. Their preliminary results showed that gamified elements did not solve the skewed effort distribution common to citizen science projects. The majority of users played a single game and most results came from a motivated few[ibid]. To reduce skew, albeit while introducing other motivational questions, my research was introduced in schools, where time allocated for play ensured data for comparison. The authors designed a subsequent citizen science data-analysis game with diegetic elements to identify variations in motivation between points- and story-based games where an in-game currency was earned for the same species-classification tasks as in their gamified designs; the diegetic game produced weaker scientific results because players cheated[179]. My designs avoid diegesis, and instead explore the motivational affordances of clearly defined mechanics and goal-state, conceptually complex mechanics and goal-state, and open-ended play mechanics

[13]Their games taught professionally curated ground-truth and asked participants to annotate observations using a constrained vocabulary, methods my research has replicated.

without goal-state. From this, I propose an arc of engagement across game-mechanics and goal-state complexity and compare results for predicting quality data.

2.5.2 Designing for engagement & flow

Prestopnik identifies positive player experience dimensions including flow, competence, immersion, and challenge[179]. Flow, "a state of maximal immersion and concentration at which optimal intrinsic motivation, enjoyment and high task performance are achieved", engages participants, potentially through games, increasing data-quality[53, 73, 202]. A flow state occurs when tasks can be focussed upon and completed in a space of time which allows concern for self to disappear and re-emerge upon goal completion. Short interactions supporting dabbling can, if repeated, lead to flow. Sweetser and Wyth's GameFlow model includes concentration, challenge, skills, control, clear goals, feedback, and immersion[222], all of which my designs support. While the game studies literature predominantly focusses on interface elements, mechanics, and game-play heuristics, few models exist for assessing player enjoyment relevant to engaging citizen science participants with games. Refined task granularity, also supporting dabbling, is tied to intrinsic motivation associated with flow, leading to game success; my designs provide this across gamified and playful interactions.

Jennett et al. propose a model (fig. 2.6, pg. 52) encompassing motivation, learning, and creativity for engaging long-tail citizen science participants[110]. They found that learning affects participation, as increasing content-knowledge and scientific literacy leads to behavioural changes, but increased scientific knowledge is only correlated with increased social interaction, not increased data contribution[ibid]. This model is similar to Iacovide's Gaming Involvement and Informal Learning framework as both "are influenced by the communities of practice literature ... emphasi[sing] the iterative relationship between learning and identity and how this is reinforced through participation in a range of practices"[106, 110, 241]. Jennet et al. propose that engagement increases when participants are offered creative outlets through project interactions; my open-ended play design explores this contention.

2.5.3 Case studies

Prior work in the collection and processing of data for citizen science and other disciplines using gamification, GWAPs, and gameful engagement are summarised and limitations identified.

2.5.3.1 Data-agreement games

While not citizen science, early acoustic GWAPs designed around music analysis included output-agreement mechanics which rewarded players when descriptive annotations were

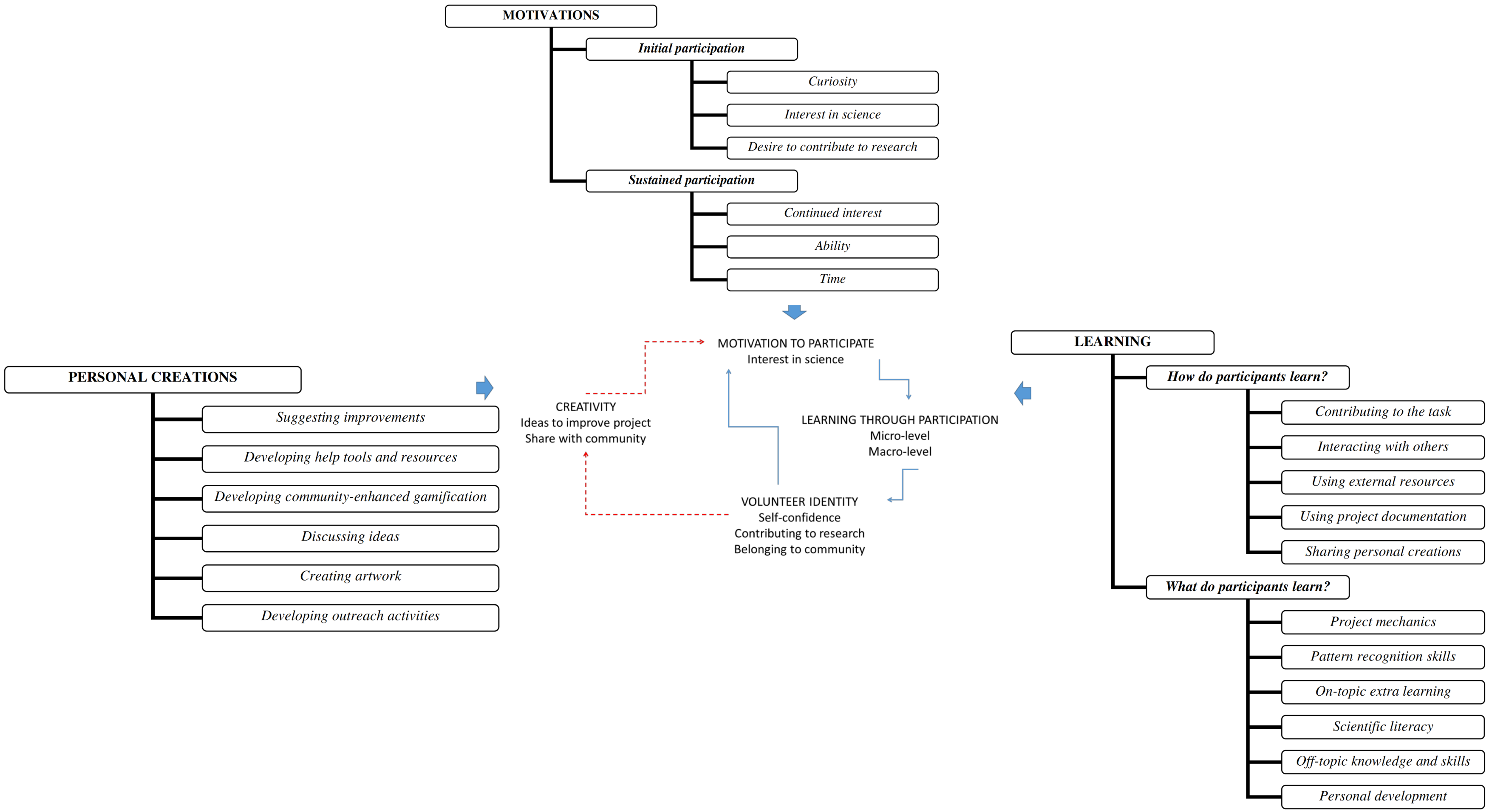

Figure 2.6: Framework for analysis of citizen science contributions in the context of motivation, learning, and creativity; adapted from[110].

Comparing Gamer & Citizen Scientist Motivation		
Motivation	Gamers	Citizen Scientists
Intrinsic Reward	X	X
Personal Interest	X	X
Learning / Education		X
Contribution to Science		X
Contribution to Public Good		X
Community Involvement	X	X
General Socialization	X	X
Personal Performance	X	X
Peer Competition	X	X

Table 2.2: A comparison of motivations for gamers and citizen scientists. Categories adapted from [25]. Selections in blue indicate where gamification can increase citizen scientist motivation.

novel and in agreement[145]. Law et al. designed an input-agreement game which collected comparative responses to questions of preference, similarity, and perception[132, 133]. Barrington et al. described game designs which functioned without ground-truth by applying social and demographic context to semantic descriptions[15]. Such designs have yet to be applied to bioacoustic analysis, but offer viable mechanisms for increasing engagement while maintaining data-quality. Although my research has not focussed on such implementations, future work involves building ground-truth for games in the absence of expert annotation.

2.5.3.2 Data-processing games

Bowser et al. designed a gamified app with badges to feed the *Budburst*[14] database which tracks plant flowering. They compare their artefact's motivational affordances between gamers and citizen scientists and suggest gamification, social features, and education as general motivators (table 2.2, pg. 53)[25]. Statistically significant results showed that 'learning', 'community involvement', and 'socialisation with other participants' motivated both groups, while 'interest in plants', 'contributing to science', and 'contributing to public good' did not[ibid]. While their gamified design saw personal interest and intrinsic reward motivating both groups, gamers require designs which support play and community as personal-development and norm-oriented motives do not motivate them.

Building on SDT and motivational arcs, Tinati et al. examined player motivation in *eyewire*, a game with points, badges, leaderboards, and social competitions[229]. They posit an analysis framework which articulates 18 types of motivation, grouped into intrinsic or extrinsic and coded into 4 categories: desire to contribute; desire to learn; desire to be part of a community; and desire to be challenged or entertained, or to play. The long tail

[14] https://budburst.org

was significant in their game with 1% of players producing over 50% of output; while the authors contend that incorporating game elements and defining granular tasks enhanced participant retention, it was insufficient.

Prestopnik et al., comparing gamified (targeting achievers) and diegetic (targeting explorers) games, found data-quality stable between versions and dabblers' data sufficiently accurate[175–179]. While they posit that diegetic rewards are intrinsic to meaningful games, learning did not affect data-quality; cheating occurred in their diegetic game, diminishing data-quality. As in other projects, they report a long tail of participants and high user attrition; while interest in science and nature correlated to data-quality, interest in games did not.

2.5.3.3 Data-collection games

Han et al., describing project Budburst, posit that gamification can motivate participants with little or no prior interest in the underlying science, thus increasing the participant cohort[99]; however, their results were inconclusive. Bowser et al. describe *floracaching*, a gamified geo-caching extension to the Budburst database[15], designed using PLACE, an iterative co-design approach to Prototyping Location, Activities, and Collective Experience[26]. Comparing motivations of gamers and nature enthusiasts, they reported that gamer-participants desired more guidance and feedback than nature-participants, concluding that designing a single interface for multiple user archetypes should support task decomposition. Marti et al. designed two gamified mobile apps for gathering noise pollution data, the first encouraged users to take noise measurements *in situ*, the second rewarded users for repeating other players' geo-located measurements[146]; transient avian bioacoustic signals are poorly suited to geo-caching games. Xue et al. describe *avicaching*, a game extension for reducing bias in data-collection in the eBird database[247] which incentivised participants to visit under-sampled areas; a 19% shift in surveys towards under-sampled areas resulted from higher point incentives. Point incentives were tied to the leaderboard-based competition structure already present in their interface; unfortunately, only high-level participants chose to be involved: those who participated in the extension already produced 64% of total surveys for the region. Regardless of mechanisms for incentivising data-collection, data-representation must either be familiar to the participants or readily comprehensible.

[15]Geo-caching is a game format in which users are required to perform an action at a defined real-world location.

2.6 Audio interaction design & interface affordances

Bioacoustic data-collection or analysis requires interfaces supporting visual and audio interaction. Training users to recognise birdsong can be augmented with visual depictions of sound. Projects where participants identify birdsong from field recordings entail a source separation problem, followed by a metadata-creation problem. Motivating such tasks requires interactions which meet the requirements of initial simplicity, rapid familiarity and ensuing comfort. Computational auditory scene analysis (CASA) approaches exist for automating source separation, but human results remain consistently better than machines'[39, 71]. Most citizen scientists have no background in, or desire to learn about, properties of sound[48], yet participation in data-analysis tasks requires basic familiarity with visual representations of audio data. An HCI design approach to citizen scientist audio interaction with bioacoustic data requires considering affordances and constraints of sound-representation interfaces[165].

2.6.1 Visualising sound

Visualising complex data such as sound involves emphasising some types of information at the expense of others. Useful representations must consider human perception and visual cognition[181, 231]. Waveforms, plotting energy over time, although familiar from music playback interfaces[16], are limited to amplitude information comprising onset, energy, and duration. Databases such as Xeno-canto[17] sometimes augment birdsong recordings with spectrograms, a visual representation of the amplitude at each audio frequency projected onto a two dimensional graph of frequency over time, for visualising avian bioacoustics. Spectral data-representations have an intrinsic modelling trade-off between frequency and time resolutions, depending upon windowing function and hop size selected[208]; nevertheless these are the most common representations for frequency visualisation. While spectrogram time and frequency scales are self-evident, when hop size is specified, choropleth mappings[18] for signal energy content at a given frequency and time are not standardised. My research explores various audio data-representations, and provides novel results regarding visual interpretation of choropleth maps.

2.6.2 Choropleth mapping

Ubiquity of spectrograms for bioacoustics research notwithstanding, little prior work has considered the effect of visualisation parameters on region of interest (ROI) selection in a spectrogram, despite relevant human factors. The most common non-greyscale

[16] *e.g.* https://soundcloud.com

[17] A citizen science portal for amateur and expert ornithologists. http://www.xeno-canto.org

[18] Relating colours to measurements of displayed variables.

spectrograms in the literature use a rainbow choropleth mapping[231], although spectral maps, which increase data-carrying capacity, exist[127]. Light et al. note that "[c]olor has the potential to enhance communication, but design mistakes can result in color figures that are less effective than grayscale" and further propose a heuristic to "[a]void the use of spectral schemes to represent sequential data because the spectral order of light carries no inherent magnitude"[137]. In addition "[t]he rainbow color map confuses viewers through its lack of perceptual ordering, obscures data through ... uncontrolled luminance variation, and actively misleads interpretation through the introduction of non-data-dependent gradients"[107]. Effective transitions in choropleth colour maps have been examined by the authors of the *ColorBrewer* website[19] who propose sequential[20] and divergent[21] colour schemes for conveying continuous data where thresholds matter, as with spectrogram amplitude. Gardner analysed ColorBrewer choropleth maps, preferring "sequential schemes that use more than one hue in their transition"[85] and found that diverging schemes are better suited for colour-vision-impaired users. Recent research testing the information content of choropleth mappings compared greyscale to diverging mappings posited to leverage human perception of luminance and chromaticity[157], with inconclusive results. My research investigates various spectrogram choropleth maps for tasks of bioacoustic ROI selection.

2.6.3 Interacting with metadata

In addition to acoustic data, recording metadata, such as time and location, provides valuable information. Slaney contends that metadata may be more useful than content-data[207] for model formation and search. Standardisation helps to define consistent data-quality metrics and simplifies collaboration across datasets. Fegraus et al. proposed ecological metadata language (EML)[79] to contextualise descriptions of biodiversity assessment data, whether collected by professional or citizen scientists. My data-collection software input protocols apply these standards.

2.7 Frameworks for design & analysis of citizen science games

Based on the literature reviewed previously in this chapter, I now develop two frameworks for the design and analysis of citizen science games. The first framework defines dimensions for discussing how games support engagement whilst increasing knowledge beneficial to

[19] http://www.colorbrewer2.org

[20] Dark to light, either mono- or multi-chromatic.

[21] Where shifts are conveyed with hue and lightness, in the dichromatic form, between two primary colours.

scientific stakeholders. The second framework defines how games of varying complexity motivate participation, depending upon participants' underlying interest in biodiversity monitoring and sources of motivation.

2.7.1 Game interaction framework

This game interaction framework presents dimensions for discussing how games can motivate learning requisite for participation in biodiversity assessment (fig. 2.7, pg. 57). Dimensions are *game difficulty* — encompassing game-mechanic complexity and knowledge

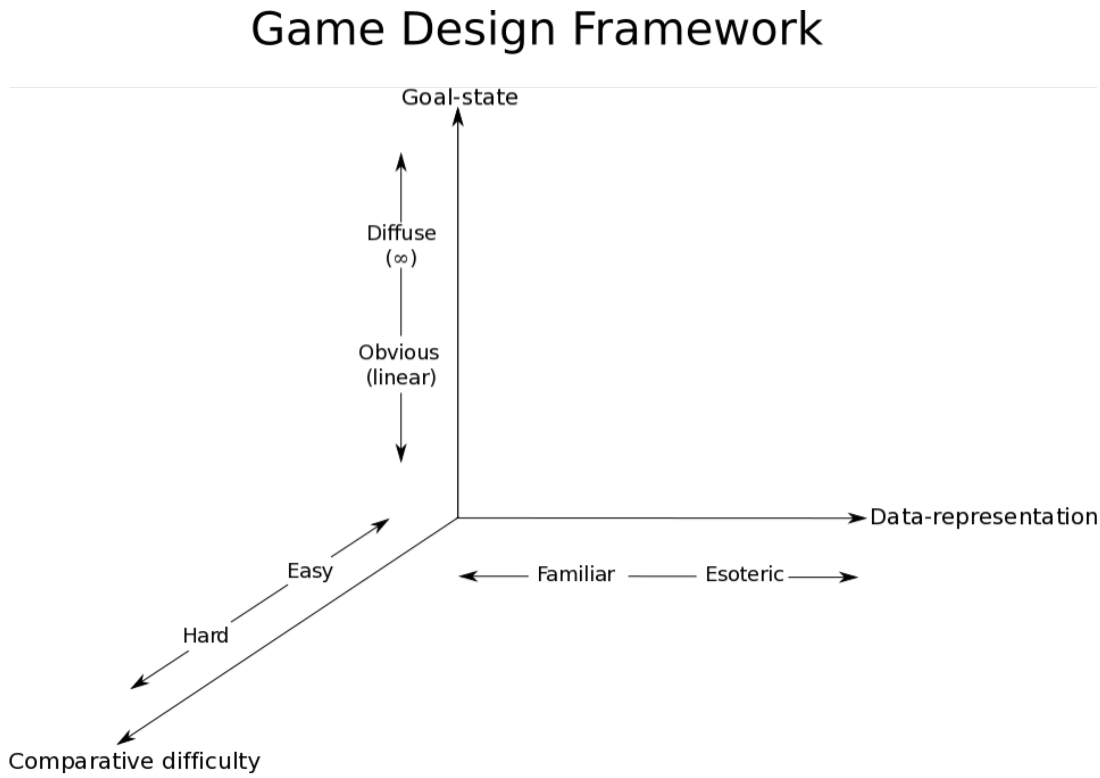

Figure 2.7: A framework comprising dimensions along which games vary, supporting multiple modes of motivation.

familiarity, *representation* — encompassing data familiarity and dimensionality reduction, and *goal-state complexity* — the degree of autonomy game-mechanics provide.

2.7.1.1 Qualifying difficulty

The difficulty dimension encourages discussion of how solving increasingly challenging tasks may benefit from gamification and how (un)familiar game-mechanics drive engagement. Increasing difficulty can involve increasing the number of granular tasks required within a game or increasing the complexity of an atomically granular task. The speed with which a series of tasks must be completed may vary non-linearly with granular task completion, further affecting difficulty. Performance rewards in this dimension are extrinsic and positive, provided regulation is introjected or identified, in the context of variable regulation of extrinsic motivation (see table §2.1, pg. 36). Rewards for progress along the difficulty dimension may be internalised if either the user is predisposed to favour control or the extrinsic motivation regulation mode enabled by game-mechanics leads to integrated rather than identified regulation. Integrated regulation occurs when users' conceptualisation of self-worth increases because rewards are given for task completion at increased difficulty which leads to congruence with project goals.

2.7.1.2 Data-representation

In designing my games for teaching acoustic information, the data-representations must enable relevant interactions with underlying data. Familiar visual audio data-representations reduce audio content dimensionality, limiting types of interactions between users and data. Exploration of prior audio data-representation familiarity is introduced in preliminary experiments in §4.2. Introducing esoteric visual representations of audio requires users to undergo a learning process prior to effective interaction; if this process is too complex, games based on such representations will inevitably lose the majority of players who, as Cottman noted[48], engage with the project for reasons unrelated to learning data-representations. On the data-representation dimension, effective interface designs must balance familiarity against completeness, providing sufficiently complex visualisations to allow users to elicit useful results without significantly reducing the participant base.

2.7.1.3 Goal-state complexity

The goal-state complexity of a game may vary in either the complexity of granular tasks leading to the goal, or the degree to which the goal-state is comprehensible and the path apparent. Simple games offer a single goal where success is the predictable result of a series of moves which can be identified as leading the user closer to the goal-state; viewed as a graph-traversal problem, each turn either directly decrements the number of turns necessary for completion or leaves the user equidistant from the goal, loops are not possible. More complex games require mechanics which allow cyclical trajectories, where moves bring the user closer to winning, further from winning, or equidistant from the target state, without guarantee of finding a terminal winning goal-state. Both such games are premised upon the existence of a terminal state. Further along this axis lie games or toys with diffuse goal-states where it is neither evident nor necessary that a terminal state exist, and users accrue benefits from engagement and output during play, rather than upon interaction completion.

2.7.2 Engagement framework

Motivating engagement through games requires understanding and predicting motivation elicited by rewards arising from varying game-complexity and identifying underlying motivation elicited by participants' desire to perceive value associated with their actions in the broader project context. I propose a framework for discussing motivating engagement (fig. 2.8, pg. 59). The game design complexity dimension is a composite of dimensions

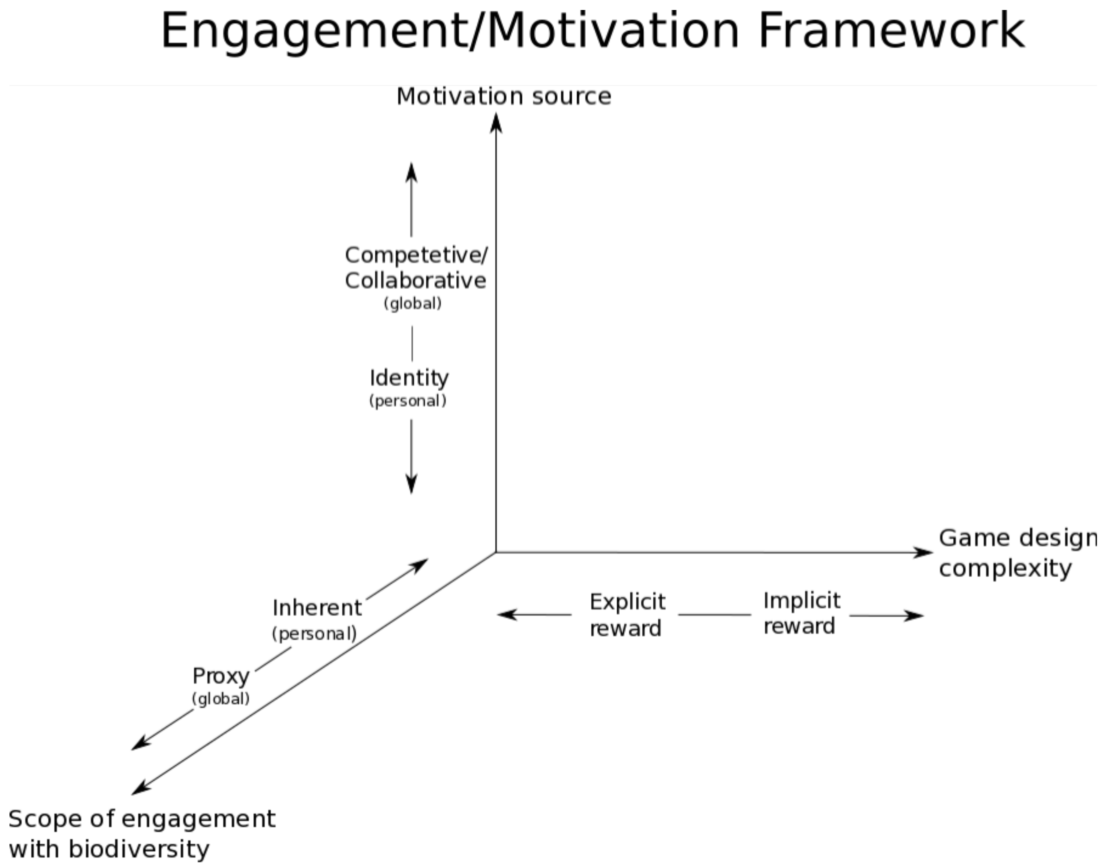

Figure 2.8: A framework for discussing interaction dimensions along which citizen science participant motivations may be considered.

from the previous framework where goal-state complexity and game difficulty combine to implicitly or explicitly reward user performance.

2.7.2.1 Identifying & supporting users' motivations

While prolonged interaction with citizen science projects may be enhanced by extrinsic motivation for continued involvement, initial decisions to participate are primarily intrinsic. This project, allied with the broader social construct of school curricula, introduces an exception, which is that projects in environments that encourage participation without intrinsic motivation may still yield quality data. Games can provide extrinsic rewards from gamified features and intrinsic rewards if users internalise and appreciate learning from the game; citizen science project design benefits from contextualising participant motivation, even absent game structures. Ongoing participant engagement varies depending, in part, upon the degree to which initial interactions are intrinsically or extrinsically motivated. Intrinsic motivations for engagement may be characterised as either identity-related — resulting from a sense of autonomous self, or socially-related — resulting from the need for collaborative and competitive reinforcement. Intrinsic motivations are not necessarily triggered by self-centred behaviour, or from self-community relationships, but may be elicited from self-environment relationships, particularly in conservation projects.

2.7.2.2 Engaging users with biodiversity

Interactions with the environment may provide participants with intrinsic motivation for engagement with biodiversity on a local, immediate, personal scale — enjoyment of nature, or by subsequent perceived global-scale benefit to the environment resulting from their engagement — contributing to conservation. If participants identify sufficiently with a project, awareness that their interactions benefit distant locales may enhance intrinsic motivation. Games designed to promote collaborative success enhance user engagement when participants are capable of perceiving the proxy value of their actions. Engagement likewise increases when interactions with a game community provide introjected regulation which moderates extrinsic motivation. Continued engagement with biodiversity projects may be discussed through the stages of the motivational arc, once sufficient motivators have been introduced to elicit an initial contribution.

CHAPTER 3

Biodiversity monitoring, bioacoustics, & citizen scientists

The previous chapter developed an analysis framework for describing the motivation of citizen scientists. This chapter explores the context of bioacoustic citizen science encompassing biodiversity monitoring, bioacoustics, and previous technical approaches to citizen science.

3.1 Biodiversity monitoring theory & practice

The oldest continuous biodiversity monitoring dataset was collected for most of its history by people who were neither professional, nor considered citizen, scientists. This dataset documents the timing of flowering of the cherry blossoms in Kyoto, Japan from the 9^{th} Century to the present[11]. Kobori et al. identify a change in perception of the purpose of phenological[1] data-collection, noting that "observations made by volunteers, [historically] thought of primarily as a cultural practice, are now used to evaluate the effects of climate change"[122]. In the context of data-collection protocols, the cherry trees are an indicator species; presence and prevalence provide data for ecological models and ring information augments environmental models. Historically, most data obtained for biodiversity monitoring indices were aggregated from data collected by hypothesis-driven scientists working on limited spatio-temporal scales; such work was slow as professionals were needed for collection and analysis. Moving from hypothesis-based biodiversity assessment to data-driven modelling necessitates increased collection of data via sensor networks of varying scope and topology. Data-driven ecological models draw on increased depth and breadth of information, providing insights unavailable from indicator-species-

[1]Phenology deals with the influence of climate on the recurrence of annual phenomena of animal and plant life such as budding and bird migrations or range shifts over time.

driven modelling.

Data acquired through static continuous sensor networks yield more consistent spatio-temporal information than do data collected by individuals, but costs increase significantly when scaling spatial coverage. Outside of biodiversity assessment, participatory mobile networks treat volunteers as participant sensors and are popular for volunteered geographical information (VGI) data-collection projects, including noise mapping and pollution monitoring. Design of experiments for biodiversity monitoring remains largely in the static sensor network domain where data-forwarding algorithms and protocol considerations are paramount[172]. Static sensor networks of acoustic autonomous recording units (ARUs), which provide most current raw bioacoustic data, are comprised of acoustic data-logger nodes which record and forward sensed data to a centralised database (fig. 3.1, pg. 62)[246]. Mobile networks (fig. 3.2, pg. 63) introduce alternate considerations:

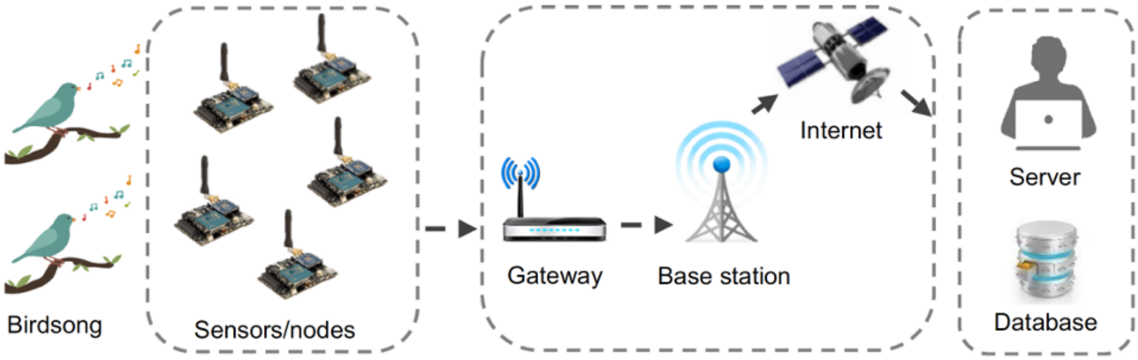

Figure 3.1: System diagram from Boulmaiz et al.: birdsong is recorded by sensors at static locations[24]. In such a network, transmission protocols are fixed.

will on-device processing occur prior to transmission to databases?; will sensing occur at personal, group, or community scale?; will users actively engage in data-collection, or will their devices determine, based on location and a signal from a centralised server, what data to collect?; and how secure must sensed data be to ensure end-user privacy?[130]. In both opportunistic and participatory mobile sensing infrastructures, resource allocation must be balanced between the need for sensing location, recording data, audio or otherwise, and end-user battery constraints.

Participatory data-collection conservation projects leverage user contributions, while biodiversity assessment projects need users for data-analysis. As end-user populations become more diverse, defining archetypal users' knowledge becomes a requisite step prior to integration of collected data into databases for analysis[131]. Haklay, characterising spatial VGI data, proposes a minimum bound over which user-generated data may be considered of equal value to professionally-generated maps, noting that the relationship between the number of contributors and data-quality is non-linear[98]. Increasing the number of participants is important for mobile collection of spatio-temporal data for

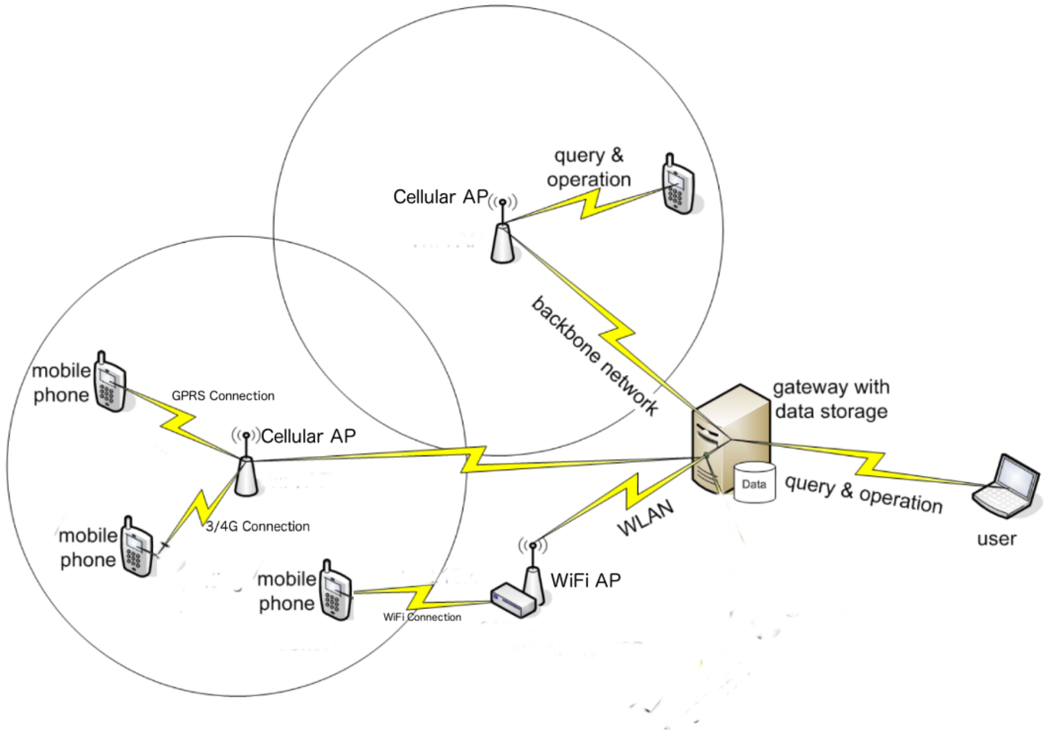

Figure 3.2: Mobile sensor network adapted from Ruan et al.: birdsong is recorded to mobile sensors and transmission protocols may be opportunistic, depending on access point and mobile device capabilities[192].

biological surveys where reproducing a timed, documented sighting is impossible.

3.1.1 Biodiversity indicators & assessment

A limited population of professionals, with limited time in the field, has limited capacity to collect data about all species present in an ecosystem. Aggregated results of global biodiversity indicators summarising changes, in the Wild Bird Index (since 1980) and the Waterbird Population Status Index (since 1985), show the majority of indicator species in decline[33]. The United Nations' 2010 biodiversity targets failed and Butchart et al. noted that "[while] indicator development has progressed substantially since the 2010 target was set, ...there are considerable gaps and heterogeneity in geographic, taxonomic, and temporal coverage of existing indicators"[ibid]. Biodiversity metrics and ecological models often derive expected presence of diverse species, given known presence of indicator species. Engaging citizen scientists to increase data- and metadata-collection simultaneously reduces data-gaps for ecosystem model development and provides a population willing to support conservation policies.

3.1.1.1 Biodiversity metrics

The input parameters to biodiversity models have often originated as aggregations of otherwise independent indices or metrics. Indices proposed to measure species diversity and abundance include general density, relative abundance, specific richness, and the Simpson and Shannon indices as general entropy metrics[62, 119]. Borges et al. applied landscape metrics to assess farmland bird declines and found Simpson's diversity index the strongest, albeit inversely, correlated metric for predicting target species' presence; increased diversity correlated with reduced single species occurrence[23]. They note that "addressing species diversity or habitat suitability requires different sets of landscape metrics", depending upon a given location's potential populations[ibid]. Depraetere et al. propose indices of "α diversity, which measures the diversity within areas, and β diversity which evaluates differences amongst areas which provide information on the turnover of specific diversity"[62]. Static passive acoustic monitoring provides the necessary data-volume to generate such summary biodiversity indices with reduced data-collection effort, but costs are significant. They posit an acoustic richness (AR) metric which, when combined with a dissimilarity index (D) between locations, can help assess α and β diversity from such data[ibid]. Riede et al., inventorying *Orthoptera* species and communities, found acoustic parameters sufficient for identifying recognisable taxonomic units (RTUs): "[b]ioacoustic diversity is a first estimate for species richness, and provides baseline data . . . prerequisite for conservation"[187]. Despite the decreased granularity of RTUs, they do provide biodiversity assessment value; such approaches have not been applied to birds. The construction of a geographical information system (GIS)-annotated avian acoustic dataset is a long-term scientific goal, however, motivating consistent spatio-temporal data acquisition is prerequisite; to this end I propose data-collection games.

3.1.1.2 Modelling ecology & biodiversity

Several forms of ecological modelling exist, from those focussing on single populations of a single species, to landscape levels, to entire ecosystems. Gontier et al. note that there are few well-developed methods for quantifying and predicting impacts of fragmentation on biodiversity and meta-populations, the set of local populations that may exchange individuals through dispersal[91]. Habitat suitability modelling (HSM) of distributions predicts occurrence of target species in fragmented partially connected habitats; meta-population models calculate population dynamics and viability of populations[226]. My research collaborators in the Nidderdale Area of Outstanding Natural Beauty (AONB) include ecologists pursuing biodiversity assessments based upon HSM.

GIS modelling tools for biodiversity assessment incorporate geo-location of recorded data extending spatio-temporal datasets for model construction. However, adding GIS parameters alone is insufficient for biodiversity assessments, HSM and otherwise, as models are predicated upon the existence of sufficient data[101]. My research investigates increasing the volume and veracity of the data needed to support analysis through HSM with training games for mobile data-collection and analysis.

3.1.2 Data-acquisition: volume, standards, & biases

Biodiversity assessment is transitioning from hypothesis to data-driven approaches. Kelling et al. contend that "a data-driven approach is necessary because of the complexity of ecological systems, particularly when viewed at large spatial and temporal scales"[118]. Conclusions from hypothesis-based experiments conducted on small spatio-temporal scales are inherently limited. Volunteer data-collection approaches, while increasing data-volume, lead to spatial bias: "[i]nterior forest[s] and other areas of low human population density are frequently under-sampled in large-scale monitoring programs"[121]. Although citizen science initiatives provide increasing data-volume, expanding spatio-temporal participant population through novel engagement mechanisms remains challenging. Geo-caching games to reduce avian bioacoustic under-sampling are complicated by the transient nature of birds.

Globally, "the accumulation rates of non-bird species occurrence records stored in the Global Biodiversity Information Facility (GBIF)[2] have not improved . . . over the past three decades"[9]. In the case of bird observations, eBird has yielded the largest single-project dataset to date, with nearly a third of a million participants contributing over 25 million surveys. These data are significantly skewed to North America and assimilate inconsistently with global databases. Transitioning from knowledge-driven to data-driven biodiversity modelling requires interdisciplinary standardisation of data-collection protocols across multiple fields for effective model formulation. The absence of inter-community standards creates fragmented, heterogeneous datasets, inhibiting data-integration into generalised models. My research output formats into ecological metadata language (EML), standardising participant output.

3.1.2.1 Collection methods

Common avian survey methods include point-counts, either by humans or acoustic ARUs[3] and transects; video monitoring techniques from marine studies are rarely used for birds. point-counts, where professional or citizen scientists document identifiable species within visible or auditory range from a single location, remain the predominant avian data-collection approach for humans and ARUs. While the eBird data-submission interface

[2] https://www.gbif.org/en/

[3] ARUs can be passive, providing constant recording, or active, triggering recording when acoustic energy exceeds a threshold.

allows for one- and two-dimensional transects, survey complexity means most professional and avocational contributors collect data from point locations, such as hides. Leach et al., comparing point-counts performed by professionals with passive acoustic monitoring, found that, for fixed duration observations, point-counts by professionals detected more species by visual recognition[134]. While increasing ARU recording duration can mitigate this, in the absence of automated processing analysis requirements increase[ibid]. My research explores training avocational collectors to exceed passive monitoring standards by qualifying trust in their active acoustic monitoring performance where audio is intentionally recorded, given prior identification-game success. My designs provide the benefits of active acoustic data-collection, where participants perform monitoring as a human intelligence task (HIT) by recording only data containing signals of interest, eliminating the annotation complexity inherent in passive monitoring systems.

3.1.2.2 Limitations of eco-informatics & computational sustainability models

Both predictive and causal models are needed for ecological science. Many ecology problems are superficially similar to previously-studied problems involving object recognition, density estimation, model fitting, and optimisation, but existing solutions are rarely directly applicable. Eco-informatics experiments, often erroneously premised on the existence of sufficient data, examine ecological subsystems or pathways using computational methods to build predictive and causal models, applying inference on sensor data instead of hypothesis-based experiments[69]. Gomes et al. describe the burgeoning interdisciplinary field of computational sustainability which explores "modelling complex species distributions and developing conservation strategies [which] require ... stochastic optimisation methods [for biodiversity preservation]"[69, 89]. Designing spatially-expansive, static, passive sensor networks to collect sufficient data presents significant monetary, infrastructure, and computational challenges.

I contend that participatory approaches to data-collection and analysis for biodiversity modelling have better scaling potential. Citizen observation networks can contribute the datasets necessary for such modelling and enable the public to engage in scientific investigation, develop problem-solving skills, and engage with policy-formulation processes. My research investigates processes for generating the data necessary for eco-informatics models by supporting citizen scientists' geo-located contributions.

3.1.2.3 Stakeholder engagement

With increasing demand for geographically-broad surveys, volunteer data-collection provides an ancillary benefit: the involvement of citizen scientists enhances public appreciation of science and conservation. However, data-quality remains contentious. Danielsen et al. found that public participation in environmental monitoring positively influences the speed and spatial scale of decision-making and resulting action[56, 57]. While implementing regional, national, and international conventions driven by scientist-executed monitoring takes 3 – 9 years, small-scale monitoring involving local stakeholders frequently produces decisions within the first year[57]. Unfortunately, data-quality suffers — in studies of four locally-based monitoring schemes, only one yielded results which correlated with professional biologists'. Locally-based monitoring can increase local empowerment for natural resource management, including the application and generation of local knowledge[ibid]. My research explores geographically-localised games, collaboratively designed with local stakeholders who contributed knowledge, to build engagement and increase result reliability through targeted learning.

3.1.2.4 Dealing with incomplete metadata

With large datasets come new challenges in data-management and analysis. Caruna et al. "develop and evaluate practical strategies for automatically identifying subsets of important features" and visualise their effects on Cornell Lab of Ornithology (CLO) citizen science datasets, such as eBird, where metadata have many attributes, and recordings are frequently noisy and missing values[35]. Their approaches support incorporation of records created by dabblers and others with incomplete knowledge. Zhang et al., failing to support data-standardisation, proposed tools supporting the utilisation of audio datasets for environmental monitoring to record and upload audio data and metadata[251]. The proposed design of their unbuilt system for data-collection, management, event-processing, and data-mining would allow queries at the spatial, temporal, annotation, and project levels, worthwhile goals. Constrained entry protocols reduce incomplete metadata; my survey designs, described in §5.4.1, iteratively evolved to reduce incomplete submissions. Labelling recorded utterances, described in §4.3.2, forced data-entry using local common names or scientific names.

3.1.2.5 Engaging scientists with citizen science data

Building scientifically sound models requires validating trust in user-generated data. Theobald et al. authored a "quantitative review of biodiversity-related citizen science to determine whether data collected by these projects can be, and are currently being, effectively used in biodiversity research"[225]. They define citizen scientists as participants in projects which collect and process quantifiable information related to a specific issue or question and consider the type of data collected, the rigour with which they are collected, and the availability of data collected for subsequent analysis. They identify lack of data-quality metrics as limiting professional use of citizen science data. This diminishes the resultant number of peer-reviewed scientific publications. They warn that "potential in citizen science will not be fully realised if citizen science data do not reach the peer-reviewed scientific literature ... [and] if biodiversity science does not engage nonscientists"[ibid]. My constrained entry protocols mitigate the risks associated with trusting amateurs' data, while my interaction designs incorporate mechanisms for building trust in user-generated data through games.

3.1.2.6 Designing for trust

One way to build professional trust in citizen science datasets is to support amateur and professional domain experts' collaborative design of biodiversity assessment protocols. Cottman et al. note collaborative protocol design processes create agency and power for end-users, while benefiting researchers by increasing subsequent trust in collected data[48]. Their research output comprises web-based tools and systems for monitoring biodiversity using acoustic ARUs; they note that "[al]though new tools and systems offer the power to capture more data, human collaboration, analysis and stewardship are required to extract useful information"[48]. Most citizen scientists have minimal prior acoustics knowledge; however they have collaborative and competitive tendencies for sharing new discoveries and building species lists, both of which I consider in my design research.

Games are useful tools for developing trust metrics associated with user-generated data. While digital technologies enable citizen scientists to collect, analyse, and disseminate data, the design of technologies for citizen science projects and communities has rarely benefited from human computer interaction (HCI) theory[173]. Smartphone interactions offer opportunities for HCI practitioners to incorporate game elements which integrate data from commonly available sensors into interaction designs for citizen science. However, incomplete knowledge, resulting from the *'don't know'* trap, introduced in §2.2.1.2, creates tension between adherence to data-collection protocols and motivations for participation in scientific research. Interaction modalities such as games, which increase trust associated with collected data, can concurrently develop requisite user knowledge and increase engagement in conservation projects.

3.1.3 Case studies: prior work

Numerous studies of flying vocalising species, primarily avian, exist for population and ecosystem modelling. Here I critique the scalability of collection protocols, identify the value in developing new methods of data-acquisition, and discuss how HCI theory guides my designs.

3.1.3.1 Automated acoustic monitoring

Several researchers have leveraged ARUs for acoustic habitat-monitoring sensor networks for target classification[171, 223, 239]. Some have attempted spectral cross-correlation for automation, while others have simply collected recordings for post-hoc human identification and labelling. Automated monitoring, while useful for collection, creates data-sets which require human annotation. Although automated acoustic analysis techniques offer some tools, building human-in-the-loop (HitL) annotation interfaces benefits from HCI theory.

In acoustically-rich environments Taylor et. al.'s classification software failed, so they proposed an online reference allowing project participants to "update the underlying taxonomic framework to cope appropriately where the taxonomy is unstable/incomplete"[223]. While Wang et. al were capable of real-time event-driven processing for a small corpus of vocalisations by a single target species, scaling to multi-target classification and localisation failed; therefore autonomous biodiversity scanning was infeasible[239]. These results support my characterisation of data-analysis for model development as a HIT which can benefit from the contributions of a cohort of engaged citizen scientists regardless of data source; however, none of these systems was functional past a small set of target species. I therefore avoid automated monitoring, proposing instead that a cohort of engaged data-analysis participants may concurrently function as a source for data-collection.

3.1.3.2 Population estimation

If the primary goal is the constrained task of estimating a single species population from a single defined utterance, acoustic analysis can reduce the need for human validation. In a controlled environment Terry et al. built a census system using neural networks and back-propagation, concluding that "neural networks were able accurately to count and re-identify individuals within populations that varied in size and composition"[224]. This did not scale to the wild as a single-species environment was prerequisite[ibid]. Fisher et al. compared permanent point-counts, randomised point-counts, and stationary acoustic transects and introduced a mobile acoustic transect method for counting populations that employ an aggregation response such as flocking[80]. Using mobile transects the authors identified twice the population of stationary counts because flocks, where spatial abundance density is highly variable, are difficult to survey and monitor with methods such as point-counts and ARUsibid. The comparative efficacy of population estimation from mobile recordings guides my research, but building transect approaches into games remains future work.

3.1.3.3 Community estimation

Single species population estimates provide a single feature for biodiversity modelling; community estimates better support data-driven approaches to biodiversity assessment. Researchers comparing point-counts with ARUs for community estimation found that detection, whether automated or by humans, declines with distance when data are collected from a point source[82, 121]. When spatial biases are inherent to the data-collection method, this is reflected in derived models. Klingbeil et al. concluded that "both [point-counts and ARUs] provide similar estimates of species richness and composition"[121]. Given few human experts capable of performing such surveys, they proposed that acoustic ARUs may supplement or replace point-counts at scale, cost notwithstanding[ibid]. Acoustic ARUs provide cleaner baseline data than citizen scientists, given human error and sampling bias, for population modelling in regions where the infrastructure requirements are feasible. However, given difficulty scaling ARU infrastructure, I propose gamification as a means to build expert human capacity.

3.1.3.4 Habitat suitability estimation

My fieldwork, described in chapter 5, involves collaboration with biologists developing a regional HSM for which they lacked sufficient input data. Applying computational auditory scene analysis (CASA) methods to geo-located recordings may provide a basis for acoustic mapping and survey techniques with reduced human input[14, 39]. Ultimately, HSM requires human-contributed data-analysis in the absence of significant increases in automated classification capacity. Bioacoustic monitoring is suitable for habitat mapping, although source segmentation in low signal-to-noise ratio (SNR) environments remains primarily a HIT. Bardeli et al. estimate noise in frequency bands known not to contain bird vocalisations, for noise reduction in an algorithm designed to detect vocalisations; their system output produced both significant false-positive and false-negative elisions[14]. Identifying a baseline noise floor as a HIT contribution to acoustic habitat mapping for biodiversity assessment is introduced in my experiments in §4.2.2.1 and developed in §4.3.1.3. My work focusses on the design of interfaces and interactions which incorporate bioacoustic dimensionality-reduction and efficient data-representation techniques to build knowledge capacity and enhance trust in amateur data.

3.2 Bioacoustic signal processing

Bioacoustic content from ARUs and human-collected recordings, while a significant source of survey data, must be processed and analysed prior to inclusion in biodiversity models. This process involves a combination of human expertise and automated signal processing. Many approaches to bioacoustic signal processing find their origins in automated speech recognition (ASR) methods which better resolve features tuned to human speech than to animal vocalisations. Automatic call recognition (ACR), applying ASR techniques modified for avian vocalisations, aids processing of bioacoustic data but results do not yet scale to replace human analysis. Automated bioacoustic analysis is split into acoustic preprocessing and subsequent inference of the number of classes present in the signal; a class can represent a call type, an individual of a species, a species within a community, or a less precise soundscape feature as a parameter to a biodiversity model.

3.2.1 Processing bioacoustic signals

Bioacoustic signal analysis, regardless of source, by either human or machine, begins with preprocessing for removal or reduction of the noise floor and non-target signals. This is followed by partitioning target signals for subsequent analysis and may include signal transforms or extraction of summary statistics.

3.2.1.1 Noise reduction & endpoint detection

Whether identifying an individual in a population or a species in a community, preprocessing includes noise floor removal and possible reduction to overlapping non-target signals, as necessary. A noise floor can be eliminated through high-pass filtering at an arbitrary threshold. However, Gage et. al.'s assumption that biophonies, sounds emitted by birds or other animals, occur above 2kHz, while anthropophonies, sounds of humans and machines, are generally below 2kHz, removes potential birds of interest, bitterns for instance[83]. Band-pass filters tuned for *a priori* expected target species' vocalisations provide more efficient noise removal, and spectro-temporal box filters (STBFs) such as the 2-D Gabor filter may be used to clarify further regions of interest (ROIs) in a signal, albeit with greater computational demands. While filtering reduces signal complexity in the frequency domain, endpoint detection, to the level of note, syllable, phrase, call, or song, is less computationally intensive in the time domain, where energy thresholds generally suffice.

3.2.1.2 Extracting audio features: signal transforms

Time domain representation of an audio recording is typically graphed as signal energy over time. Such depictions hide signal frequency content, but in high SNR environments they provide sufficient information for endpoint detection. In order to derive a signal's frequency information, a Fourier transform (FT) is used to calculate frequency content for windows in the time dimension (fig. 3.3, pg. 72). Suppose that $x = [x_0, \ldots, x_{N-1}]$ is an N dimensional complex vector representing input audio. Let $\omega = \exp(-2\pi\frac{i}{N})$. Then the discrete Fourier transform (DFT), $c = \mathcal{F}_N(\S)$ is given by:

$$c_k = \frac{i}{N} \sum_{j=0}^{j=N-1} x_j \omega^{jk}. \tag{3.1}$$

While information is lost in the transform because of the trade-off between frequency and time resolution, fast Fourier transform (FFT) implementations are commonly used to obtain frequency information prior to filtering in bioacoustic signal processing. FFT output is a spectrogram with x-axis representing time while y-axis maps frequency bins. Spectrograms augment birdsong recordings in some modern databases[4]; however historic

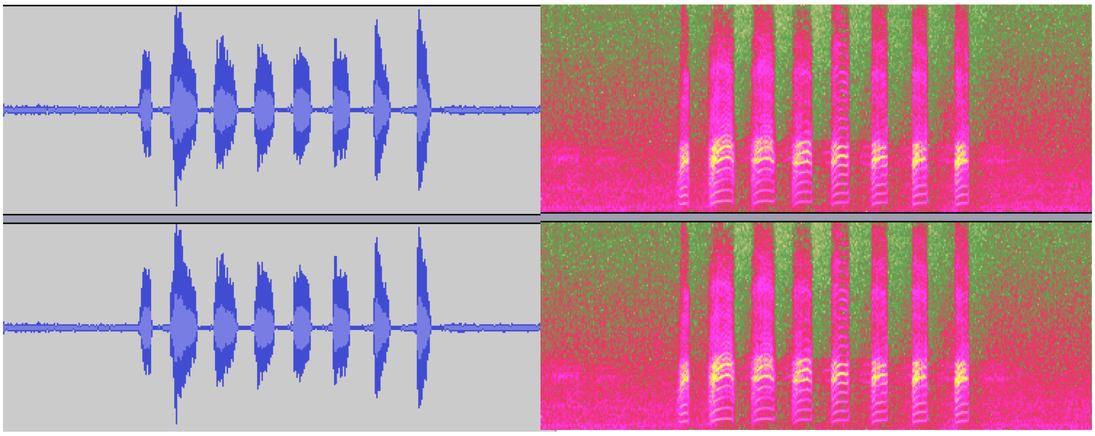

Figure 3.3: Post-noise-reduction energy (left) and spectrogram (right) plots of the laughing call of a female mallard, *Anas platyrhyncos*. While syllable endpoints are clear in the energy plot, frequency information is only present in the spectrogram.

larger databases such as the Macaulay Library's collection at the CLO focus primarily on audio recordings[5]. Spectrograms can be represented consistently on a computer or mobile screen but choropleth energy mappings are not standardised. I will investigate this in chapter 4.

Another transform applied to bioacoustic signals derived from ASR research are Mel-frequency cepstral coefficients (MFCCs), a set of coefficients derived by taking the discrete cosine transform (DCT) of the frequency output of a DFT represented on the Mel scale for frequencies. MFCCs were designed for parameterisation of human speech and scale poorly to the higher frequencies of many avian utterances. Early research comparing the efficacy of sinusoidal models, descriptive features, and MFCCs from syllables as classifier input vectors found the best classification of multiple species resulted from an MFCC trajectory model[209]. Aligning MFCC coefficients from a syllable time series with a wavelet transform produces wavelet Mel-frequency cepstral coefficients (WMFCCs) which retain temporal information and are useful input features for the recognition of inharmonic and transient avian vocalisations. These features suffer from time dependence but, with the inclusion of shift-invariant parameters, outperform MFCCs when provided as input to equivalent classifiers[40, 201]. Due to the non-stationary nature of harmonic birdsong, FFT representations work better for short duration calls than songs; my training games, introduced in chapter 5, teach calls of <2 seconds. Wavelets and related multi-scale analyses are better suited to modelling non-stationary phenomena such as transients

[4] *e.g.* Xeno-Canto
[5] https://www.macaulaylibrary.org

and discontinuities found in songs. Chirplet transforms extend wavelets and represent a windowed sinusoid over monotonically time-varying frequency; they have been used to generate vocalisation syllable dictionaries modelled as piecewise linear approximations which perform reasonably as input to standard classification algorithms[217, 218]. However, visual representations for all but FFT output are not standardised and thus less relevant to my design research.

3.2.1.3 Additional acoustic features

In addition to the output of signal transforms, further acoustic features may be computed across an entire utterance or narrower signal windows for visualisation or as parametric input to classifiers. Nelson et al. note that a feature's central tendency relative to other species in the local acoustic environment is often necessary, although sometimes redundant, for utterance recognition[160]. Fagerlund et al. propose numerous features, including signal bandwidth, spectral centroid, flux, and frequency range for parameterising bioacoustic signals[77]. My experiment in §4.2.2.2 compares a visual depiction combining several features, including centroid[6] and flux[7]. For calculating the spectral centroid, x(n) represents the weighted frequency value, or magnitude, of the n^{th} bin, and f(n) represents the centre frequency of that bin.

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{3.2}$$

Spectral flux, how rapidly the signal's power spectrum changes, is a common measure for timbre[8]. In my research I implement the L_2 norm for the normalisation coefficients μ_k and μ_{k-1} as follows:

$$S_{flux}^{(k)} = \sum_{m=m_l}^{m_u} \left(\frac{X^k(m)}{\mu_k} - \frac{X^{k-1}(m)}{\mu_k - 1} \right)^2 \tag{3.3}$$

$$\mu_k^{L_2} = \sqrt{\sum_{m=m_l}^{M_u} (X^{(k)}(m))^2}. \tag{3.4}$$

Others have explored computationally expensive approaches, including increasing the number of filter banks, but have found that optimal time resolution remains species-dependent[93]. Stowell et al. posit that "[frequency modulation] data encodes aspects of ecologically pertinent information such as species identity ... [and that a] simple technique based on spectrogram data is sufficient to extract information pertinent to species"[215, 219]. Therefore, the crux of my research leverages FFT output for training

[6] The spectrum centre of mass, perceptually associated with sound brightness.
[7] Measure of the rate of change in signal spectrum.
[8] Flux equations adapted from a preprint of a book by Eyben[76].

avocational participants in utterance recognition, as efficient computation permits real-time data-representation on low-powered mobile devices.

3.2.2 Vocalisation classes: calls & songs

Avian vocalisations fall into two categories: songs are frequently harmonic, melodic, complicated, and varied, while calls, used for contact and alerts, are generally monotonous, brief, and of fixed melody[40]. While hearing songs may provide aesthetic reward and corresponding intrinsic motivation for public engagement, calls are better suited to training avocational participants in species recognition, due to their consistency across individuals and limited complexity. Many species have a varied call vocabulary, including flight calls, threat response calls, nesting calls, and calls to young. Most of these are short or repetitive, simplifying selection for human recognition training. Calls generally lack phrasing and contain comparatively more inharmonic content than songs as they lack phrase level harmonic progression. Despite similar vocal production bio-mechanics in birds and humans[227], identifying an avian individual in a noisy population or community remains challenging for humans, who nonetheless outperform machines. Avian vocalisations are produced by an excitation source shaped by resonances of the trachea, larynx, mouth and beak; compared to human vocalisations, avian output results in more widely spaced spectral partials and more narrowly spaced resonances[38]. Songs are comprised of phrases, syllables, and notes[41] and sinusoidal modelling, using harmonic structures to classify call syllables, has been shown to improve ACR, albeit not to human levels[100].

3.2.3 Labelling & classifying signals

Preprocessing reduces data-volume whilst retaining information pertinent to recognition of an individual, call type, or species from a bioacoustic signal. While in many instances labelling and classifying remain HITs, statistical approaches have had constrained success and play a role in future designs; my games are designed to support a HitL model for bioacoustic analysis. The BirdCLEF automated species recognition competition, evolved from the 2013 ICML Bird Challenge[9], launched in 2014 as a competition for measuring success of statistical approaches to avian species labelling. It has continued annually, now evaluating systems designed to classify 501 bird species from Xeno-Canto recordings, roughly 5% of species worldwide. Winning results to date were mean average precision (MAP) 0.453 on a set with background species possible, and 0.511 with background species removed, both far below the capabilities of domain experts and many amateurs in the field[88, 112]. These results are insufficient to support purely data-driven approaches to

[9]Attempts were made to classify 35 species. https://www.kaggle.com/c/the-icml-2013-bird-challenge

identifying species in a soundscape, however, automated approaches remain popular and warrant introduction as some statistical approaches simplify HitL computations.

3.2.3.1 Dictionary-based template-matching approaches

Template-matching approaches to avian vocalisation classification involve matching a reference spectrogram across a continuous recording. However, these approaches are constrained by incomplete vocalisation dictionaries and are impractical for improvisational songs. Implemented classification systems are limited to constrained sets of species and potential calls[10, 123, 219, 239]. Template models are promoted by researchers at the CLO in their X-BAT[10] software package. In high SNR environments the exemplar dynamic time warping (DTW) algorithm, which calculates the cross-correlation of a recording's spectrogram selection against a known corpus of spectrogram templates warped with non-uniform time axis dilation and compression, performs satisfactorily for single species identification[10]. Applying DTW to spectro-temporal acoustic features for matching to a syllable dictionary succeeded in a laboratory environment[38] but required humans for final analysis[87]. Applying a DTW kernel to a support vector machine (SVM) for classifying 5 single species' call types yielded positive results, but extending the dataset to 45 calls across multiple species failed to scale[55]. The *matching pursuit (MP)* algorithm, applied by Stowell et al. to find a sparse spectrogram representation as a sum of templates, yielded the worst results in a comparative study of automated systems, none of which matched human performance[216, 219]. No template-matching systems yet implemented have scaled to achieve results comparable to experts' identification; in experiments and preliminary fieldwork introduced in the following chapter, ROI selection remains a HIT.

3.2.3.2 Limitations of multi-class/multi-label models

Given the limitations of template-matching approaches, I will discuss other statistical methods, although none replaces well-trained avocational citizen scientists or professionals. With more complete ground-truth and increased computational power, such techniques can augment future HitL computational models for species identification and population estimation. To categorise environmental sounds as deliberate communication or unintentional noise, machine-learning approaches where call recognition is a classification task are "not appropriate for many real-world applications [because] there are many cases where obtaining tagged training instances is too expensive or simply not possible"[150, 239]. My research explores interaction designs for increasing the volume of labelled data necessary for such approaches. Nevertheless, solving avian recording labelling and classification problems has been attempted using statistical methods, including neural networks, singleton-type

[10]https://code.google.com/archive/p/xbat-devel/

recurrent neural fuzzy networks (SRNFNs), SVMs, principal component analysis (PCA), glsLDA, quadratic discriminant analysis (QDA), decision tree (DT), and hidden Markov models (HMMs)[1, 6, 30, 113, 129, 139, 152, 236]. Few datasets have encompassed more than a dozen species due to the computational complexity of multi-class classification algorithms and issues building ground-truth for supervised learning. Neural network success requires significant preprocessing and network parameterisation; recurrent networks work better than feedforward networks as the input dimensions are more tractable. Springer et al. compared two classifiers, a radial basis function support vector machine (RBF-SVM) and a multi-layer perceptron (MLP), trained and tested on source-separated and non-source-separated inputs using classifier chains; they had generally weak results when classifying 2 – 4 distinct species[210]. They concluded that "the main problem to solve is not development of new classification approaches, but to develop better acoustic feature representations and source separation algorithms"[ibid]. New feature vectors showing promise in other acoustic analysis domains, such as stabilized auditory images (SAIs), have not been tested with avian signals. The open-ended nature of avian utterances limits supervised learning potential while HitL models benefit from human pattern-recognition abilities.

3.2.4 Soundscape analysis & species recognition

Soundscape analysis allows exploration of acoustic recordings without requiring pre-existing templates for dictionary matching or other *a priori* knowledge and rarely reaches species-level identification. While holistic soundscape metrics provide some measure of biodiversity, the goal of identifying multiple, potentially concurrent, species vocalisations in recordings has led to methods yielding varying degrees of scalability and levels of success.

3.2.4.1 Soundscape source separation

Gage et al.'s acoustic habitat quality index (AHQI) applies soundscape analysis to biodiversity assessment without need for species-level recognition[83]. Bioacoustic surveys, whether automated or performed by citizen or professional scientists, have previously been performed as part of a rapid assessment program (RAP), such as rapid biodiversity assessments (RBAs) and, as described in §3.1.3, can be designed to evaluate regional biodiversity, provide a foundation for estimates of occupancy modelling, or be indicators of environmental change[29, 83]. Mason et al. propose a hybrid annotation model, leveraging citizen scientists for HITs, where a community of listeners contributes to a database that is initially segmented using automated classifiers[150]. I contend that this is a more sound approach as statistical annotation techniques are yet to match human performance and incorporating end-users in data-analysis provides ancillary social good for conservation.

3.2.4.2 Species identification within a community

An approach described and patented by Agranat, despite mimicking prior art in academic research models [10, 38, 123, 239] for automated acoustic biodiversity monitoring (fig. 3.4, pg. 77), partitions species identification within a community into recording, conversion via FFT to a spectrogram, filtering, signal analysis, and classification via a comparison engine[3]. Salamon et al. investigated implementing this model for automatic classification

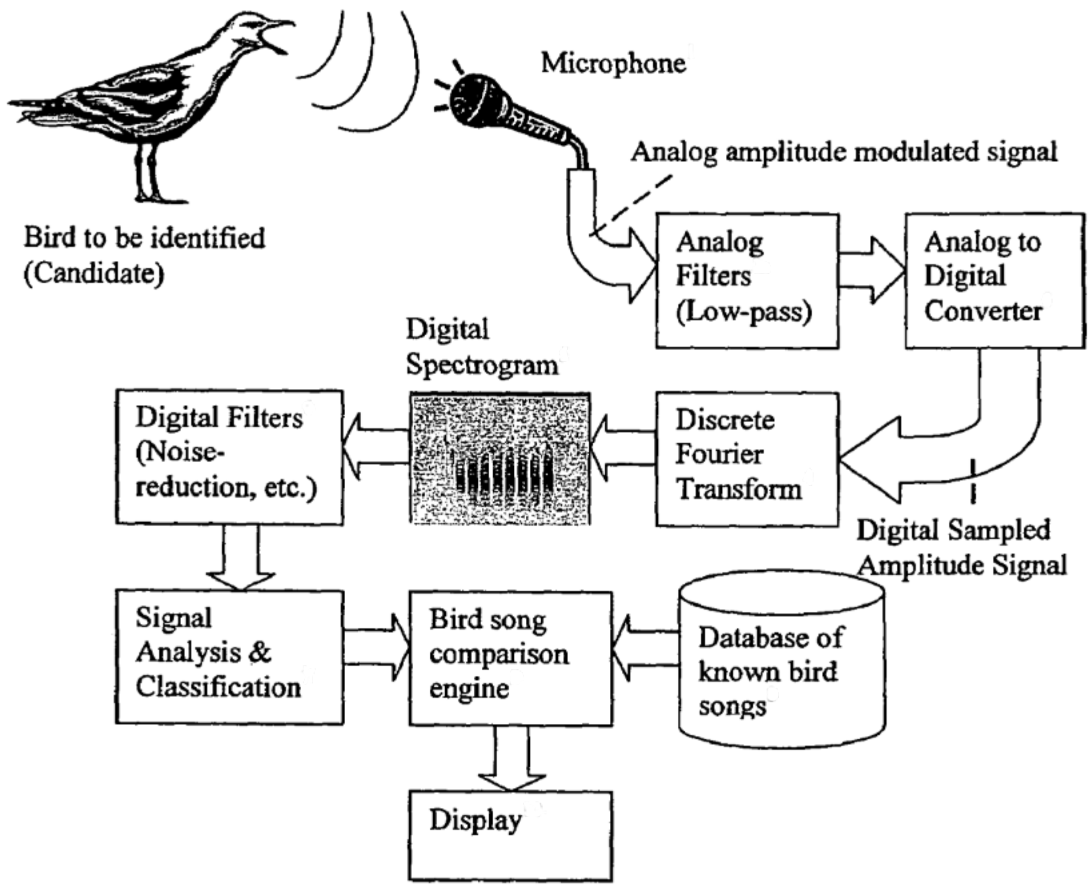

Figure 3.4: System diagram for classifying birdsong from Agranat's Patent[3]. While such designs remain common, database limitations as well as signal analysis issues remain.

of flight calls in two cases: (1) classifying a short clip containing one of a fixed set of known species (N-class problem) and (2) continuous monitoring. They achieved high model accuracy in case (1); this did not translate to case (2), as "the model is confounded by varying background noise conditions and previously unseen [sic] vocalizations"[197]. Comparison engine classification models necessitate dictionaries of all possible target species' vocalisations. For individual species, dictionaries may suffice, for biodiversity monitoring use such models fail; further discussion of these failures follows.

3.2.5 Human-in-the-loop systems

Neither statistical labelling and classification approaches, nor soundscape analysis, in themselves, provide sufficient insight to build biodiversity models from bioacoustic recordings. HitL approaches to data analysis are the most viable current methods for species identification in real-world recordings, making standardisation of metadata, such as time, temperature, and the tools used to make a recording, imperative[188]. A system covering data upload, storage, recording playback and visualisation, and automated generation of metadata using EML has been used in a number of eco-informatics analyses[142, 154, 155]. As previously noted, my research output integrates with EML databases. Development of decision support tools to bridge the gap between structured metadata production and human approaches to data organisation can increase the capabilities of automated systems[230]. Truskinger et al. note that "automatic recognition is currently intractable and manual recognition is slow and error prone ... [t]he respective strengths of human and computer analysis [can] complement one another"[ibid]. They propose recommender tools which eliminate the need for human memorisation of a large corpus of vocalisations[ibid]. Several researchers have developed acoustic data interaction tools which support user HitL annotation for analyses of bioacoustic data while running feature extraction and machine-learning algorithms[161, 233, 245]. Such systems are still nascent and will benefit from advances in systematising metadata, allowing interoperability and data exchange. My research develops tools and methods that treat avian bioacoustic data analysis for biodiversity modelling as a HitL computational problem, with particular consideration of motivating contributions through citizen science.

3.3 Citizen science

> "Traditional citizen science or volunteer programs have resulted in some of the longest ecological temporal datasets that we can access, particularly in the field of ornithology ... exploiting [current technologies], the quality, geographical range and quantity of data collected [increase]"[36].

Conceptually, citizen science may overlap with or wholly encompass other academic terms including participatory sensing, crowdsourcing, and technologically-mediated data-collection. Applying Irwin's definition, "a science which assists the need and concerns of citizens ... a form of science developed and enacted by citizens themselves"[108, 128], I discuss theoretical and practical approaches to citizen science. Designing for citizen science requires consideration of the goals enabled, stakeholders engaged, and methods implemented to realise projects, while considering the perception of citizen science by the broader academic community and the community of policy-formulators who leverage output data. I examine ethical considerations inherent to citizen science projects, including ownership of collected data, privacy needs of participants, and projects' responsibility for crediting work done. I conclude by summarising current research involving citizen science for environmental sensing and biological habitat monitoring, bioacoustic and otherwise.

3.3.1 Theory & practice

Citizen science describes a disparate set of activities involving data-collection, analysis, and public engagement. Haklay distinguishes amongst volunteer computing — donating spare computing cycles, volunteer thinking — performing HitL analysis computations, and participatory sensing — including mobile data-collection tasks[97]. Kullenberg et al. identify the dominant topic for citizen science as conservation biology, where participants collect or classify data; they note that:

> "descriptions of this purportedly new approach of science are often heard in connection with large datasets and the possibilities of mobilising crowds ... [alternatively] citizen science is a way of democratising science, aiding concerned communities in creating data to influence policy and as a way of promoting political decision processes involving environment and health"[128].

They further comment that "in the social sciences, there has been a long tradition of engaging closely with citizens as *objects* of study, especially in survey-oriented research, but this does not count as citizen science since there is no active participation or contribution from the citizens as research *subjects*"[ibid]. Biodiversity citizen science projects pursue scientific research which can be categorised into a typology of: action, conservation, investigation, virtual, and educational[244]. I implement technological interfaces which motivate project engagement and increase trust associated with data collected or classified by combining participatory sensing and volunteer thinking in a project which investigates virtual education for motivating conservation action.

3.3.1.1 Roles for technology

Citizen science projects such as the Christmas Bird Count (CBC), which the National Audubon Society (NAS) has been running since 1900, predate modern technologies. Data-collection protocols were historically premised upon expectation of handwritten documentation, with commensurate limitations. Preece et al. note that while "[c]itizen scientists may participate in all aspects of the scientific method, from problem formation, to data-collection, to analysis, and dissemination of the research results", interface usability issues hinder sustained participation[174]. Only 53% of projects they surveyed used websites for primary data-submission and only 11% had mobile applications[ibid]. Mobile applications leveraging sensor technologies combined with gameful interaction designs can streamline data-collection, improve data-management, automate quality-control, and expedite communication[163]. However, current networked project implementations suffer from a lack of design research. Despite digital interfaces, many reflect historic survey methodology limitations rather than leveraging the affordances of networked mobile devices to reduce validation time and increase engagement with data-collection processes. The PLACE approach, proposed by Preece et al. identifies key factors for designing location-based citizen science apps: Project Location, Activities, Collective experience, and Experience over time[174]. My research involves collaborative interface design to increase end-user engagement by minimising mismatches between learning or collection activities and experiential expectations from collaboration with other users and project organisers.

3.3.1.2 Participatory sensing

Participatory sensing describes a grassroots paradigm for end-user data-collection which increases motivation and enhances responses by local stakeholders. Community sensing involves the formation of communities of knowledge around sensed data[32, 57, 125]. When participation scope extends to education environments, as in my research, engagement increases with the confluence of intrinsic and extrinsic motivators. Mitigating ancillary risk of loss of stakeholder privacy must be considered during interface development as a design satisfier. Krause et al. "assert that a sensing application should weight its information need based on the expected demand for information by multiple [stakeholders]" when considering privacy ramifications related to the ultimate uses of data[125]. They describe a model for enhancing data-sharing acceptance, when collecting observations from privately held sensors, which integrates sharing preferences with probabilistic models to determine the value of probing different sensors[ibid]. Luzcak et al. found that participation in community-building features increased user output — the 1% of their users who engaged socially produced over one third of task output — and explored factors affecting user engagement, including adoption and transferability of expertise and specialised terminology[140]. My data-collection fieldwork investigates knowledge transfer from conservation practitioners to amateur ornithologists through citizen science games, while maintaining requisite privacy for educational environments. My designs treat initial trust in end-users as equivalent, and knowledge development during volunteer game-play builds end-user trust metrics. User-collected data is anonymised for privacy. My software framework allows personalisation through preference panes which can be extended to permit users to select their privacy levels, particularly of location data, within collection protocol requirements.

3.3.1.3 Volunteer thinking & protocol formulation

Attracting volunteer thinkers to a citizen science project requires interface tools, technological or otherwise, which enable and reward participation and thoughtful contribution. Poorly defined protocols or interfaces which fail to enforce data-submission constraints yield poor data, causing mistrust in results. Given program protocols appropriate for the participant audience, concisely phrased questions, and sufficiently clear instructions, citizen science data may not differ significantly from those collected by experts and benefit from inherent noise reduction with increased statistical power[21, 36]. Coordinating questions asked, collection methods, and standards for measuring educational and community-development goals also increases trust in results[21, 36]. Phenology studies, where protocols support entry of presence data, may be less affected by over- and under-reporting than studies with protocols for collecting behavioural data[86], as the latter presuppose more complex prior participant knowledge. While online or *in situ* data-collection training can increase *a posteriori* knowledge, encourage volunteer thinking, and reduce volunteer biases "it is recommended that citizen science records, particularly those involving [phenological] events . . . , should not be directly averaged across sites", given variable participant competence[90]. Datasets built by untrained amateurs are prone to spatial biases resulting from variable infrastructure and human population density but "where context data [are] available, modelling the intensity of individual observations can help [scientists to] understand and quantify how spatial biases affect the observed biological patterns"[86]. My interaction design research supports knowledge development for increasing trust in collected data while motivating user retention. It has involved identifying relevant stakeholders, increasing engagement through collaborative protocol formulation, incorporating local knowledge, and building a community of knowledge around content created by participants.

3.3.1.4 Barriers to acceptance

Despite a burgeoning number of projects, increasing user engagement in participatory sensing, advances in technologically-mediated data-collection, and attempts to formalise collection protocols, acceptance of citizen scientists' data suffers from poor integration into existing datasets and models. Specific criticisms include lack of attention to study design[124, 162], inconsistent or sub-optimal training[45], absent or problematic standardisation and verification methods[22, 31, 43, 68], and observer or sampling bias[220]. Burgess et al. categorise failures to capitalise on citizen science, including limited awareness amongst scientists of projects that match their needs; they contend that not all biodiversity science is well-suited to citizen science[31]. Overselling of the method leads to poor or inappropriate data, consequently reducing trust in the data of well-founded projects. Burgess et al. identify bias amongst scientists towards certain data sources, bemoan limited development of standards to mitigate inconsistency in data-quality, and "find limited evidence to suggest a relationship between citizen science projects that satisfy scientists' biases and data quality or probability of publication"[ibid]. Cooper et al. note that "quantitative assessment of the contributions of citizen science for its core purpose – *scientific research* – is lacking"[46]. Citation of citizen scientists is rare, collected data are frequently not attributed and the trust associated with contributions is not considered[46, 128]. This diminishes the value of citizen science data and the motivation of participants in projects which publish from such datasets. My research formalises processes for measuring trust in participant data through game performance, prior to attribution.

3.3.2 Citizen science project design

Citizen science projects require interaction interfaces, technological infrastructure, and data-collection protocols to allow effective data submission. Interaction interfaces hinder or enable participation and influence result quality, but insufficient HCI research has been performed to validate most designs.

3.3.2.1 System design & implementation considerations

Prestopnik et al. explore the technological scope of citizen science infrastructure for data-collection, analysis, and dissemination and identify interface features for each which can be categorised as satisfiers and motivators; the former prerequisite for participation, the latter motivating participant engagement through rewards[178]. They identify trade-offs between building assemblages, with capacity to tailor every system feature, and buying pre-existing components, which reduces time spent on architecture, increasing time for data-collection. Features determined to be satisfiers can be bought, reducing work associated with managing participants, whilst motivators are best built, enabling active participant recruitment and tailored project-specific interactions. Solving participant privacy concerns is a primary satisfier, with direct implications for users' trust in and engagement with the project. Primary goals include increasing data-quantity through participant recruitment and engagement while quantifying and controlling for data-quality issues such as observer variability, imperfect species detection, and spatio-temporal data-collection bias.

My game software framework leverages familiar interaction modalities, satisfiers, of the Android operating system while preliminary experiments explored unfamiliar interactions and interfaces. My design process involves exploring interactions for both satisfiers and motivators, including gamified interaction mechanics. Sullivan et al. elucidate the design thinking behind eBird "developing cooperative partnerships among experts . . . [including] conservation biologists, quantitative ecologists, statisticians, computer scientists, GIS and informatics specialists, application developers, and data administrators"[220]. The authors look at system design from a social/academic rather than technological perspective as they try to balance data-quantity and quality, while facilitating wide use of the data by developing and delivering data products and engaging diverse collaborators[ibid]. Expanding the range of participant activities beyond data-collection to include data-curation, synthesis and analysis, pattern visualisation, and community engagement supports these objectives. My designs explore methods of motivating citizen scientists to cross stakeholders' data-participation boundaries by supporting cooperative partnerships amongst multiple user groups in order to move beyond participants collecting data, researchers analysing data, agencies formulating policies based on data, and land managers taking direct conservation action.

3.3.2.2 Design of interfaces to interaction models

Bioacoustic citizen science data are rarely considered by HCI practitioners. With input from birding and online citizen science communities, the output of co-designed interfaces for citizen science projects can reflect the intentions of those collecting data, performing analysis, and building resultant models. Early analog CLO projects proposed practices for participant recruitment and training, combined with narrowly constrained data-collection protocols and forms which remain relevant to current interaction designs for collecting robust, reputable data[19]. Variable data-collection result quality produces output model error. Methods for motivating engagement while ensuring accurate and complete data submissions are necessary[116]. An early digital CLO study required bipartite content — a video and a survey; however, the data-collection protocol interface implementation did not enforce submission of both parts, and partial completion, while common, was useless for project scientists[ibid]. Cappadonna et al., exploring artefacts and practices of birdwatchers and online citizen scientists, tried co-designing user interfaces[34]. My iterative collaborative interaction designs reduce potential partial protocol completion by involving both professional and citizen scientists in protocol development.

3.3.2.3 Assemblage infrastructure & design frameworks

Shirk et al. propose that citizen science projects lie on a continuum of participation from *contractual*, to *contributory, collaborative, co-created,* and *collegial*[203]. Despite technological advances and the prevalence of mobile devices upon which data-collection interfaces can be implemented, infrastructure for management and sharing of data and metadata remains a constraint in authoring frameworks for developing tools[120, 122, 213]. Communication hierarchies inhibit bi-directional information flow in citizen science projects, limit assemblage usability, and risk decreasing motivation, thus reducing participant recruitment and retention[184]. Inhibitory motivational processes introduced in §2.1.1.1 have direct implications for citizen science project design. My research designs explore participant motivation in contributory projects — designed by ornithologists for participant data-contribution, collaborative projects — where participants contribute and analyse data and refine project design goals, and co-designed projects – where participants iteratively contribute to the design research process.

3.3.2.4 Modelling users

Archetypal participants can be classified by how they are incentivised and projects must be flexible in motivating engagement of diverse stakeholders. Even when interfaces provide data-collection, analysis, and dissemination tools, and assemblage infrastructure provides users' satisfiers and motivators, citizen science projects suffer from inconsistent stakeholder competence and buy-in. These problems frequently arise when devising natural resource management strategies from citizen science data[2]. Aceves et al. describe a five point framework known as STAKE, combining Sense of place, Tools and technology, Action, Knowledge and Economic benefits to categorise incentives motivating citizen scientists to become engaged stakeholders[ibid]. A model for user trust called the occupation-detection-expertise (ODE) model for quantifying novices and experts allowed Yu et al. to generate a trust metric for each participant's data based upon perceived project engagement[250]. My designs build stakeholder competence and buy-in through motivational rewards for knowledge development and collaborative design of conservation management protocols, while developing measures of user trust through play performance.

3.3.2.5 Ethical concerns: privacy & data-ownership

Networked communities of citizen scientists have well-founded concerns about exposure to data-mining, invasive tracking, and other forms of lost autonomy and privacy. Projects must address ethical issues including maintaining data integrity, controlling data-sharing and identifying intellectual property, without conflicts of interest or exploitation. Locations such as schools raise additional data-collection considerations where regulatory limitations can prevent the creation of usable datasets, even given confirmed data-quality. By collaborating with local administrators I have ensured the feasibility of disseminating anonymised collected results from my fieldwork. Educational environments pose the challenge of disentangling learning outcomes from participation outcomes and whether engagement results from exploiting a captive audience[237]. Regardless of context, subject recruitment for citizen science data-collection projects requires informed consent as issues arise of access to results — a motivator for local stakeholders, and data-ownership — when participants' contributions are subsumed into scientific research[27, 57, 185]. While mobile environmental sensing citizen science projects can match static systems' accuracy exist, tracking participants' sensed data and assimilating and analysing collaborative results introduces privacy concerns[67, 84, 115, 183].

Sensed and metadata contributions have the potential to infringe upon participants' privacy rights, including the right to manage access to voluntarily submitted personal data, as an unintended consequence of collection protocols. However, a relevant distinction exists between environment-centric scenarios involving continuous sensing and passive participants, and people-centric sensing where participants have greater control over sensed data and captured data are exploited at the community level[42]. In both instances, assemblage architecture must support user-level control over sensor data, including: granularity, spatio-temporal context, and eventual data recipients. Users require tools to determine which data points they are willing to share — satisfier-level components of assemblage design which should include attribute-based authentication techniques (pseudonymity) and access controls for stored data — while project designers must determine which data points can and cannot be compromised in terms of precision, public visibility, and data sharing. Those formulating protocols must respect this, whilst producing input mechanisms enabling sufficient granularity[ibid]. Protocol formulation has been explored from the standpoint of the theory of *contextual integrity*, comprising the dimensions of *appropriateness* — whether a sensed datum is required for the task at hand, and *distribution* — whether the sensed datum needs to be disseminated prior to initial analysis, with or without anonymisation[ibid]. My software framework includes pseudonymous login and secure databases for collecting user performance information. Results from my collaborative design approach suggest that bi-directional information flow increases participants' willingness to accept concomitant loss of privacy associated with data contribution.

3.3.3 Citizen science for ecological & bioacoustic monitoring

Raw audio and location data, given known sensors and system design, are high-quality, contingent upon low SNR environments; however, annotated metadata come with inherent uncertainty. Data from either static or mobile bioacoustic sensing can combine geo-location information with raw acoustic recording. Although system assemblages may skip some data, eliding information by design ought to be qualified and explained. Volunteer thinking models accumulate user-filtered selection and annotation data, either performed in the field during data-collection or subsequently, during data-analysis. Expanding passive mobile sensing infrastructure has linear cost, while participatory models, given motivating interaction design, scale data-production capabilities more efficiently.

3.3.3.1 Identifying conservation objectives & motivating users

Long-term monitoring objectives in conservation include management, awareness, education, building ecological knowledge, and improving methods, but sending people to survey can endanger the things being surveyed[156, 232]. I explore the educational value of games for interactions which raise awareness and increase ecological knowledge amongst citizen scientists. Successful designs for participatory mobile citizen science must compete with static passive sensor networks for biodiversity monitoring.

Moran et al. contend that despite "[p]otential advantages in terms of scale and engagement with the public, the turn to citizen science in biodiversity raises tensions in terms of the nature of the scientific endeavour and its current cultures and practices"[156]. They identify design considerations for citizen science projects, from data-representation familiarity, studied in chapter 4, to the level of involvement of the person holding the sensor. Project objectives and motivation sources notwithstanding, most citizen science projects produce participation curves where few users provide most data. However, long-tail participants may produce more than 50% of the aggregate data, so designs suitable for dabbling, introduced in §2.2.1.2, are justified. Observer quality is posited to undergo a *learner effect* as knowledge develops; I measure knowledge post-play, prior to assimilating participants' annotations into databases if individual data-trust metrics are sufficient.

3.3.3.2 Prior data-collection protocols & interfaces

Supporting dabblers suggests a multi-tiered approach to participant engagement and commensurate trust: project designs should "scaffold participation, recruiting a large number of participants to collect incidental information while funnelling a subset of highly committed volunteers into stricter, more labour intensive protocols"[68]. Protocols may be classified as cross-sectional — *e.g.* atlases where volunteers survey many species for a constrained period of time, and longitudinal — *e.g.* breeding bird surveys requiring on-going stratified monitoring of sites and long-term coordination[232]. Prestopnik et al. characterise citizen science projects based on participant motivation, considering interface functionality, usability features, and how data-collection tools offer intrinsic and extrinsic satisfaction[178]. Cottman et al. describe their design artefact, a website emulating a paper checklist — an interface which reflects the bird-watching community's familiarity with historic data-collection protocols[47]. The constrained affordances of such interfaces limit their design's potential by only providing intrinsic motivation to predisposed participants. My designs expand support from predisposed amateurs to dabblers with interfaces that support autonomy and competence, given minimal prior intrinsic motivation.

3.3.3.3 Interactions designed in prior system assemblages

Van et al. designed a system assemblage for processing a human-annotated vision dataset of birds and found that citizen scientists produced fewer annotation errors than did mechanical turks[11]; this reinforces the proposition that intrinsic motivation increases data-quality[103].

[11] https://www.mturk.com

In another study, citizen participants were found to be motivated by intrinsic satisfaction associated with physical birding, even in a virtual investigative environment[47]; however, gamified extrinsic motivation can also extend engagement. In developing eBird, Sullivan et al. built tools for data-collection which introduced gamified participant ranks based on individual contributions[221]. Recognition for individual effort creates competition, hypothesised to increase participation through extrinsic motivation[ibid]. Pantidi et al. describe a mobile application for bioacoustic data-collection and automatic cicada classification; despite primarily negative results, they contend that gamified mechanics can be effective motivators[169, 253]. Lepczyk et al. posit that interface designs which fail to offer virtual interaction with other participants suffer by limiting motivation for social involvement[135]. While designing for intrinsic and extrinsic motivation, I have taken into consideration the value of intrinsic motivation for generating quality data. In the following chapters I present design experiments and describe my software systems developed to explore how interface and data-representation familiarity affect engagement with mobile interfaces for data-collection, annotation, and knowledge development.

CHAPTER 4

Design theory & preliminary experiments

In this chapter I introduce interaction design theory and review citizen science design practice. Interaction design practices have rarely been applied to citizen science project development, although applying design research methods to project software interfaces ought to increase engagement. I identify prior citizen science project design approaches, discuss whether designs reflect application of human computer interaction (HCI) theory, and propose preliminary experiments to demarcate my designs for bioacoustic citizen science interactions. Initial experiments examine whether interactions can and should involve novel or familiar auditory features, explore representation of and interaction with audio data, and determine preferential interface characteristics for bioacoustic region of interest (ROI) selection. I examine the potential for novices who lack signal processing knowledge to conceptualise visual representations of sound on various interfaces, the potential for representation comprehension with increased acoustic dimensions, and avocational participants' capacity to interact with sound representations without prior explication of acoustic or interface dimensions. I explore how various representations overcome data lost from the underlying signal while supporting comprehension and users' ability to distinguish amongst signals of interest. Various audio data-representations, interaction affordances, and feature extraction mechanisms require different levels of on-device processing. While constraints have focussed many prior data-collection project designs, I contend that projects founded upon poor interaction design inevitably fail to engage participants, while technological constraints can be overcome.

I subsequently describe experiments on prototype mobile software for exploring user choropleth selection, choropleth mapping preference, time selection and frequency filtering. These results are applied to later designs described in Chapter 5 and allow increased data-transfer efficiency while maintaining relevant acoustic information in the user-selected signal. Results also show where developing for variable user preference is warranted in subsequent designs. The prototype mobile software artefact was adapted for preliminary field research, where its efficacy as a tool for augmenting bioacoustic cognition in nature was explored. This chapter concludes with a summary of experimental findings, including that mobile interfaces are most efficacious for engaging citizen scientists with avian bioacoustics and unfamiliar spectral representations of audio are comprehensible for novice users. While motivation to engage with these experiments was primarily intrinsic, even these un-gamified interactions elicited desire for further participation.

4.1 Interaction design theory & application to citizen science

> "[An] interaction design research contribution must constitute a significant invention. Interaction design researchers must demonstrate that they have produced a novel integration of various subject matters to address a specific situation. In doing so, an extensive literature review must be performed that situates the work and details the aspects that demonstrate how their contribution advances the current state of the art in the research community"[255].

In citizen science projects previously introduced (§3.3.2), discussions primarily revolved around practical aspects of participant recruitment, participant retention, and data-collection. With the exception of Yu et al.[250], who discussed modelling users (§3.3.2.4), few projects have applied interaction design research methods during development. Here I review citizen science design practice, describe interaction design research methods, and propose why the two should be combined.

4.1.1 HCI research & practice

As a design discipline, HCI explores interaction relationships between designed artefacts and people, necessitating participation and commitment by the designer, with interfaces towards academia and society[78]. HCI research comprises two forms of conduct: *design-oriented research* — wherein research is the area and design the means of producing new knowledge, and *research-oriented design* — wherein design is the area and research the means. Sato posits two design research goals: to understand acts of design and subjects of design[78, 199]. Cross's tripartite classification of design research encompasses *design epistemology* — how people design, *design praxiology* — design methods, techniques, and processes, and *design phenomenology* — study of artefacts that come out of the design process[50, 78]. HCI research techniques ought to provide grounding for practice, as interaction design researchers need appropriate design research questions as well as appropriate situated social context[214].

An interaction design research approach is suitable for projects where "the kind of knowledge and user experience sought is the kind that cannot be obtained if design — the bringing forth of an artefact such as a research prototype — is not a vital part of the research process"[78]. Zimmerman at al. contend that "design researchers focus on making the *right* thing while design practitioners focus on making *commercially successful* things", yet both follow similar development practices[255]. Rogers enumerates design methods from practice (scenarios, storyboards, low-tech and software prototyping, focus groups, interviews, fields studies, and questionnaires) and research (adding predictive modelling, Goals, Operators, Methods, and Selection rules (GOMS), and experiments)[189, 214]. Practitioners frequently adopt individual concepts — such as affordance, context, and situatedness — despite failing to apply research methods[214]. Research questions can be "reflective" — exploring experience of how a particular design technique is used, or "proactive" — seeking to change how a specific design technique is used. My design explorations involve reflective design-oriented research.

Zimmerman et al. contend that design researchers and practitioners both address under-constrained problems where the success benchmark is relevance instead of validity, with the caveat that "researchers must also articulate the preferred state their design attempts to achieve and provide support for why the community should consider this state to be preferred"[255]. They discuss Cross's contention that "normal works of practice [cannot] be regarded as a research contribution" to interaction design[ibid]. My research satisfies criteria for evaluating interaction design research contribution quality include that design artefacts "be novel integrations of theory, technology, user need, and context" and perspectives for evaluating an interaction design contribution encompass process, invention, relevance, and extensibility[50, 255].

Iterative design, the cyclic process of prototyping, testing, analysing, and refining work in progress, is both a process-based design methodology and a form of design research[254].Zimmerman posits that iterative design involves a blending of designer, user, creator, and player and involves an ongoing dialogue between designer, design, and testing audience[ibid]. His game design case studies involve the identification of *play values*, abstract principles a game ought to embody, noting that "[t]o design a game is to construct a set of rules. But the point of game design is not to have players experience rules — it is to have players experience *play* . . . [where r]ules and play are just game design terms for structure and experience"[ibid]. My research examines whether games are suitable tools for motivating engagement with citizen scientists and whether the results of motivated engagement provide professional scientists useful data. Prior to introducing game interactions, my initial experiments examine data-representation, the affordances of interaction modalities for bioacoustic analysis tasks, and interfaces best suited for both software development and citizen science use.

4.1.2 Citizen science design theory & practice

As citizen science projects affect scientific, individual, and socio-ecological outcomes, deliberate design can support sustainability, resilience and conservation outcomes while achieving scientific results. Project development should integrate elements of HCI design practice, including interviews, field studies, and questionnaires to guide interface design. However, rigorous development processes are missing from the literature. Bonney et al. classify citizen science projects by characterising public participation in scientific research (PPSR) activities[20]. Shirk et al. propose a theoretical spectrum for engagement during PPSR, mirrored in my participation degree dimensions described in §3.3.2.3, and a theory of deliberate design which reflects, without directly referencing, interaction design research methods[203]. They report that "in some PPSR fields of practice, design choices are guided by theories of participation, expertise, and democracy. In other traditions, project design is guided primarily by a growing body of practical knowledge, along with implicit assumptions about participation or expertise"[ibid]. The former provides a basis for project development grounded in interaction design theory; unfortunately, the latter is more prevalent and offers only observable case studies for design practice.

4.2 Problem formulation & initial experiments

Prior citizen science project development has rarely involved interaction design practitioners and many designs focus on technological feasibility. Many citizen science and bioacoustic data-collection project interfaces raise questions regarding underlying designs decisions made by project architects who lacked an interaction design research approach. Such project designs expose gaps between designers' assumptions and the interactions desired by users for making contributions. My initial experiments, regarding the affordances of various interfaces and representations, test prior project design assumptions to either validate such designs' choices or identify more suitable interfaces and interactions.

4.2.1 Study #1: platform choice, representation, & interaction

My first experiment was designed to identify salient characteristics of a portable device, either existent or capable of being built with current technology, for representation of and, potentially gameful, interaction with a visualisation of bioacoustic audio. For paper prototype interface depictions provided for the experiment, see fig. 4.1, pg. 94. This experiment explored the efficacy of prior visual representations of audio, introduced in

§2.6.1. Results guided optimal data-representation design for subsequent experiments considering devices' physical affordances and the context of citizen science data collection. Results from this experiment allow me to characterise motivations resulting from the rewards afforded by interface interactions as controlling or informational (§2.1.1.1).

4.2.1.1 Platform, representation, & interaction preferences

This experiment was designed to elicit: (1) a preference for a physical interface for display of and interaction with bioacoustic data; (2) a preference for the best visual representation allowing ROI identification and selection; (3) consideration of the design of a set of tools and interactions which could be performed with selected representations. Interaction designs were asked to reflect each interface's physical constraints, trade-offs between data content and comprehension in high-dimensionality representations, and ease of predicting interaction results on familiar audio visualisations.

Despite the prevalence of desktop data-entry interfaces in other citizen science projects, my first supposition is that the most preferred interface will be the most commonly available device of sufficient size and interaction capability, a small touchscreen tablet. My second conjecture posits that the preferred audio data-representation will be a spectrogram, as frequency information usefully characterises an ROI. My third speculation is that a touchscreen's affordances enable suitable interactions for ROI identification and selection without requiring training.

4.2.1.2 Prototype design & study evaluation procedures

The experiment prototype comprised: audio recordings and images of the spectral and waveform plots of avian utterances; images of the various classes of interfaces being compared; and drawing implements (coloured pencils) for the proposed tasks. I developed a low fidelity prototype where sample playback was performed using Audacity[1] for the first task. For the second task Audacity was used in the first instance, in which the participants were shown a waveform plot, while in the second instance Raven[2] was used for audio playback and spectrogram output. The study began with a short introduction to the experimental tools (*viz.* audio, images, and drawing tools), after which I observed and documented participants' task performance and use of experimental materials. Following the tasks, participants answered closed and open-ended questions about their actions and assumptions.

[1]http://www.audacityteam.org/home/
[2]http://www.birds.cornell.edu/brp/raven/RavenOverview.html

4.2.1.3 Prototype description & experimental procedure

The experiment proceeded as follows. Participants were given paper printouts depicting three possible interfaces: a smartwatch[3], a large smartphone[4], and a tangible user interface — in this case an image of the desktop inFORM interface[81], defined on the paper prototype for the purpose of this experiment as having a 10" diagonal dimension. Participants were then instructed to listen to a set of 4 sample avian utterances[5], duration ranged from 3 to 8 seconds, and frequency fundamental from 150Hz (XC120378) to 8kHz (XC42685); (2) For each interface participants drew a visual representation of the audio within interface constraints on an initial printout (fig. 4.1, pg. 94). Participants were next shown images

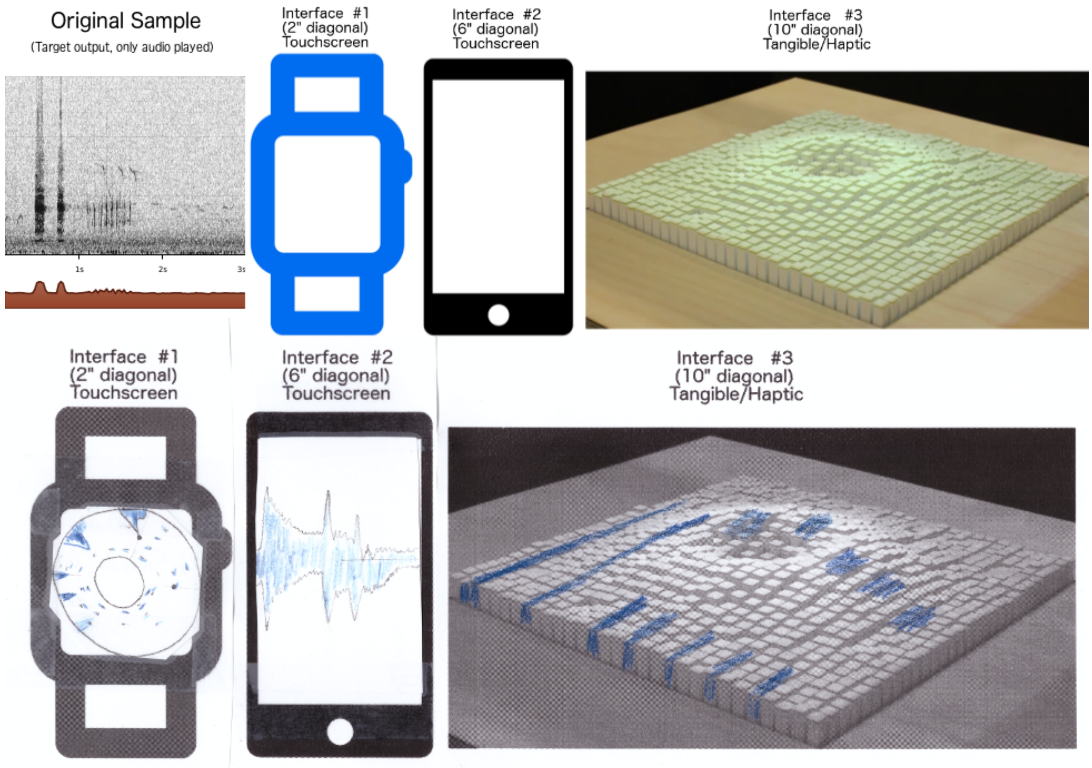

Figure 4.1: This shows the waveform and spectral output from the first avian utterance. Users listened to audio and were asked to draw how they would represent the audio (top left) on paper prototypes. Sample user output (bottom).

of spectrogram and waveform plots for each recording and instructed to draw them on two additional printouts. While depictions sometimes overlapped with output from the first task, this experimental task was designed to elicit a discussion of the relative usability of interfaces of differing sizes and dimensionalities for audio visualisations of varying complexity. Finally, participants were asked to describe tools and interactions for

[3] 2" touchscreen.
[4] 6" touchscreen.
[5] Xeno-Canto entries: XC120378, XC42685, XC66288, XC71943.

manipulating representations in time and frequency to navigate (play, seek, rewind) in the time dimension and filter (low-pass, high-pass, and bandpass) in the frequency dimension.

4.2.1.4 Data collection & participant recruitment

The experiment was administered to 8 third-year computer science students, 6 male and 2 female, who expressed an interest in HCI. It was assumed participants had prior knowledge of interface design and would be capable of interpreting the tasks. Prior audio data-visualisation familiarity was not assumed. Each session spanned 30 minutes and comprised an observational study, in which the participants' actions for each task were documented, and a set of written questions upon task completion. All participants signed consent forms, results have been anonymised.

4.2.1.5 Results

For questions asked after each task, response counts for each option are reported. Open-ended responses from each task's observational portion are summarised.

4.2.1.5.1 Interface preference

The first question, asked after participants auditioned recordings and viewed pictures of three possible interfaces, was 'which interface do you find most suitable for visualising audio?'; 6 of the participants preferred the large touchscreen interface, 2 preferred the touchscreen watch (fig. 4.2, pg. 95). For 5 of the participants the second choice of

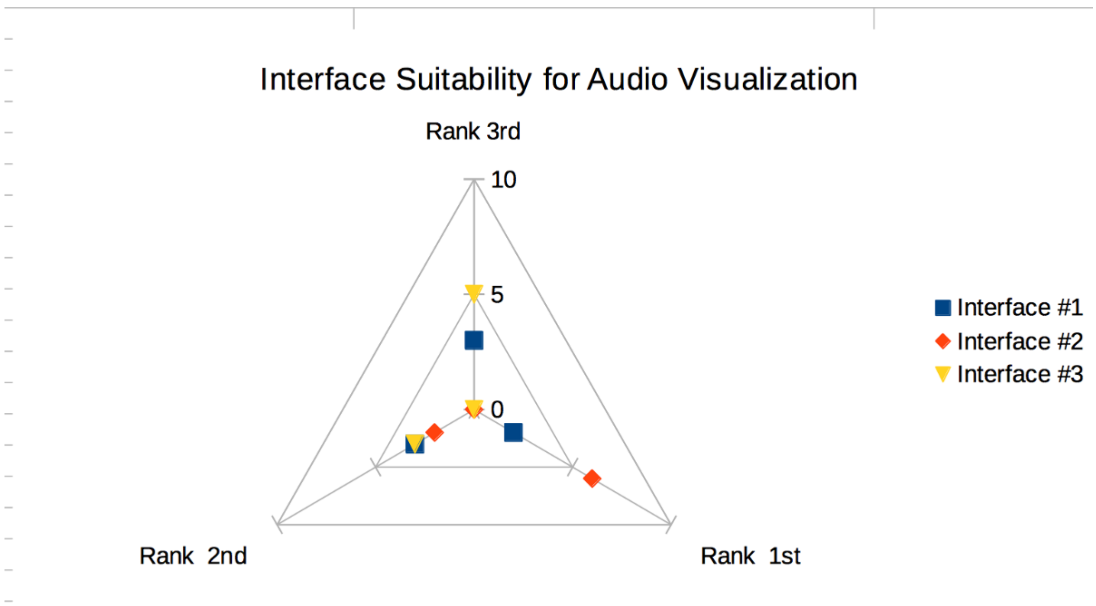

Figure 4.2: Participant interface preference; each participant was asked to rank each platform by interface suitability for visualisation of a representation of audio.

interface was the smartwatch, 3 participants stated that, if existent, a portable tangible user interface (TUI) of higher dimensionality was preferable. The interface least liked (by 5 participants) was the TUI, whilst 3 most disliked the watch, given its constrained size. All visualisations proposed were either a spectrogram (4 participants) or a waveform (4). One participant drew a waveform and posited that the drawing conveyed frequency information, but was unable to clarify when interviewed how that information was expressed.

4.2.1.5.2 Representation preference

When asked to characterise a sound representation that contained sufficient information to allow interactions, as yet unstated what these might be, 6 identified the spectrogram as containing more pertinent information, allowing more complex interactions, 2 stated that a waveform, showing energy over time, sufficed for the interactions they conceived (fig. 4.3, pg. 97). One of the two participants who preferred the waveform, although describing mapping a representation of sound energy at a given frequency bin to the TUI elevation dimension, expressed confusion when interpreting the choropleth map colour channels for spectrogram energy depicted on both touchscreen interfaces; Raven choropleth output does not include a key. Two of the participants who preferred the spectrogram on the touchscreen interfaces expressed confusion about mapping either representation to the TUI.

4.2.1.5.3 Interaction potential

The third task asked the participants to describe tools that they might implement, and interactions afforded by such, to (1) navigate in the time domain of each visualisation on each interface for the purpose of moving through a sample to find an ROI and (2) to filter noisy frequencies from an ROI on the spectrogram visualisation. Proposed tools, familiar from image processing programs, included a hand, pointer, selection rectangle, and magnifying glass. Participants proposed that selecting either the magnifying glass or the hand might enable pinch-to-zoom interactions. Both the pointer and the selection rectangle tools were proposed to enable section selection. Both the hand and the pointer were proposed to enable touch-to-drag interactions. The participants were next asked which visualisation on which interface allowed the simplest interactions to support navigating audio in frequency and time. The spectrogram representation on a large touchscreen was identified by 5 participants as best suited for both such interactions, 1 agreed for frequency selection but noted that the waveform on the watch sufficed for time selection. Both remaining participants thought the TUI was better suited to selection in the frequency domain, while one preferred the waveform on the watch and the other preferred the larger touchscreen for the time selection task.

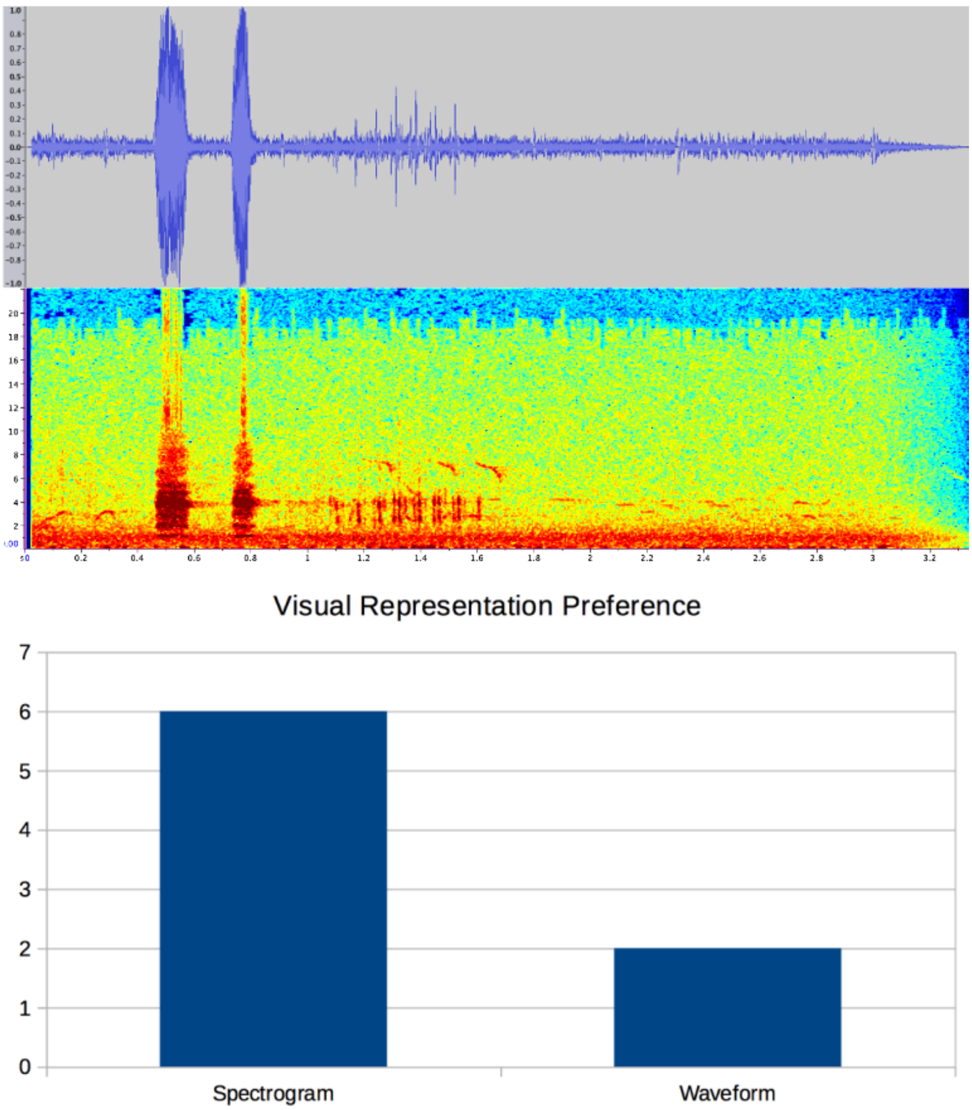

Figure 4.3: Results of participants' data-representation preferences given two visual representations of audio: above, waveform from Audacity, below, spectrogram from Raven.

4.2.1.6 Design analysis & evaluation of results

Despite similar touch-based affordances on tablet and watch interfaces, the larger visual interface was preferred unless interactions were constrained to a few seconds of displayed audio information. In instances where interactions involve longer-duration samples, larger devices, while marginally less portable, are preferable; non-square devices introduce orientation preferences subsequently examined in §4.3.4. Results support my first supposition, given observable preference for larger touchscreens; consequently ensuing design choices presuppose implementation on small tablet-sized touchscreens. Despite low TUI preference, discussion after third task completion indicated that the potential of higher-dimensionality representations remains compelling; potential tangible interactions, albeit at the bounds of portability with existing technology, introduce considerations of what information might be displayed, given additional physical dimensions.

My second conjecture, that the spectrogram representation is preferable, was supported by results from the second and third tasks; this informs future designs which endeavour to make spectrogram depictions easily comprehensible, through relevant choropleth mappings. First task results were balanced, with equal numbers of participants proposing a waveform and spectrogram, despite spectrogram familiarity not being expected as not having yet been explicitly presented. A subsequent experiment, §4.2.2, will determine whether this result was anomalous.

My final speculative proposition, that the interaction affordances of a touchscreen would enable sufficiently complex interactions for analysis in time and frequency, was supported; this informs future designs which endeavour to make spectrogram depictions easily navigable through familiar touch interactions. Interestingly, participants described tools coming from non-multi-touch-enabled 2D interfaces; all four tools described exist in desktop image processing software[6]. However, some participants availed themselves of additional interactions afforded on a touchscreen to select a ROI including swiping and dragging as means of scrolling and selecting and pinching to zoom and refocus attention. As conventional audio interactions with associated iconography (play, pause, search) have been present in playback hardware for decades, the introduction of new iconography for such interactions is unnecessary. Conversely, interactions such as selecting a region on a visual representation of sound in the time or frequency domain for time-stretching, pitch-shifting, and filtering, while comparatively simple to implement on a computer or mobile device, have historically been implemented with dedicated software or hardware lacking standardised iconography to represent underlying mechanisms by which these interactions occur. Characterising requirements for development of a platform for citizen science data-collection, interpretation, and analysis which can be implemented on a smartphone is therefore the continued goal of this research.

4.2.2 Study #2: representation & annotation choices

The second experiment explores visual representations of sound, introduced in §2.6.1, and how represented data can be interpreted, selected, and annotated by citizen scientists when depicted on a 2D touch-screen interface, chosen in light of first experiment preference results. Participants were shown paper prototypes of a set of audio representations, asked to complete a set of tasks using the prototypes and provide feedback regarding their decisions. Subsequent design choices are guided by analyses of these results which elucidate user expectations on the dimension of data-representation, introduced in §2.7.1.2.

[6] *e.g.* Adobe® Photoshop, Gimp, &c.

4.2.2.1 Hypotheses: representation familiarity & ROI selection

This experiment was designed to determine whether non-specialist users could discern and extract ROIs in visual sound representations without prior explication, particularly for the identification of energy and frequency thresholds, and perceptual effects of computational noise-reduction techniques. The experiment explores how participants identify mapped variables on different representations and whether colour provides viable axes for informative variables. ROI selection is the initial step for identifying a baseline noise floor for acoustic habitat mapping or when cleaning data for biodiversity assessment.

My first hypothesis is that linear spectrograms, the preferable representation for visually conveying information about sound from the first experiment, is the most familiar representation. My second hypothesis states that the waveforms provide most familiarity and afford the simplest selection process for selecting a timing ROI. My third hypothesis contends that waveforms provide the most familiarity and afford the simplest selection process for selecting an energy-bounded ROI. My fourth hypothesis is that linear spectrograms provide the most familiarity and afford the simplest selection process for selecting a frequency-bounded ROI.

4.2.2.2 Prototype design & study evaluation procedures

The experimental prototype comprised coloured pencils for participants to carry out proposed tasks and paper prototypes upon which were printed three visual representations of a short segment of sound. These prototypes depict a waveform representation of energy over time, a spectrogram representation of frequency over time, and a similarity matrix derived from work by Siedenburg[204] which involves "color-coding and superimposing similarity matrices (based on euclidean distances) of three 4-d feature vectors including centroid and flux. The intensity of colors corresponds to the distances between feature vectors over time . . . forming one complex higher-dimensional, less reductive representation of feature-time-series." Twelve participants saw an image of a linear spectrogram and 11 of a circular spectrogram, introduced in prior bioacoustics research by Pantidi et al., see §3.3.3.3[169] (fig. 4.4, pg. 100).

The experimental protocol involved introduction to the tools (*viz.* images, and drawing implements), each participant was provided paper prototypes depicting three visual representations of a short segment of audio and coloured pencils for annotation. Evaluation was performed on responses to questions regarding interface familiarity and data-complexity in the visual representation and an observational study documenting participant performance during each task. Likert survey questions investigating prior interface familiarity were followed by specific questions regarding the efficacy of each interface for performing selection tasks and the perceived complexity of performing such selection[138]. Participants were

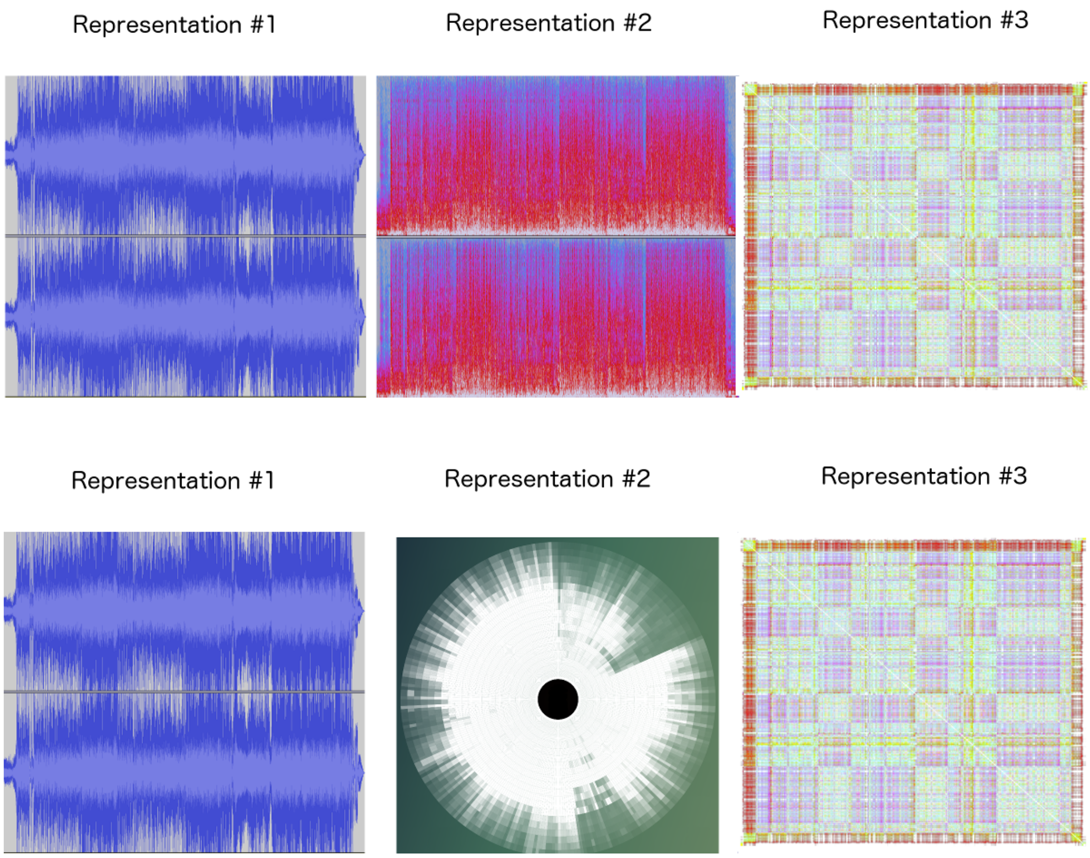

Figure 4.4: Paper Prototypes for the Representation Selection Experiment. Representation #1 is a waveform plot, #2, a spectrogram (linear above, circular below), and #3, a similarity matrix.

asked the following questions about selection of (1) an arbitrary time domain ROI, (2) the highest total energy ROI, and (3) the highest frequency ROI: 'do all prototype interfaces allow selection in this domain?'; and 'how would you make a selection?'. Participants' understanding of the dimensionality of each representation was discussed after all tasks were completed.

4.2.2.3 Data-collection & participant recruitment

Twenty-three undergraduate students, one of whom studied computer science and some of whose studies presupposed signal processing familiarity, performed the experiment. Experimental data were collected with questionnaires surveying participants' prior exposure to audio visualisations and asking closed and open-ended questions regarding their actions, as part of an observational study on tasks involving moving through samples and annotating selections. Participants signed a consent form and results have been anonymised.

4.2.2.4 Results

I define selection simplicity as use of the fewest possible delimiting marks; potential marks' implementation viability was not considered at this juncture. Representation familiarity was intended to identify whether and to what degree participants had experienced previous exposure to and interactions with presented depictions.

4.2.2.4.1 Representation familiarity

Upon viewing the prototype representations, participants were asked, on a Likert scale from 1, *Strongly Disagree* to 5, *Strongly Agree*, to respond to the statement 'I am familiar with this representation of sound' (fig. 4.5, pg.101). My first hypothesis was not supported.

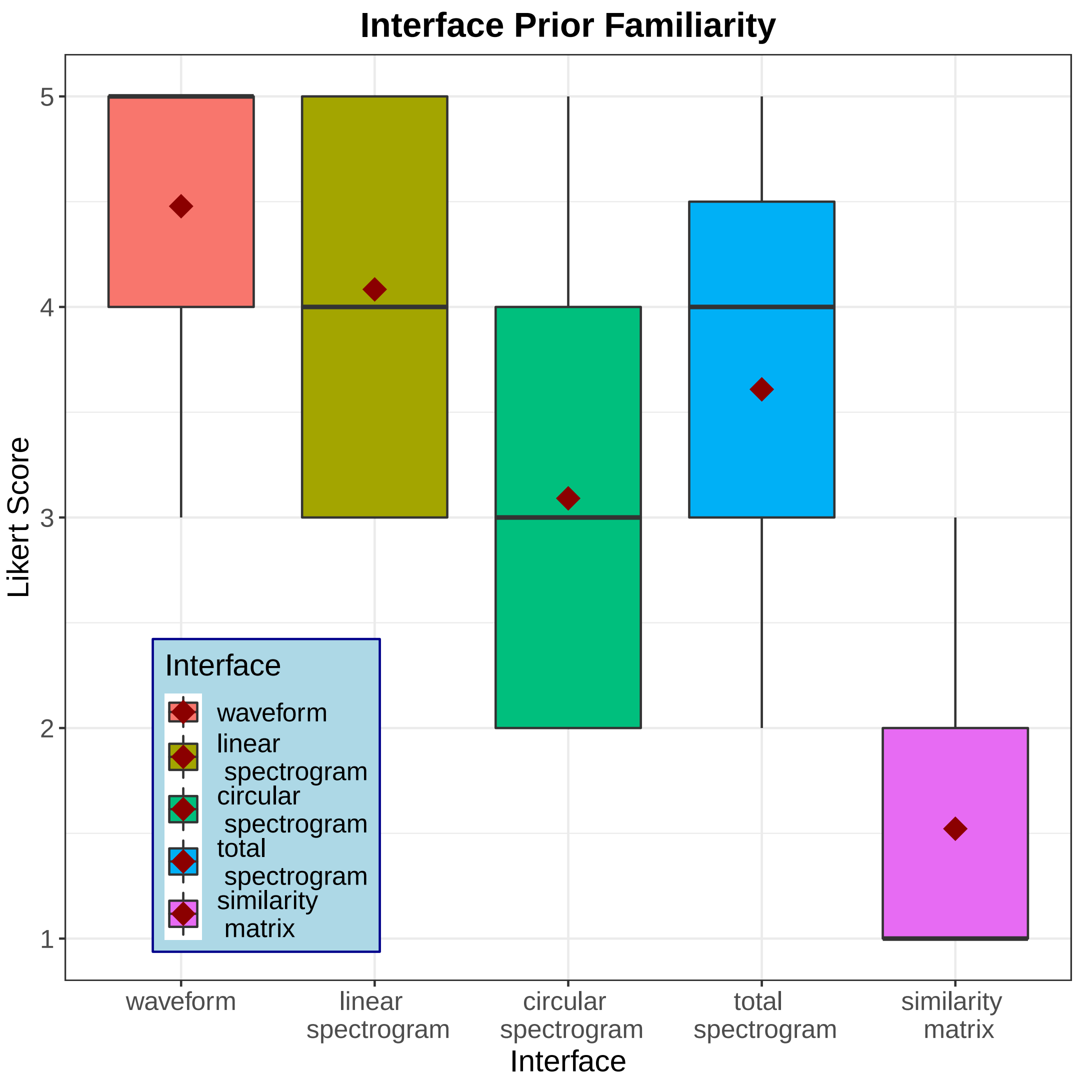

Figure 4.5: Summary of the participants' data-representation familiarity on a 5-point Likert scale. Plots show mean, denoted as diamonds, medians as black lines, and coloured range. As expected, the similarity matrix is unfamiliar.

The waveform representation mean familiarity score was highest (μ=4.48, σ=0.67, N=23).

Waveform familiarity was reported as resulting from visualisers common to audio playback applications and websites[7]. The linear spectrogram familiarity score was μ=4.08 (σ=0.90, N=12), while the circular representation score was μ=3.09 (σ=1.04, N=11), yielding combined spectrogram familiarity μ=3.61 (σ=1.08, N=23). While prior bioacoustics platforms have used the circular representation, it is not common and familiarity differed significantly (Wilcoxon's p=0.03) from that of linear spectrograms, diminishing overall familiarity. For the 3 of 12 participants who expressed neutral familiarity with the linear spectrogram, subsequent discussion revealed that they had seen such depictions but were uncertain what was represented. Although the remaining 9 participants expressed a degree of familiarity with linear spectrograms, the question formulation limits comprehension of the range of prior exposures participant responses encompass; subsequent selection tasks attempt to clarify whether claimed familiarity correlates with increasingly effective interactions. This initial result guides further experiments using spectral depictions of audio with minimal introduction for engaging participants with sound frequency dimensions, impossible with waveforms. The similarity matrix was least familiar (μ=1.52, σ=0.73, N=23); since this representation is from academic research, such result was expected.

4.2.2.4.2 Time selection task

The participants were asked whether (1) a section of time could be annotated on each representation and (2) how many marks needed to be drawn. All 23 participants believed this could be done with the waveform, 19 believed it could be drawn on a spectrogram — 11 with the linear representation and 8 with the circular, and 9 believed it could be depicted on the similarity matrix (fig. 4.6, pg.103, top). Results show the time dimension: is most evident on the waveform; benefits from being a horizontal linear spectrogram dimension, whereas radial depictions in circular spectrograms are less intuitive; is not intuitive in the similarity matrix as it lies on the descending diagonal and is simultaneously rescaled along horizontal and vertical axes (fig. 4.6, pg. 103, bottom). The simplest waveform time selection with straight lines involves drawing paired parallel lines perpendicular to the time axis; 21 participants selected time in this manner while 2, who consistently annotated thus across all questions, drew circles[8]. Of the 19 participants who thought spectrogram time-selection viable, 16 drew two lines, the aforementioned 2 drew circles, and 1 drew rectangular bounds. The one who drew rectangular bounds connected the top and bottom of the prototype screen, neither adding nor excluding data. Of the 6 who drew two lines on a circular spectrogram, 2 drew parallel lines, while 4 correctly assessed the need to draw radii. Of the 9 who thought time selection possible on the similarity matrix, the aforementioned 2 drew circles, while 7 drew two lines; of those 7, only 2 correctly identified

[7] iTunes®, soundcloud, &c.

[8] In subsequent discussion, they correctly reiterated that this was the minimum number of lines necessary in this case.

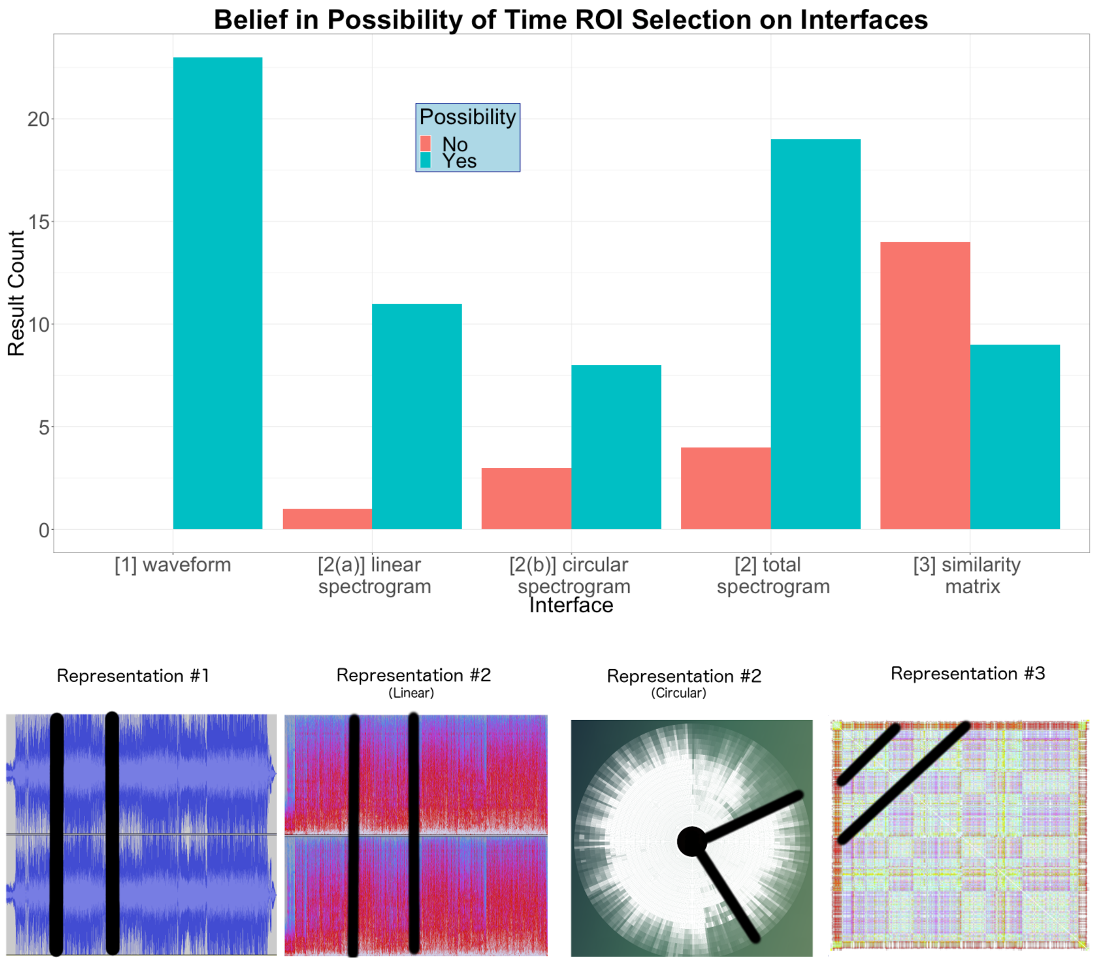

Figure 4.6: Above: participant belief that each representation allowed for time selection; below: most common correct time ROI selection lines.

and drew bounds perpendicular to the descending diagonal, the primary time axis. The remaining 5 drew lines perpendicular to the screen's bottom. Interestingly, the 2 who drew circles belonged to the circular spectrogram subset. Subsequent discussion was inconclusive in determining whether this influenced their selection approach.

4.2.2.4.3 Energy selection

Participants were subsequently asked (a) whether they thought each representation supported selection of the highest energy region and (b) how many marks were required. All 23 participants believed an energy ROI could be selected on the waveform, 20 believed it could be drawn on a spectrogram — 11 with the linear representation and 9 with the circular, and 6 believed it could be depicted on the similarity matrix (fig. 4.7, pg. 104). On the waveform, 2 drew circles, 20 drew parallel lines, and 1 drew a triangle. With the circles and triangle, participants enclosed the point of highest energy, while of those who drew lines, 18 surrounded the appropriate point but 2 failed.

Of the 20 who thought spectrogram representations expressed information about the

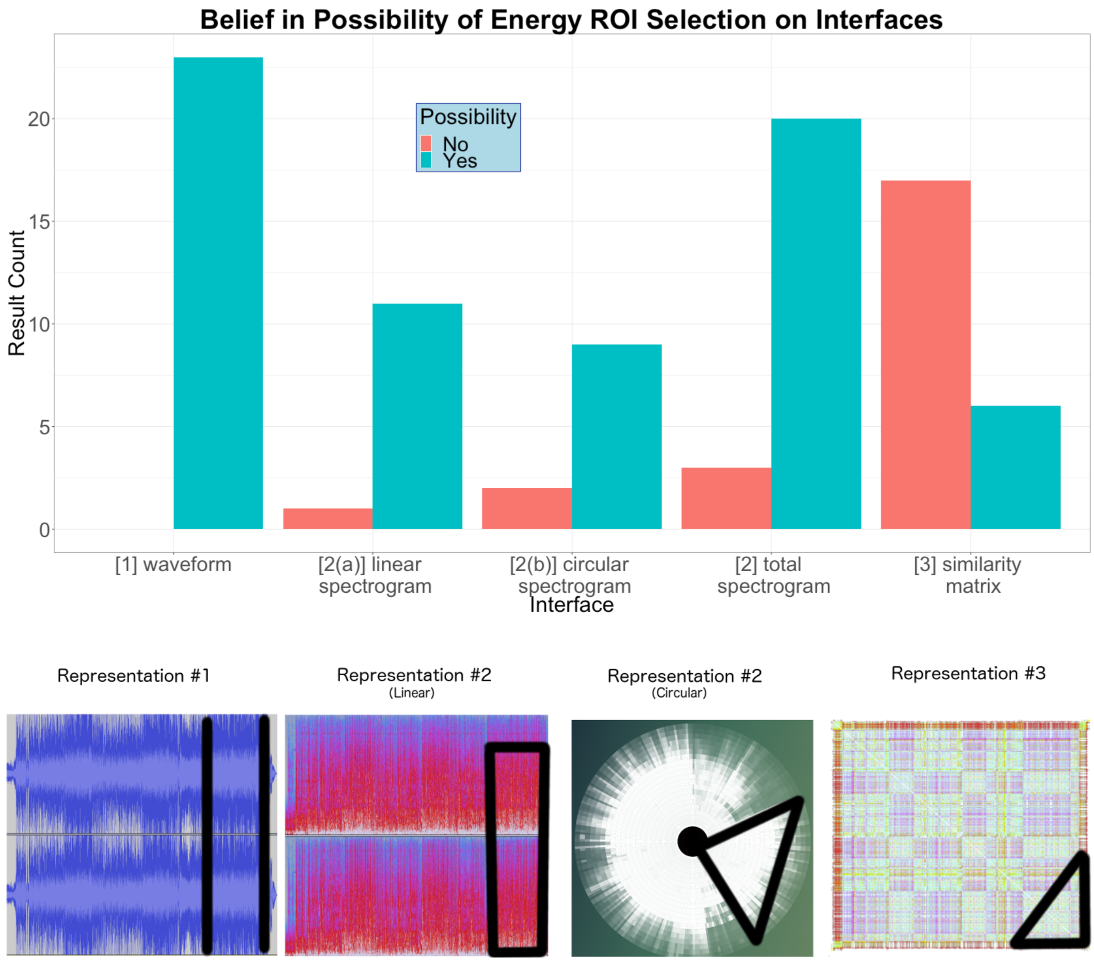

Figure 4.7: Above: participant belief that each representation allowed for energy selection; below: summary of common selection lines depicting locations of highest energy.

point of highest energy, 14 drew rectangular selection bounds, 3 drew triangles, 1 drew a pair of lines, and 2 drew circles. Both who drew circles and 1 who drew a triangle saw the circular spectrogram and accurately identified the point of highest energy based on choropleth intensity mapping. However, the participant who drew parallel lines, and 4 who drew rectangular bounds (1 on linear, 3 on circular) failed to identify the target ROI.

Of the 6 who identified a region on the third representation, 2 each drew circles, triangles, and rectangular bounds. The highest energy point in this representation is choropleth mapped to the magenta channel of the representation, furthest along the time diagonal. No participants successfully identified this although the selected triangles did encompass the salient space.

4.2.2.4.4 Frequency selection

The participants were finally asked whether (a) each representation allowed selection of the highest frequency ROI, and (b) how many marks were needed. Three participants erroneously believed waveforms depict frequency, 20 believed that the point of highest

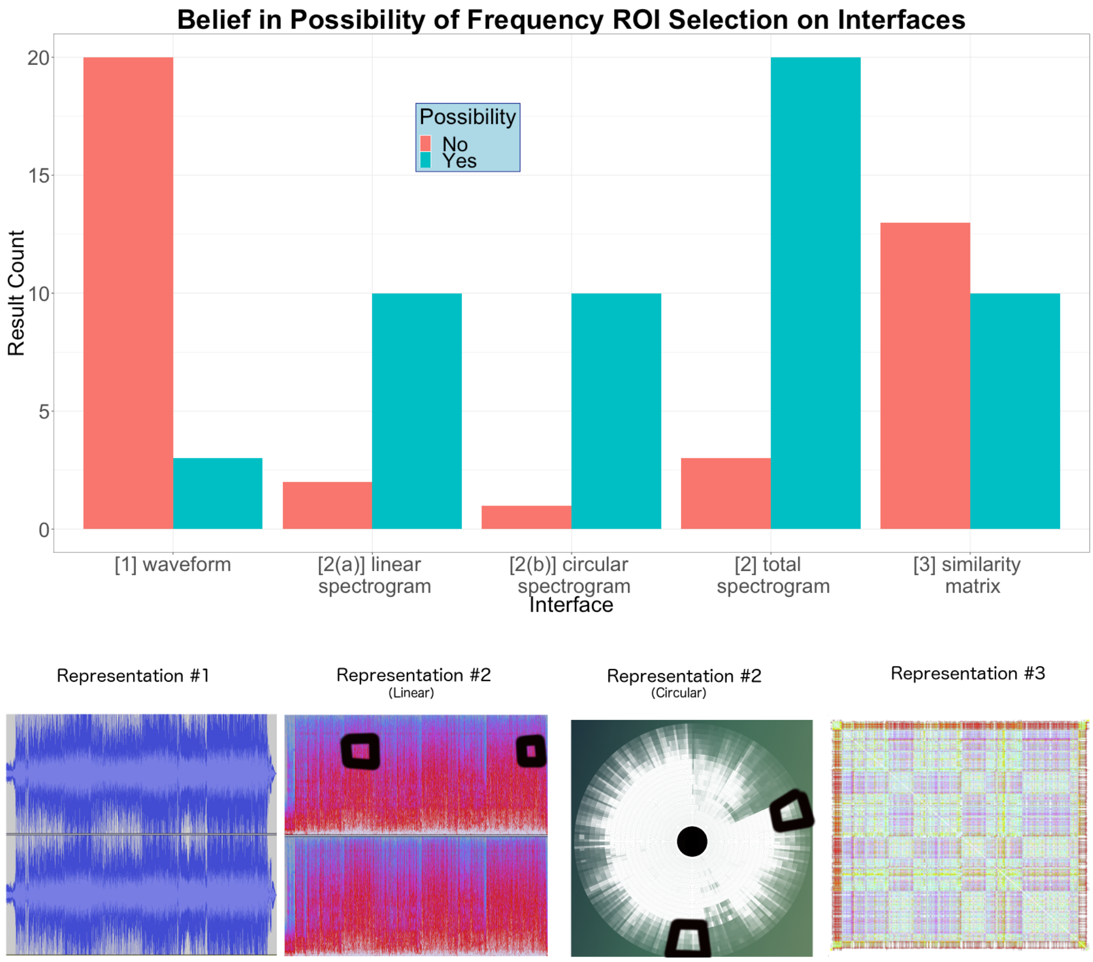

Figure 4.8: Above: participant belief that each representation allowed for selection of highest frequency; below: summary of common selection lines depicting locations of highest frequency as applicable.

frequency could be drawn on a spectrogram — 10 each with the linear and circular representations, and 10 believed it could be drawn on the similarity matrix (fig. 4.8, pg. 105). Of the 3 who drew a region on the waveform, 2 drew parallel lines and 1 a triangle; all were incorrect.

Of those who selected a region on the spectrograms, 2 drew circles, 1 drew parallel lines, 3 drew triangles, and 14 drew rectangular bounds. The region of highest frequency is the point past a choropleth-mapped threshold denoting the signal highest on the vertical axis of the linear spectrogram and furthest out the radial axis on the circular spectrogram. Participants who drew circles and parallel lines both correctly identified the ROI, although parallel lines do a poor job of selecting only the ROI. Only the participant who drew a triangle on the circular spectrogram failed to identify a high-frequency region, while on the linear spectrogram 2 participants, one each who applied triangular and rectangular bounds, missed a useful ROI.

Of the 10 participants who attempted to select the region of highest frequency on the third representation, 2 drew circles, 1 a triangle, and 7 rectangles. Since frequency is only a component used to calculate timbre encoded in the turquoise channel it is not evident that there is a region of highest frequency on the representation, so all attempts were necessarily flawed.

4.2.2.5 Analysis & directions for subsequent investigation

The hypothesis that linear spectrograms would be most familiar was ill-founded, as it had not considered the prevalence of waveform representations in consumer digital audio playback interfaces. The spectrogram preference exposed in the first experiment likely resulted from participant selection bias. Subsequent designs endeavour to familiarise participants with spectral data-representations from initial interface interactions.

The hypothesis that time ROI selections on waveform representations are simplest, as the primary axis denotes time, was warranted. While participants who tried succeeded with time selection on the linear spectrogram, 1 was confused by the axes and did not attempt a selection. The circular spectrogram's depiction of the time axis along the inner diameter of the plot created confusion. The descending diagonal for the time axis on the third representation was sufficiently abstruse to limit effective selection by most participants, although some identified scaled horizontal and vertical time axes.

The hypothesis that waveform representations suffice for selecting highest energy ROIs was supported, although overall spectrogram comprehension was similarly high. The error rate for correctly identifying the region of highest energy on the spectrogram (17.4%) was twice as high as for the waveform (8.7%). While energy, depicted as loudness, was directly encoded as magenta on the similarity matrix, the participants' conceptualisation of the keyless choropleth map was limited and results correspondingly poor.

The final hypothesis, that selecting the highest frequency ROI would be easiest on the linear spectrogram, was supported. Few participants laboured under the misconception that waveform depictions contain frequency information. The error rate on the circular spectrogram (9%) was half as high as that on the linear spectrogram (16.7%). While nearly half the participants believed frequency information was encoded on the similarity matrix, none could correctly discern the feature as it was not directly mapped to colour or location.

In light of these results, it is apparent that attempts to use spectrograms as primary representations for visualising audio requires explication before ROIs can be appropriately selected. Demonstrating spectrograms' usefulness necessitates explaining waveform representation limitations to users who, despite familiarity, may fail to conceptualise waveform axes. For higher-dimensionality representations of audio to be useful, intuitive representation dimensions must be explored, as the similarity matrix, comprised primarily of colour dimensions, caused confusion. I conclude that a spectrogram representation on a smartphone interface will be most easily comprehended by avocational users such as citizen scientists, although the affordances of higher-dimensionality representations and interfaces warrant exploration and choropleth mappings may need explication.

4.3 Analysis of a spectrogram recording & annotation interface implementation

I collaborated with Ben Elliott, a final-year undergraduate student, who implemented a mobile interface enabling geo-located field recording and real-time spectrogram visualisation from my design. I subsequently designed experiments to explore the efficacy of this smartphone-based mobile sensing platform for amateur avocational bioacousticians to record, annotate, and store a library of bioacoustic spectrogram samples[74] as contributory data-collection and data-processing citizen science. The mobile interface implementation enabled users to view spectrograms with multiple choropleth maps, leading to the preliminary question, in light of discussion in §2.6.2, of whether one choropleth map was optimal, or whether multiple options should be available. To this end, I mixed a set of samples of single and multiple avian utterances for repeatable analysis to treat identifying a ROI noise floor for biodiversity assessment as a human-in-the-loop (HitL) task, as introduced in the previous experiment. I designed and implemented an experimental testbed in MatLab[9] to examine users' perceptions of audio information visualised on the mobile interface. This enabled exploration of the degree to which users considered their selections and preferences sufficient and appropriate. These experiments consider source separation as a human intelligence task (HIT), rather than a data-driven inference task. Results influence the degree to which client-side processing can reduce data-transfer requirements when scaling designs. Results locate spectral images along the data-representation dimension of the framework introduced in §2.7.1.2.

4.3.1 Research questions

Can a crowd curate segmented biophonies? Interaction experiments were designed to determine user preference for visual representations of audio signals and whether users deemed their visual selections to be acoustically accurate. Results determine the content stored and transferred to a centralised database and the amount of processing to be performed on-device.

[9] https://uk.mathworks.com

4.3.1.1 Choropleth selection

My first experiment explores choropleth mappings of frequency energy information in spectrograms and the effectiveness of colour dimensions for on-screen data-representation. For the purpose of this study, participants were repeatedly shown spectrograms drawn with three different choropleth maps: a monochromatic sequential map, a multi-chromatic sequential map, and a divergent dichromatic map. Both chromatic maps used colours designed to work for standard colour-blind participants. Designs apply recommendations from [107, 137, 157], introduced in §2.6.2 (fig. 4.9, pg. 108). Initial hypotheses state that

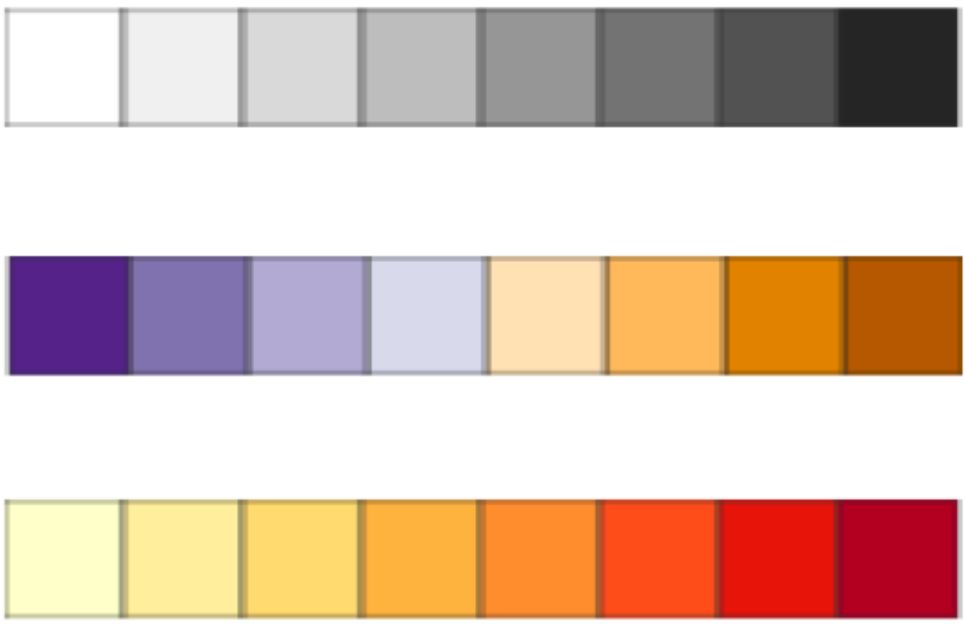

Figure 4.9: Three colour-blind-suitable choropleth maps selected for the interface: (top) monochromatic sequential, (middle) multi-chromatic sequential, and (bottom) divergent dichromatic; designs apply recommendations from [107, 137, 157], introduced in §2.6.2.

a monochromatic sequential choropleth map — greyscale — will be preferable, will be perceived to depict most accurately the audio, and will yield most efficient ROI selection.

4.3.1.2 Time selection

The next experiment examines the degree of perceived selection accuracy in time. I hypothesise that users pad time-domain selections, preferring silence before and after the target signal. If correct, then prior to dispatch to a server, given suitable automated onset detection, the ROI may be shrunk in the time dimension, reducing transmission bandwidth requirements without causing perceptual issues.

4.3.1.3 Frequency selection

Finally, I observe user bandpass filtering of bioacoustic recordings and explore whether filtering enables perceptually better noise reduction and simplified ROI identification. I hypothesise that users frequency-domain filter to minimise noise outside of ROI bandwidth. Selection should find the lower filter bound at the ROI fundamental frequency — given that bioacoustic harmonics rarely produce the psychoacoustic effect of missing fundamentals[136] — and that selection will include removal of higher harmonics, assuming that timbral brightness is still perceived.

4.3.2 Prototype design & evaluation study

I supervised development of an Android[10] audio recording and real-time spectrogram visualisation interface. It enabled user selection and annotation, using constrained data-entry protocols, introduced in §3.1.2.4, of a geo-tagged ROI depicting a species' utterance (fig. 4.10, pg. 109). I proposed this design, given characteristics of expert analysis tools as developed at the Cornell Lab of Ornithology (CLO), having identified a reduced parameter set for novice users. Iterative interface designs consider a series of selection tools in response to user feedback. Presenting complete short-time Fourier transform (STFT) parameterisation hindered the underlying goal of designing for avocational bioacousticians. I supervised a user study, building on observations of bioacousticians' and acoustic

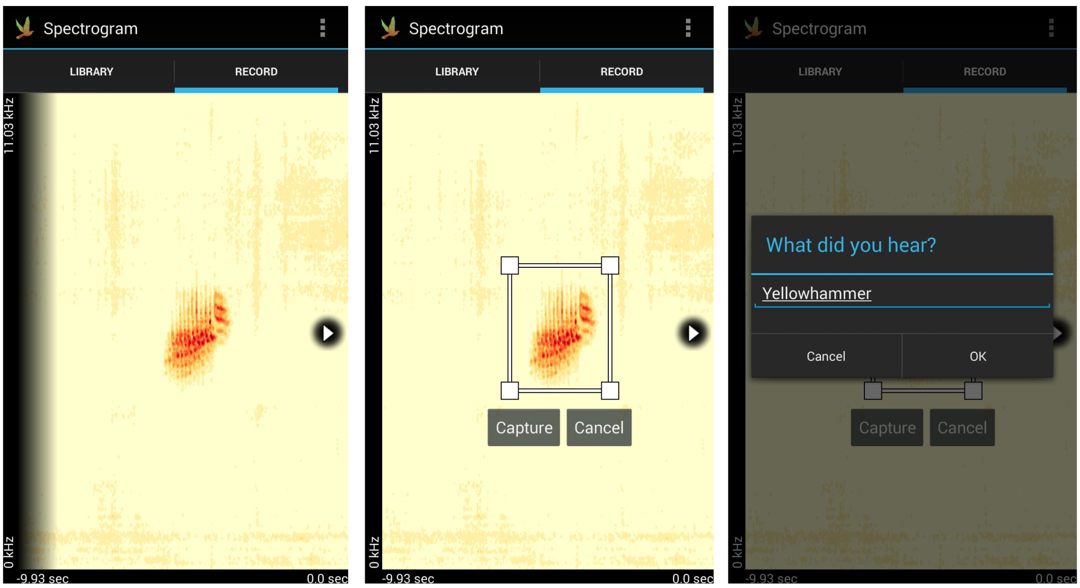

Figure 4.10: The mobile interface depicting spectrogram visualisation, ROI selection, and annotation screens.

ecologists' interpretations of visual representations of audio, examining whether novice

[10]https://www.android.com

users could mimic experts when performing bioacoustic ROI selection from spectrogram representations of noisy avian utterances. For ground-truth source material I generated a series of synthetic mixtures of single and multiple avian utterances[11] and geophonies using Audacity (fig. 4.11, pg. 110). High signal-to-noise ratio (SNR) avian utterances

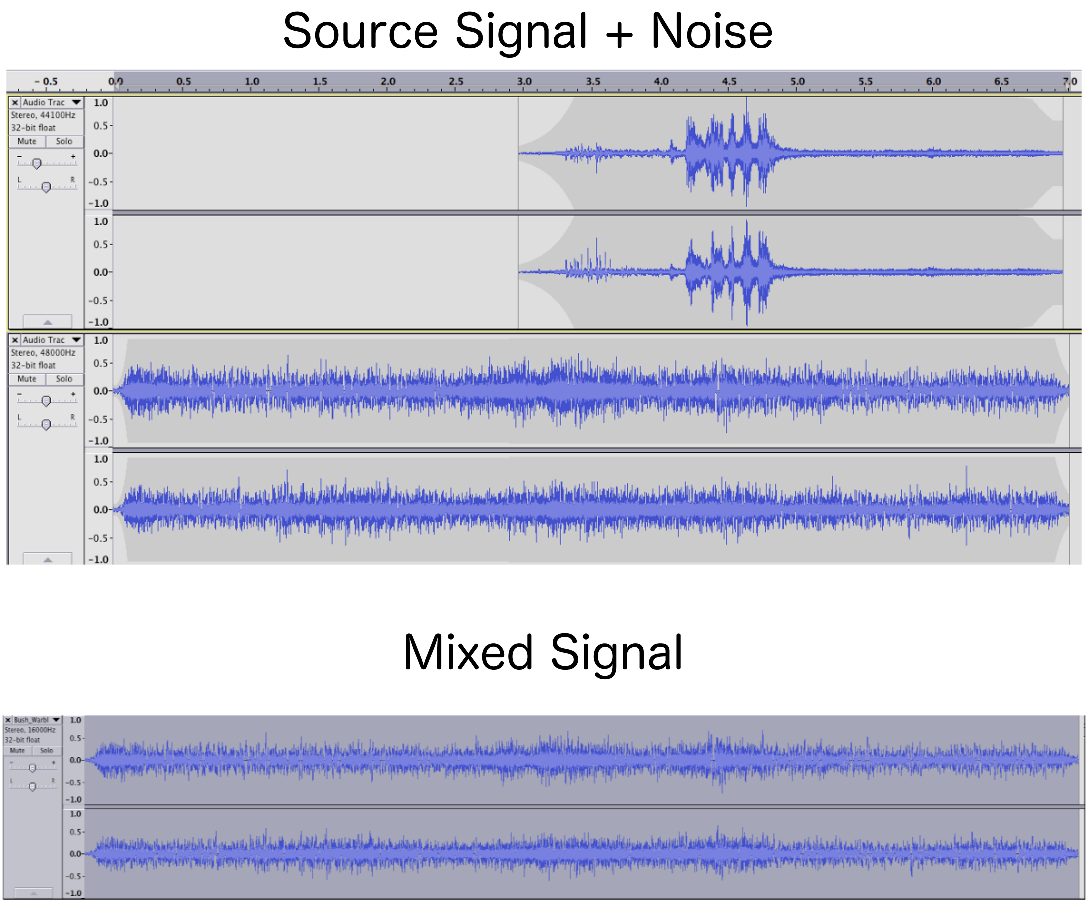

Figure 4.11: Creation of source signals, the waveform-plot representation of the mixed signal is too noisy to identify the ROI. Experiments explored whether source remains identifiable when visualised on a spectrogram.

(the top waveform), mixed at predetermined onsets with wide bandwidth noise (the waveform below), result in mixed samples (bottom); these waveforms are insufficient for original utterance endpoint detection. Samples were designed to reflect the likelihood that smartphone recordings contain environmental noise. A useful data-representation must support identification of low SNR bioacoustic signals by avocational users.

Each participant was asked to perform a number of repeated tasks with varying choropleth maps. The study addressed whether the identification of a ROI which represents a band-limited energy detector (BLED) was an effective HIT (fig. 4.12, pg. 111). My experimental framework recorded user output from time- and frequency-bound selection

[11]Sourced from Xeno-Canto: XC42685, XC52211, XC66288, XC71943, XC77206, XC77334, XC77354, XC120378, XC144821.

tasks for comparison with ground-truth known from synthetic sample construction. Subsequently, selected bounds were shifted and randomised and users ranked resulting samples by perceived accuracy. Each participant auditioned 2 of 6 available mixes containing

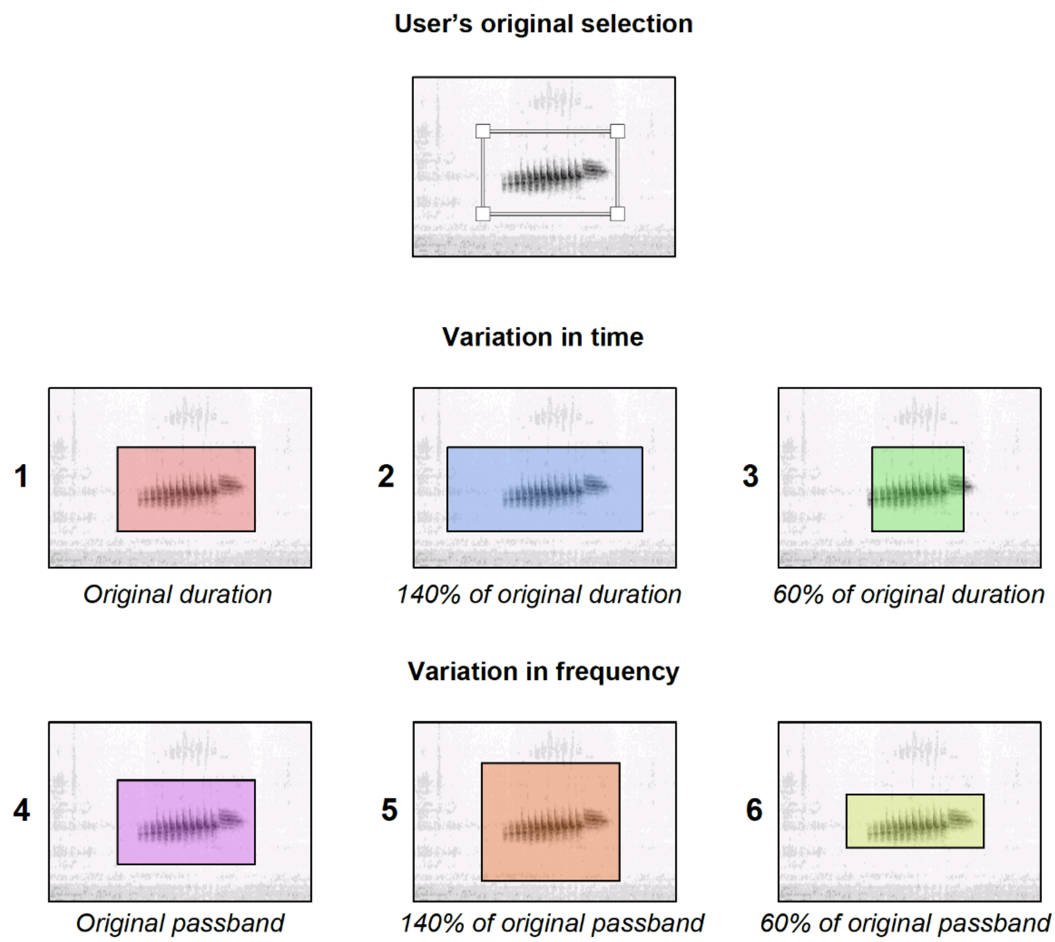

Figure 4.12: Creation of several ROIs formed by altering the time- and frequency-bounds of a BLED; from[74].

single or multiple avian utterances at different starting points within the sample. Mixes were shuffled; each mix was heard by 4 participants and each participant heard each mix 6 times, twice for each choropleth map. Selection-bound coordinates for each task were recorded for comparison with ground-truth from the synthetic samples. Upon ROI selection, participants were asked to audition 3 versions of their selections, altered in time and frequency. Altered time- and frequency-bounds presented in each experimental comparison task included participants' initial selections, one narrower, and one wider selection. Upon conclusion, participants discussed questions pertaining to their impressions of the selection features, ranking strategies, and choropleth preferences.

4.3.3 Experimental interface design & procedure

Discussing choropleth-mapping variations for data visualisation provides useful responses regarding participants' perceptions of the information presented during interaction with the application interface. Elliott describes in detail the prototype mobile application interface for spectrogram visualisation and ROI selection[74] (fig. 4.10, pg. 109). My experimental interface asks participants to audition and rank variably constrained time-bound and band-pass filtered signals associated with the spectrogram, based on users' initial ROI bounds (fig. 4.13, pg. 113). The initial experimental interface screen instructs users to play a sample whilst observing the Android application interface on which they select ROI time and frequency bounds (fig. 4.13, pg. 113); time taken to do this was recorded. BLED vertical marks on the Android application interface selection box bound call-onset and conclusion times; horizontal edges are parameters to an 8^{th} order Butterworth bandpass filter which reduces noise outside the frequencies of interest. Users were prompted to audition and rank three order-randomised clips which reflect the initial selections' time bounds and two versions which provide +/- 20% bound variation. Then they were instructed to audition and rank three order-randomised clips which reflect the initial selections' frequency bounds and two versions which provide +/- 20% variability on those bounds. Ranked results identified the optimal degree of noise reduction for human perception; this provided a baseline for on-device filtering to reduce data-transmission requirements prior to sending a recorded sample, corresponding spectrogram, and metadata to a centralised server, requisite for citizen science data-collection projects.

4.3.4 Participant recruitment & data-collection

A demographically varied group of 12 participants, representative of the application's diverse targets for age, gender, prior scientific knowledge, and avocational approach to ornithology, was sought. Participants ranged in age from 18 to 45, 5 were female, 6 had studied a technical subject and 10 gave a positive response when asked if they were interested in birds. One participant was affected by deuteranopia[12]; as their results fell within the remaining population's bounds they were not excluded from statistical analyses. My experimental method required participants to annotate depictions of 2 mixed samples for each of 2 orientations, landscape and portrait. Participants identified ROIs in representations with 3 different choropleth mappings: monochromatic sequential, multi-chromatic sequential, and divergent dichromatic. This yielded 144 sets of selection bounds and preferences for statistical analysis. To conclude the experiment, participants were asked open-ended questions pertaining to their actions and beliefs about the application interface, to identify their preferred mode of interaction and to offer feedback regarding

[12] Red-green colour-blindness.

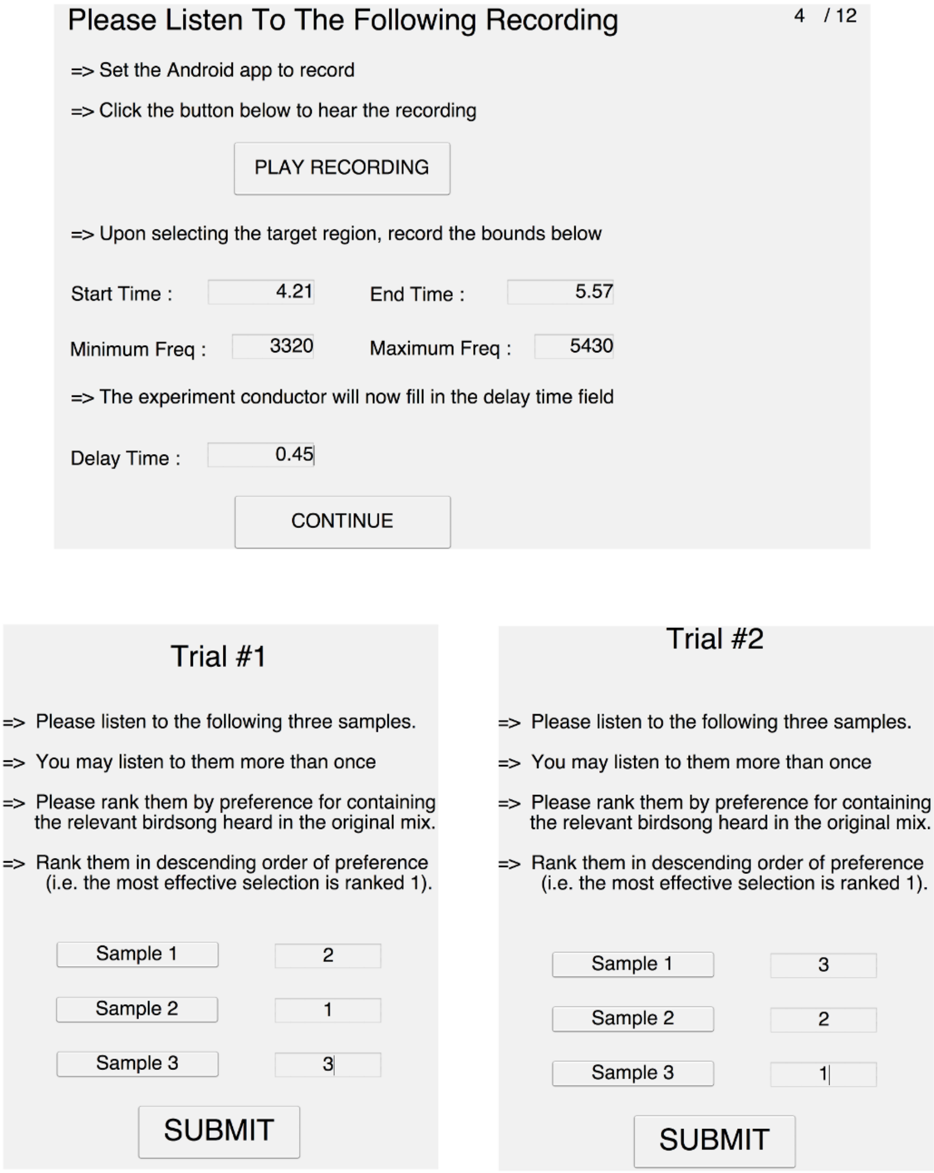

Figure 4.13: The user interface (UI) for the BLED's bounds selection-ranking experiment. Upon auditioning an initial sample and making a ROI selection on the Android interface, users are prompted to rank time- and frequency-varied selections.

application design. Participants signed consent forms and results have been anonymised.

4.3.5 Observations & results

The 144 sets of results produce preferences for choropleth and orientation when selecting time and frequency ROI bounds.

4.3.5.1 Choropleth preference

To test the hypothesis that the monochromatic sequential choropleth map would be preferred, participants ranked choropleth map preference from 1 to 3, with 1 the highest. Results show that the mono-chromatic sequential map was most preferred (μ=1.42), followed by the multi-chromatic sequential map (μ=2.08), and the divergent dichromatic map (μ=2.25). Thus the hypothesis is well-founded.

To test the hypothesis that the monochromatic sequential choropleth map most accurately depicts the ROI, participants were asked to rank perceived selection accuracy with each choropleth map. The hypothesis is supported as results showed the monochromatic sequential map ranked highest (μ=1.33). The divergent dichromatic map (μ=2.08), whilst less preferred, was considered more likely to correctly encapsulate ROI content than the multi-chromatic sequential map which was deemed least likely to reflect the target ROI (μ=2.17).

To test the hypothesis that the monochromatic sequential choropleth map yields most efficient ROI selection, timing data between hearing and selecting were collected. Selection onset timing data follow a log-normal distribution, so a parametric analysis of variance (ANOVA) test was performed on the logarithm of selection time taken, with results showing no significant differences across choropleth maps. The hypothesis that a mapping might lead to faster, and by implication easier, ROI selection was not supported.

4.3.5.2 Timing selection preference & error

The hypothesis that users would pad selection in the time domain, preferring silence before and after the signal in the ROI, was tested by observing preference ranks for three variably time-bound playback samples. Results showed strong dislike for clips with shortened time bounds (μ=2.63). However, there was only slight variation between preference for lengthened clips (μ=1.75) and users' original selections (μ=1.63) (fig. 4.14, pg. 115). Post-experiment interviews identified that the lower preference for shortened selections resulted from instances where bounds adjustment clipped relevant signal. Most participants' padded ROIs before and after the sound. Those whose bounds were close to signal onset generally preferred lengthened sample versions.

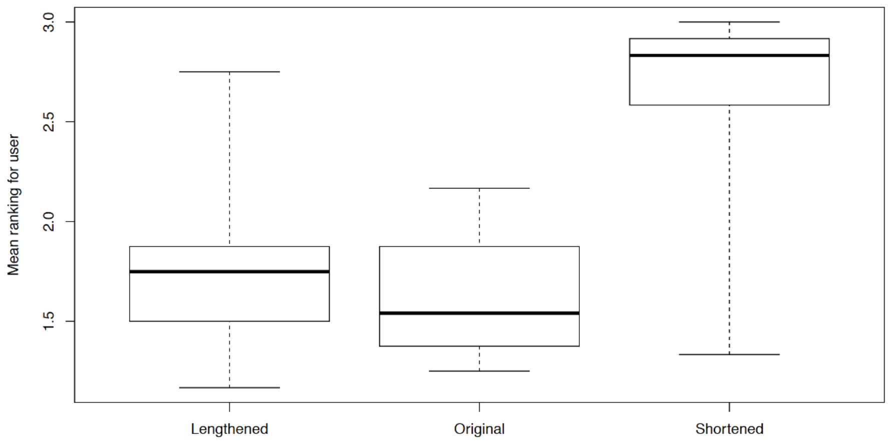

Figure 4.14: Participants' timing-selection preferences; a lower rank means the clip is preferred[74].

Ground-truth start and end time-bounds were calculated, by finding the first and last energy components within the original sample with energy amplitude above a 30% threshold of the mix's highest energy, for the original bioacoustic recordings prior to incorporation into mixed samples. Bound preferences were compared to bound errors,

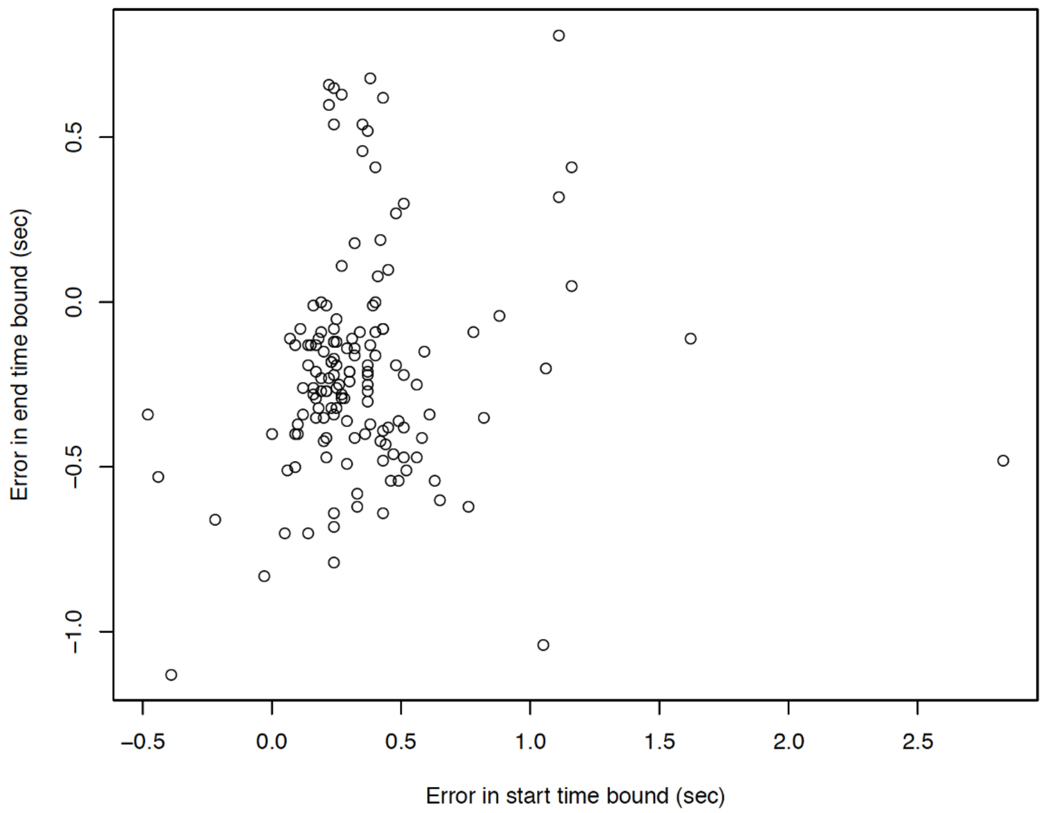

Figure 4.15: Participants' time-selection error; from[74].

determined as the difference between ground-truth and selected start and end times (fig.

4.15, pg. 115). No correlation was observed between start- and end-bound errors in user selection. Errors in end-bound selection were determined, in the interview stage, to result from individual variations in what participants considered the threshold above which a signal was believed to continue. The hypothesis that participants would prefer padding the time selection was minimally supported: generally user selections sufficed and did not significantly contribute to errors.

4.3.5.3 Frequency selection preference & error

The hypothesis that users filter in the frequency domain to minimise external noise was tested by observing preference ranks for three variably frequency-filtered playback samples. Slight preference appeared for clips narrowed in pass-band (μ=1.72, σ=0.69), compared

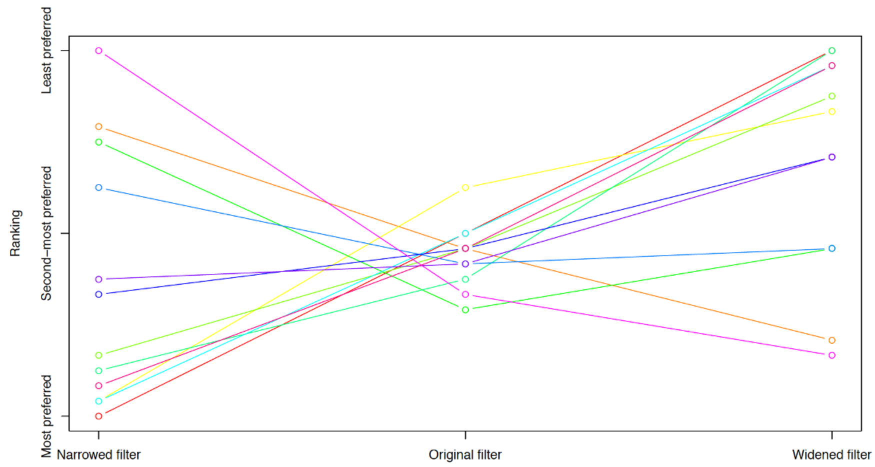

Figure 4.16: Participants' frequency selection adjustment preference results; each participant is drawn in a different colour[74].

to original selections (μ=1.88, σ=0.17). Preference diminished for samples widened in pass-band (μ=2.38, σ=0.60) (fig. 4.16, pg. 116). Wide variances for narrowed and widened pass-bands present bimodal data, with participants falling into two distinct groups: one strongly preferring over-filtered clips, the other strongly preferring under-filtered clips. In post-experiment interviews, the former identified desire for maximum background noise reduction, while the latter preferred background noise to contrast with bioacoustic signals even in low SNR recordings.

Ground-truth frequency bounds were calculated in Audacity for original bioacoustic recordings prior to sample mixing. This involved classifying minimum and maximum frequency bounds as highest and lowest frequency bins above an amplitude threshold. Ground-truth was compared to user lower and upper frequency-bound selections, yielding error values for each (fig. 4.17, pg. 117 and fig. 4.18, pg. 117). Spearman's correlation coefficient relating these two errors was -0.58 (p=0.00); bound errors were either wider

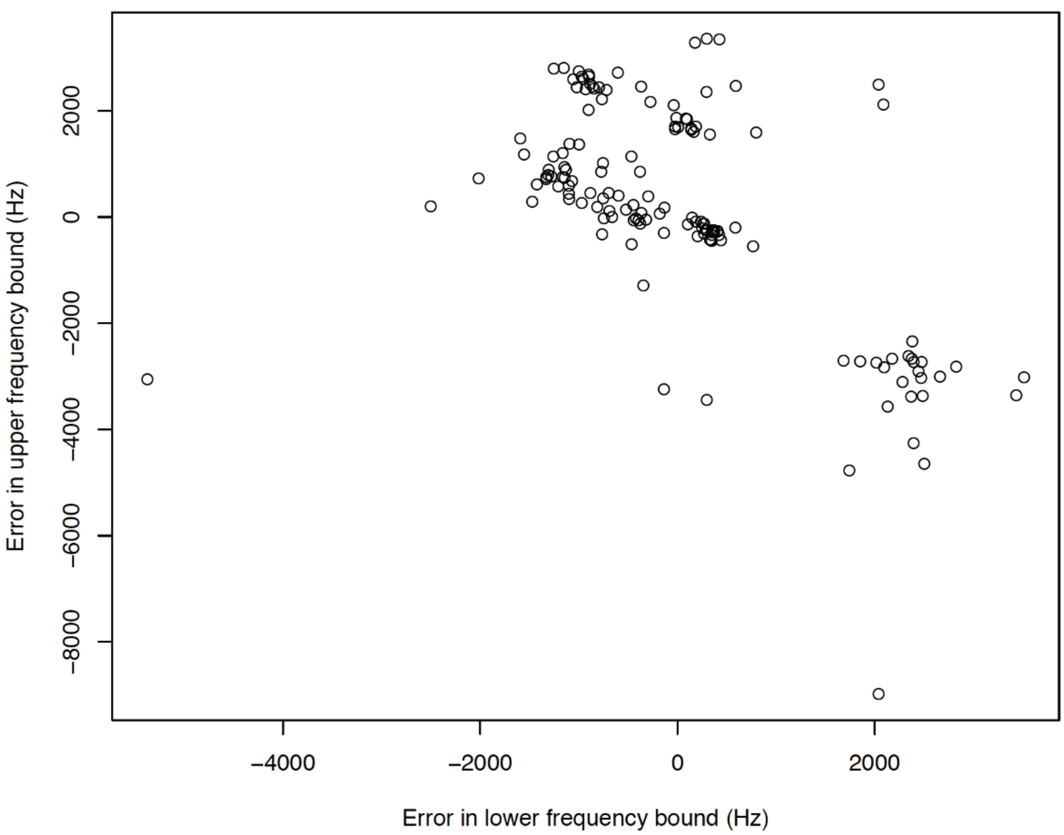

Figure 4.17: Participants' frequency selection error results[74].

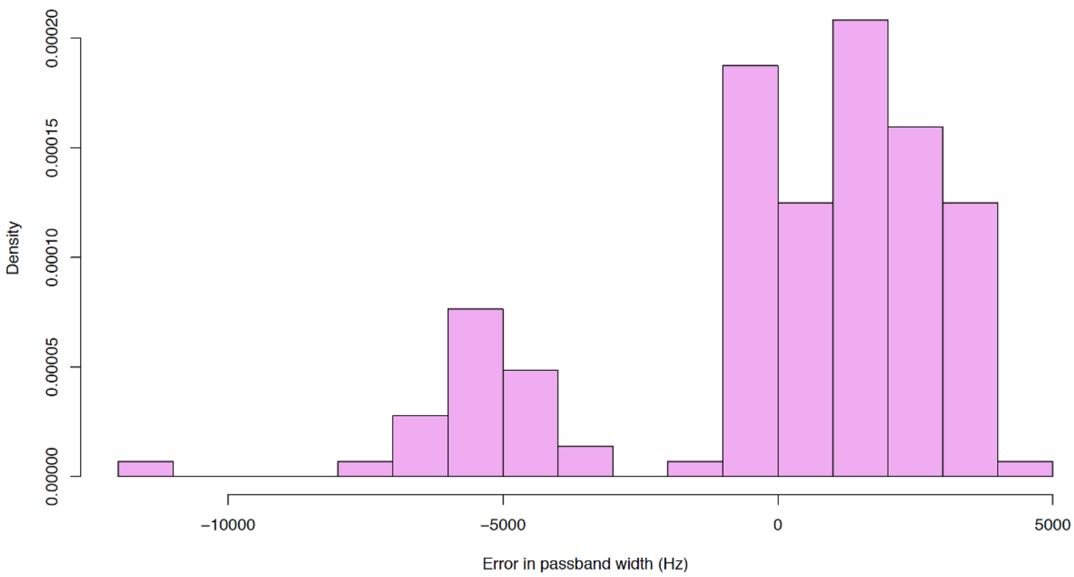

Figure 4.18: Participants' frequency selection error distributions is bimodal; negative values reflect a widened pass-band[74].

or narrower relative to a central frequency rather then consistently above or below true bounds. While consideration of bandwidth error is warranted, the bimodal distribution of error in pass-band width does not support my hypothesis that users filter consistently.

4.3.6 Analysis & discussion

Choropleth map ranking results showed consensus: the monochromatic sequential mapping was preferred. Thus in subsequent designs I use such mappings for spectrogram representations. The landscape orientation was considered more efficacious for spectrogram scrolling; this preference constraint guided future field deployment. While additional colour channels may convey additional audio information, the minimally sufficient choropleth map suffices. Such maps require fewer bits per pixel than chromatic maps, reducing transmission bandwidth for data upload to centralised databases from the field. Furthermore, inconsistent choropleth mapping of additional audio features necessitates unwelcome additional user training, as noted in §4.2.2.5. As noted in §2.6, citizen scientists interested in ornithology bring little intrinsic motivation to learn theoretical acoustics. Given a lower-complexity choropleth map, such an interface will be more usable in the field as transmission and battery requirements are reduced.

This experiment provides novel data regarding users' preference and accuracy in BLED bounds selection. Preferences for and correspondingly low error with original time domain selection, show that it is not necessary to perform additional on-device processing. Reduced processing requirements diminish ancillary battery drain in the field, of concern to users. Avocational users perform better with single-source than multi-source mixes; participants have produced more accurate time than frequency selections. Bimodal frequency selection task error means that subsequent designs could incorporate filtering preference parameters to support both user archetypes. For the group which prefers higher SNR, post-filtering results will contain fewer transmittable data while retaining salient information.

4.4 Preliminary field deployment

I organised a preliminary field deployment using a constrained version of the mobile data-collection and annotation application, reflecting user-testing biases for landscape orientation and monochromatic sequential choropleth maps. I designed a user-selectable filter switch for recordings prior to transmission, reducing client-side memory requirements. For this study no collection server was implemented. My research explores use of the prototype interface for initial participant engagement in the citizen science involvement cycle[191] (see fig. 2.3, section §2.1.2). The experimental context explores the value of community through interaction with others, as internally motivating, (see fig. 2.5, §2.5.1.1). This experiment explores whether personal motivation and personal engagement with biodiversity, (see fig. 2.8, §2.7.2) sufficiently encourage bioacoustic data-collection.

4.4.1 Site selection: RSPB Minsmere

Fieldwork was performed at Royal Society for the Protection of Birds (RSPB) Minsmere in collaboration with Planet Birdsong (PBS)[13], a multi-disciplinary initiative which teaches the science and music of birdsong to citizen scientists and students. I designed the deployment in collaboration with the RSPB Minsmere Learning Officer and PBS as a contribution to RSPB educational activity sessions to augment teaching outside school term. The Minsmere Youth Education and Families Manager noted that their site is undergoing a transition, common to RSPB sites, to reduce the role of guided walks within their educational model by providing patrons with mobile devices to augment their visits instead of booking guided tours. Although sending people to survey can endanger the thing being surveyed[156, 232] this risk was mitigated by directing participants to pre-existing paths and hides in Minsmere.

4.4.2 Exploratory questions & evaluation methods

Primary questions investigated through initial public user-testing were: what prior avian knowledge might avocational birdwatchers bring as participants to a citizen science project?; what prior exposure does an avocational birdwatcher have to technologically-enhanced avian interaction?; and does my platform provide participants valuable content as part of a birding experience? Prior to interacting with the application, participants completed a short survey about: prior avian knowledge by sight and sound; whether for identification purposes senses could be combined; prior touchscreen interface familiarity; prior use of digital tools for birding; and demographic data. After interacting with the application in the field, participants answered questions about: comfort interacting with live spectrograms, prior exposure to such representations, and likelihood of subsequent use.

4.4.3 Participant recruitment & data-collection

Over two days, walk-in participants joined one of four PBS educational programmes which included a lesson on the physics of avian utterances and an introduction to spectrograms using Raven. After signing consent forms, they completed the aforementioned pre-surveys and went on unguided site tours to observe and record what they saw and heard on mobile phones running my application. Upon returning to the Education Centre, participants completed post-surveys. For three of the four sessions the 5 available devices sufficed, however the final session had 27 attendees so only one family engaged with the application. Overall 15 participants engaged with the application, results have been anonymised.

[13] http://www.planetbirdsong.org

4.4.4 Results & observations

Each of the 4 groups comprised a single family; age distributions were logically bimodal, with a total of 7 parents, aged 33 – 46 (μ=40.3), and 8 children, aged 7 – 14 (μ=10.5). The duration participants had identified as birdwatchers was correspondingly split by age, although one family and several children did not self-identify. For those who did, the 6 adults had spent $\mu = 15$ years birdwatching, the 5 children $\mu = 4.2$ years. From self-reporting of total numbers of birds known, the adults' mean was 76.4 (σ=31.2) species, the children's mean was 19.4 (σ=10.3). When asked to count species known by sound, adults reported mean 35.0 (σ=15.6), children reported mean 6.3 (σ=5.2). Although self-reporting is potentially unreliable, participants' estimates were identified by the Minsmere Learning Officer as similar to prior survey results. When asked if they kept bird lists, 7 participants said yes and 8, no — interestingly, this split was neither along age boundaries nor family groups. When asked if they would record either hearing or seeing a bird, of the 15 participants, 8 said yes, 4, maybe, and 3, no. All of the adults reported having experience with touchscreens, with a mean experience of 5.43 years (σ=1.99), while only three children had such experience (μ=3.33 years, excluding those with no prior experience). Only two adults and no children reported prior use of digital tools to augment birding.

Before the introduction by PBS, 11 participants had not previously seen spectrogram representations, while 4 had some exposure. In response to the question of whether, when using the application, they could visually identify the utterances that they were concurrently hearing while making a recording, 8 said yes, 4 –sometimes, 1 –maybe, and 2 –no. When asked if they could recognise the call that they were hearing by the shape of the spectrogram on the screen, 5 said yes, 4 –'in a few instances', 1 –maybe, and 5 –no. Finally, when asked if they would use such an application in the field again, 11 said yes and 4 –maybe; of the latter 4, 3 were adults.

4.4.5 Analysis

While these participants had self-selected to engage with nature, given their presence at RSPB Minsmere, prior knowledge that avocational birdwatchers bring to citizen science projects ranges widely, and few use digital tools. All participants were able to identify more species by sight than sound, although small sample size prevents significance calculations. Participants stated that acquiring better acoustic species identification was a valuable skill and that the introduction of visual sound representations, giving users static images encapsulating entire utterances which could be played back, aided comprehension. While prior touchscreen-interaction familiarity was not presupposed, younger participants expressed more interest in continued interface use, perhaps due to minimal prior experience with such interfaces. Since a RSPB goal involves increasing younger demographics' engagement with nature, this interface was considered a valuable addition to their education programmes and as a field data-collection tool. The platform adds value for participants in a birdwatching experience as well as for non-governmental organisations (NGOs) hoping to increase collected data. However, for use on participants' personal devices, a better dynamic interface design is needed to account for variable screen dimensions.

4.5 Summary discussion & guidance for subsequent investigation

These preliminary experiments, testing assumptions made in the design of prior citizen science data-collection interfaces, have guided my subsequent research by demarcating the design space prior to iterative software prototype development. I have applied interaction design research methods involving low-fidelity and software prototypes to elicit user data from questionnaires, interviews, fieldwork, and experiments in order to contextualise my designs. This iterative research practice has provided insight into user expectation and comfort before I engaged fieldwork participants in the situated space of bioacoustic data collection.

4.5.1 Summary of findings

While prior citizen science data-collection interfaces have rarely been mobile, such devices' affordances, increased prevalence, and diminished cost, suggest exploration into designing interfaces with suitable interaction potential. Prior projects have infrequently presented spectral audio visualisations for user interaction. Those that have, such as BirdSongHero[14], have inconsistent scales for representation comparison.

My initial experiments provide novel exploration of platform selection, data-representation visualisation, and interaction preferences for citizen science. Results include observations of a preference for large portable touchscreen devices over wearable or tangible interfaces. Such devices afford sufficiently complex interactions for the manipulation of spectrograms which were found to be useful representations for capturing frequency data, albeit not necessarily sufficiently familiar for eliciting interactions without additional explication. Subsequent results showed that while spectrograms were unfamiliar, they were recognised as encapsulating information necessary when identifying the highest frequency ROI; waveforms sufficed for time- and energy-based ROI selection. Designs which involve spectrogram visualisations may necessitate representation explication.

[14] https://academy.allaboutbirds.org/bird-song-hero/

As application of the aforementioned design guidance for ROI selection has yet to be extended to public-facing interfaces, my subsequent research artefacts explore novel context. Additional experiments extend my research into interaction preference, given consistent data-representation, while introducing avocational users to bioacoustic selection tasks. These experiments constrain data-visualisation variability, enforce spectral representations while extending research into choropleth perception, and examine my contention that bioacoustic ROI selection is a HIT. Preferences for monochromatic sequential choropleth mappings and landscape orientation were identified. Analysis of ROI selection accuracy for BLED bounds showed user preference corresponding with lower error, albeit better in the time than the frequency domain where bimodal task error results specify two archetypal user behaviours. Variable user preference has informed design guidance that filter parameters ought to be settings for a citizen science bioacoustic interaction interface. Treating time, energy, and frequency selection as HITs avoids issues common to prior automated approaches discussed in §3.2.3.2. Likewise, classifying ROI selection as a HitL computational task introduces the potential for motivational rewards to engage participants through project interfaces. Initial fieldwork survey results showed that prior use of digital tools by avocational bird-watchers was rare, as was prior exposure to spectral representations; however, across age groups users expressed confidence that the mobile interface enhanced visual comprehension of avian utterances.

4.5.2 Guidance for subsequent investigation

Having investigated the efficacy of prototype software for motivating engagement with a contributory project in both experimental and field conditions, engaging avocational participants, scientists, and conservation practitioners with my designs and their output remains integral to my continued collaborative design research. Having identified optimal interface data-representation dimensionality for supporting interactions with bioacoustic signal visualisations, assessing motivation for such interactions remains to be explored. My next design iterations therefore advance the premise that motivating citizen scientists' engagement through play will simultaneously elicit greater quantity and quality of data even from those whose primary motivation for engagement is not intrinsic. As noted in §2.2.2 that novices, lacking preconceptions, are more trainable, I will investigate the minimal interaction complexity necessary to engage those without prior bioacoustic or avian knowledge, while supporting the data-collection needs of professional scientists and conservation practitioners through collaborative design.

CHAPTER 5

Designing games for citizen science

Mobile prototypes described in §4.3 for field trials in §4.4 provide a data collection and annotation interface which offer neither extrinsic rewards nor participant knowledge validation. Absent prior knowledge or intrinsic motivation, I posit that games serve to engage participants in citizen science projects by providing motivation while building knowledge, discussed in §2.5. This chapter describes my collaborative design, drawing on focus groups and user surveys, and iterative development of a set of games for engaging students with bioacoustic data through play in co-created curriculum-based community science. I contextualise designing for engagement through play, specifically for these implementations, which I place within my game-design framework. I identify data collection and evaluation procedures and introduce research questions to discuss from play results.

5.1 Designing for engagement through play

My iterative design of interaction artefacts for engaging participants with bioacoustic citizen science considers processes whereby participants become and remain engaged with project activities through motivations afforded by playful interactions. *Initial contribution*, in the context of the motivational arc (fig. 2.4, pg. 39), requires participants' *attention to need to contribute* and *perceived capacity to contribute*. Games provide external motivation, enhancing the former and, by rewarding learning, increasing the latter. I designed an extensible software framework for rapid prototyping which supports differentiation along dimensions of the parameters proposed for my game analysis framework. My software framework supports various game mechanics on a single platform, allowing diverse emergent gameful and playful interaction dynamics and aesthetic exploration of bioacoustic data.

5.1.1 Citizen science participant engagement progression

Citizen science projects engage participants in myriad ways; my design research focusses on motivating engagement given various levels of prior intrinsic motivation and expectations of prior knowledge, through games. My participant engagement progression model (fig. 5.1, pg. 124) adapts and extends earlier models, including Rotman's engagement cycle (fig. 2.3, pg. 37) and Crowston and Fagnot's motivational arc (fig. 2.4, pg. 39), by supporting divergent engagement paths through diverse activities, depending on motivation for data interaction. Motivating potential participant engagement necessitates initial outreach

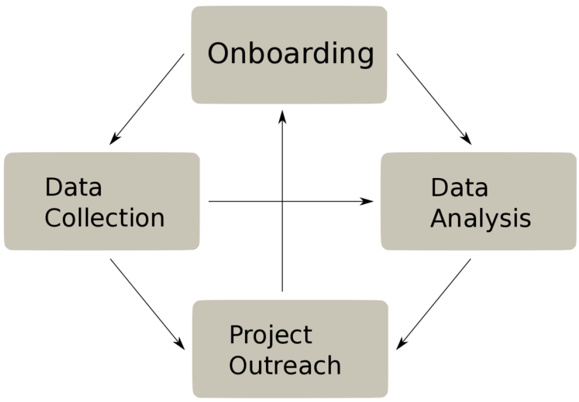

Figure 5.1: A directed graph of processes by which participants engage with citizen science projects and encourage project growth.

to stakeholder communities. Once engaged, participants may perform data analysis or collection tasks, depending on interest, perceived capabilities, prior knowledge, and project requirements, in contrast with approaches wherein volunteers only collect data and scientists perform annotation and analysis. Data collection requires prior knowledge if users produce metadata annotations without verifiable ground-truth, *e.g.* for species recognition claims. Games motivate participants whilst providing a platform for developing and validating knowledge and support participants transitioning from data collection to analysis. Data analysis requires prior knowledge, except in projects leveraging participants' spare computer processing cycles, or where a substitute for domain knowledge, allowing interaction without comprehension of content, exists[1]. Engagement with collection and analysis may further motivate participants to support project outreach, increasing stakeholder on-boarding.

5.1.2 Project interaction cycle

My game-design framework supports bioacoustic data interactions designed to drive participants along a knowledge development trajectory, prerequisite to providing quality data, while motivating participation through the project cycle (fig. 5.2, pg. 125). Hearing

[1] *e.g.* Galaxy Zoo, where solutions are geometric patterns rather than astrophysical.

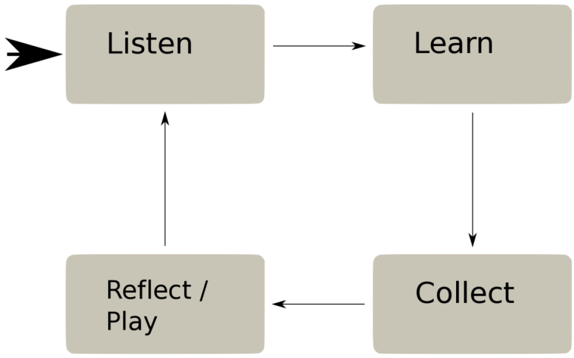

Figure 5.2: A directed graph of processes by which participants engage with bioacoustic data as they progress through my games, while increasing their potential for collaborative project contribution.

nature, prerequisite to bioacoustic citizen science project participation, can occur *in situ* or, with sounds incorporated into games, wherever play transpires. Passive participation involves collection of unidentified avian utterances, while active participation includes metadata creation requiring that participants can identify what they hear. Collection interfaces should support both active and passive contribution. Insufficient knowledge, linked to decreased motivation for collection, also inhibits metadata creation. Games may develop knowledge and confidence. Deeper comprehension of avian utterance characteristics results when participants reflect on data. This can be elicited through engagement with games and sound toys that encourage data interaction prior to field data-collection. However, over-reliance on mobile applications for engagement through play risks reducing the value of time spent in nature for collection and listening.

5.1.3 Artefacts motivating & enabling engagement

Chapter §4.3 introduced my mobile application design which engages participants with field-recording collection and annotation. Effective artefact use relied on participants' prior knowledge for annotation and intrinsic motivation for recording and library curation. Motivational rewards for engaging interaction with the initial prototype are presented in the following cycle, along with the game mechanics and dynamics supported by my software framework (fig. 5.3, pg. 126). My design artefacts support data creation and validation, reifying the learning and creation components of Jennet's MLC model (fig. 2.6, pg. 52). I designed and implemented a software framework encompassing game classes which supports learning, exploration, and play, and research whether compiled artefacts motivate interaction and enable participant learning while enhancing engagement with birdsong. The first class, memory games, involves learning to identify species visually and audibly from calls. The framework supports extensible datasets providing motivation for

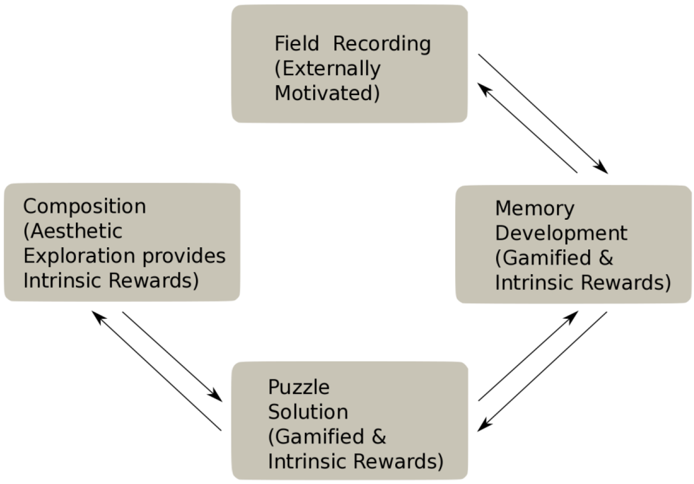

Figure 5.3: Graph of interaction processes between participants and my software artefacts. Whether software motivates, and degrees of freedom during play, as applicable, are identified. While initial recording requires intrinsic motivation, this is external to the software.

users to collect field recordings to increase the scope of learning by increasing baseline knowledge. Upon baseline knowledge acquisition and spectrogram familiarisation from the first game class, participants engage with a second game class, puzzles, designed to reward problem-solving while increasing data-representation and bioacoustic ground-truth comprehension. When participants achieve puzzle success, a third interaction class, sound-toys for open-ended play, introduces aesthetic creation, mixing choruses from an utterance library users have learned and, potentially, curated.

5.2 Design context & implementation

Experiments conducted on my software were pursued under the auspices of Planet Birdsong (PBS) as part of a Heritage Lottery Fund (HLF)-funded project, the *Wild Watch*, in collaboration with the Nidderdale Area of Outstanding Natural Beauty (AONB). My project design brief involved building games to engage primary school children and validate their avian bioacoustic learning, prerequisite to active project contribution through data collection and analysis. Various project stakeholders participated in collaborative design, contributing knowledge to influence project goals.

5.2.1 The Wild Watch

The Wild Watch project[2] was a three year project, the largest wildlife survey heretofore in the Nidderdale AONB. Launched in June 2017, the project trains citizen scientists to collect and analyse species-prevalence data for 50 target species within the AONB; of these,

[2] https://www.nidderdaleaonb.org.uk/wild-watch-0

23 are birds. Project goals include improving public knowledge and understanding of local wildlife, enabling volunteer data-collection surveys, and identifying priorities for habitat improvement and creation using habitat suitability modelling (HSM). As project training and data collection needs evolved, I implemented two iterations of games using my design framework. First-year training data comprised 10 primary calls of 10 target species, the second-year dataset comprised 14 utterances of the 4 owl species present in the region.

5.2.2 Motivation by design

My research contributions include design implementation of sound games and toys to motivate project engagement through learning and play considering Crowston and Prestopnik's concepts of satisfiers and motivators[178]. I designed the initial class of memory games to build participants' local avian knowledge necessary for accurate survey data contributions. Gamified elements extrinsically motivated goal-achievement in a pattern-matching memory game, while goal-state, introduced as a dimension in 2.7, pg. 57, was clearly defined. The puzzle game design engages participants with learned data-representations, providing intrinsic motivational rewards from problem solving; multiple winning goal-states exist. My play interaction class manifested as a *sound toy*[72]. I provide a playful set of composition interactions whereby participants find aesthetic value and intrinsic motivation, applying knowledge acquired in previous games to creating novel avian choruses.

5.2.3 Design implementation

Each implemented artefact uniquely fills a region within my game-design framework (fig. 2.7, pg. 57) along axes of variable difficulty, data-representation, and goal-state complexity (fig. 5.4, pg. 128). In the memory games, participants succeeded by matching cards encapsulating various visual representations and related audio content for target avian species. This game was furnished in three modes presenting different visual representations for identical audio. Each mode had varying difficulty levels but maintained identical goal-states – to find matching sets of species. Puzzle game implementation took the form of a multi-image reconstruction task with audio feedback. This game offered varying levels of complexity, necessitated a comprehensive understanding of spectrogram data-representations – taught in the memory games – and allowed multiple winning goal-states. My sound toy implementation comprised a composition toy in the form of an audio tracker[3], an interface enabling production of multi-track audio from a library of samples – here the library of target avian utterances. While this interface lacks explicit difficulty levels, complexity varies with the number of concurrent playback tracks and library samples users select for composition. Library sample data-representations are individually auditionable

[3]https://en.wikipedia.org/wiki/Music_tracker/

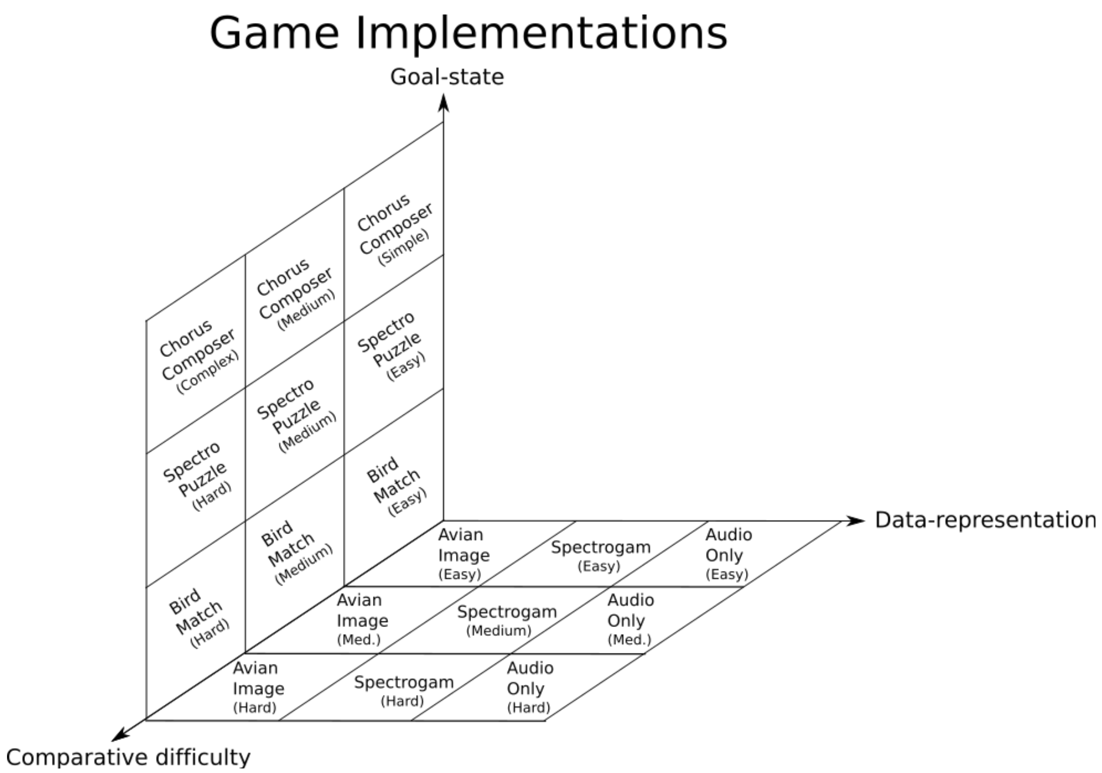

Figure 5.4: Where my implemented designs lie on the game-design framework described in §2.7.1.

spectrograms. Goal-state complexity is diffuse as user-defined composition completion is open-ended.

5.2.4 Fieldwork context

I have created designs which fulfil the United Kingdom (UK) academic prerequisites for a *structured experience unit* teaching children about *working scientifically* which includes methods of data collection, data analysis, and extracting meaning from data. Initially games were designed to target key stage[4] 2 classrooms (ages 7 – 10), but implementations were adapted and simplified as the Wild Watch project's target audience was revised to incorporate key stage 1 (ages 5 – 7). In collaboration with PBS, which provides a classroom introduction to spectrograms and combines birdsong with music education, my games were introduced into schools as training tools preparing students to engage with the Wild Watch's citizen science data-collection surveys. Over two week-long periods in June 2017 and June 2018 ∼240 students in years 1 to 5 played my games. Iterative interaction design practice explored how my games augment learning while motivating engagement with the Wild Watch project's needs. I also ran focus groups in both years to collect feedback from adult participant stakeholders who might contribute to and play the games. This feedback has informed implementation modifications and designs of games in

[4]https://www.gov.uk/national-curriculum

development.

5.3 Game design framework & implementations

Given interface and direct manipulation preferences determined from the design experiments presented in chapter 4, I built a development framework for rapidly prototyping the game classes introduced in section §5.2.2, implemented for Android. This permitted open-source availability of my research output[5]. My framework targets OS v.4.4 and higher, comprising over 96% of Android devices currently in use[6]. While Android has historic low-latency audio development issues due to their internal audio pipeline, for my concurrent playback purposes latency is sufficiently low.

5.3.1 Software design framework & initial prototypes

Applying standard interaction design research methods, each game class was prototyped and introduced prior to trials in schools to small groups of variable demographics comprised of 3 male and 3 female post-graduate university students, none of whom studied computer science or design. Each game class satisfied different game-design framework criteria (fig. 5.4, pg. 128) and therefore filled different spaces within my motivational framework (fig. 2.8, pg. 59).

5.3.1.1 Memory game prototype

Prior to implementing the memory game for Android, I researched prior studies on memory games with varying data-representations in structured learning environments[8, 182, 238]. I developed a mixed-media prototype incorporating paper flash cards depicting target species images and images of the corresponding call spectrograms; audio playback for each card was via Audacity. (fig. 5.5, pg. 130). For this prototype I used the 10 most common species from the 2014 Big Garden Birdwatch (BGBW)[7]; as participants were in Cambridge I presumed neither interest in nor familiarity with Nidderdale AONB target species. Participants performed a set of matching tasks, with images of birds and utterance audio, with spectrogram images and utterances, and finally with blank cards and utterances. Post-trial interview responses identified game mechanics as suitable for learning and the second mode, where spectrograms were seen and corresponding audio heard, as viable for introducing spectrograms.

[5] Game software is available for download at http://bioacoustic.games/ and the underlying source code is available on GitHub at https://github.com/isakh/BridgeGames and https://github.com/isakh/BridgeOwls

[6] https://developer.android.com/about/dashboards/index.html accessed 06/21/2019

[7] Species selected were: *Passer domesticus, Cyanistes caeruleus, Sturnus vulgaris, Turdus merula, Columba palumbus, Fringilla coelebs, Carduelis carduelis, Parus major, Streptopelia decaocto*, and *Erithacus rubecula.*

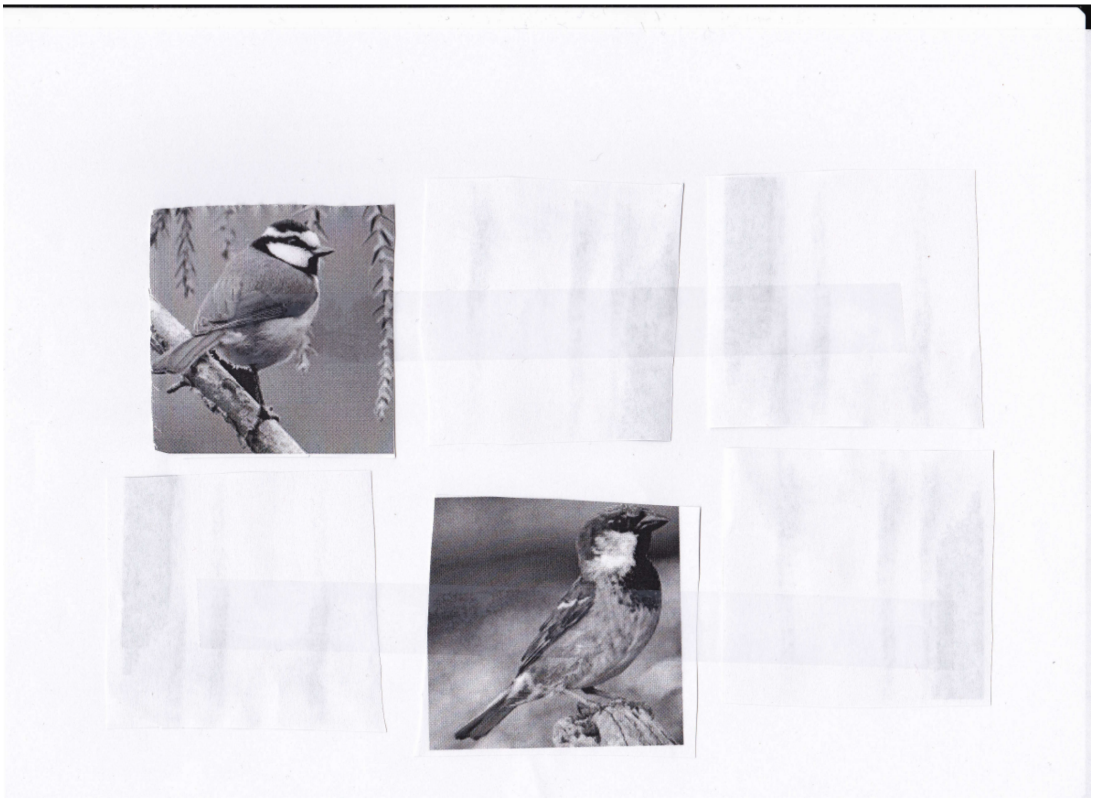

Figure 5.5: Paper prototype showing the visual components for the memory game, images of birds are depicted; here identical bird images were used, in subsequent mobile versions, male and female images were used. Additional versions were tested with spectrograms, always identical, and blank cards.

5.3.1.2 Puzzle game prototypes

I implemented a desktop Java[8] version of a 2D combination puzzle which partitions and shuffles tiles from an initial image (fig. 5.6, pg. 131). Such combination puzzles constrain users to move tiles into an adjacent open space. Upon correct reorganisation the final tile becomes visible and the entire image is depicted. The first image presented in the experiment was a bird, presumed to be familiar. The second image a spectrogram, presumed to be novel. Combination puzzles are commonly created from familiar images. During testing, spectrogram image unfamiliarity yielded initial poor performance with this prototype.

Thus I iterated the combination puzzle game design to relax the constraint that tiles be moved into an empty adjacent space, support tile-swapping, ensure target image familiarity by showing it to participants prior to tile shuffling, and allow success upon image row reconstruction, without row order mattering, relaxing goal-state. I implemented a mixed-media prototype for this design and designed an experiment exploring the potential for flow states, resulting from intrinsic exploration motivation, in games with multiple goal-

[8]https://www.java.com/en/

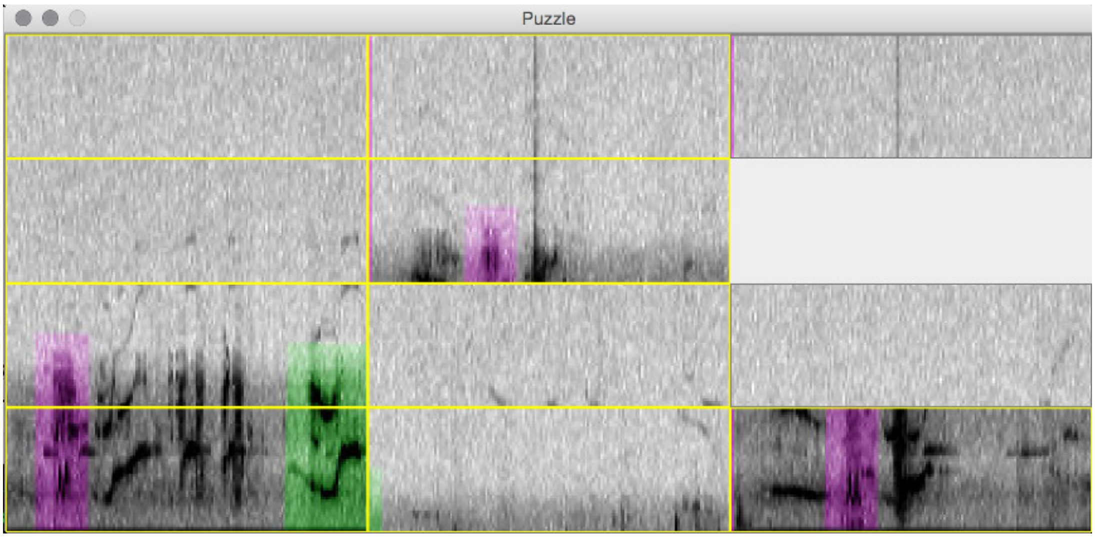

Figure 5.6: Prototype sliding-tile puzzle game, desktop Java implementation; adjacent leftmost bottom two tiles have moved towards a solution.

states[202]. This experiment required visual reconstruction of a set of song spectrograms from cards upon which syllable sections of each utterance were printed (see fig. 5.7, pg. 131, for design visual components). Play continued as follows: cards were distributed

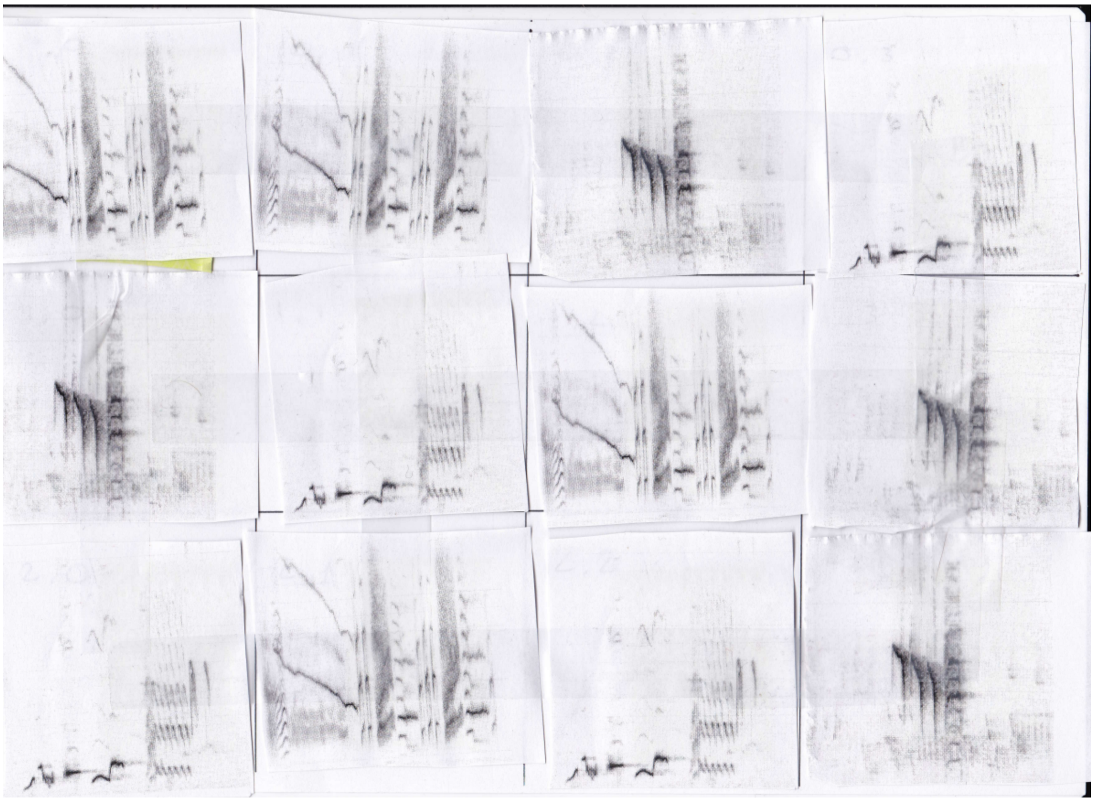

Figure 5.7: A paper prototype for the puzzle game. Cards depicting spectrograms of species' syllables are randomly placed on the board. Card pairs may be swapped. Winning occurs when images in each row reconstruct a single species' utterance, row order is not constrained.

in a grid face-up; turns involved swapping an arbitrary user-selected pair of cards; the goal was to ensure that each board row comprised spectrograms depicting a single species' utterance. Post-move, each row's audio was played back with Logic[9]. In a harder version of play, participants were instructed that each utterance had to be ordered within the target row to accurately reflect repeated call patterns in the original utterances. For some utterances this was tractable, but for those of the collared dove[10] and wood pigeon[11], where only number of syllable iterations within the call varies, arbitrary tile segmentation made ordered reconstruction inordinately complex.

5.3.1.3 Composition interface prototype

Trackers[12] have long provided a software interface model for computer music composition. Conceptually, trackers afford placement of samples on a grid where columns represent time and rows represent individual tracks, potentially linked to instruments within a composition. I designed a mixed-media prototype comprising a stack of sample spectrogram images, 4 duplicates of each of 10 samples, and a 4-row (tracks or channels) by 8-column (time steps) board. As participants placed spectrogram tiles on the board, I loaded corresponding avian utterance audio files into PocketSampler[13], a commercial Android sample playback application, which I manually triggered to reflect prototype board state. Observing participants' interactions led me to conclude that a board of similar dimensions could elicit musical complexity, leading to an intrinsically motivating flow state in chorus composition.

5.3.2 Software implementation structure

Having considered prototype implementation feedback, I designed an Android application encompassing all three game types (see Appendix A, fig. A.I.1, pg. 250, for application structure). The architecture supports project-specific deployment of underlying utterance datasets for localised training. Upon application launch, users are presented with a registration screen where registration name becomes the foreign key to several SQL database tables which store game play results. During experimental deployment in schools, all databases are implemented locally. Databases for each game class store user performance variables for each game played, others store content paths for serving the underlying dataset. Upon login users are presented with a choice of games, names were not descriptive to mitigate confounding results when names implied mechanics. Upon game selection, a launch screen appears with an option to trigger a settings pop-up containing game-applicable variables (the composition toy has no variable settings).

[9] https://en.wikipedia.org/wiki/Logic_Pro

[10] *Streptopelia decaocto*

[11] *Columba palumbus*

[12] The Tracker History Graphing Project visualises paradigm evolution. http://helllabs.org/tracker-history/

[13] https://play.google.com/store/apps/details?id=info.superkiki.pocket.sampler&hl=en

5.3.3 BirdMatch

I implemented the flash card memory game prototype as BirdMatch, presented as *Game 1* after the login screen. Three game modes provide users with target species' audio samples linked to one of three visual data-representations. Mode 1 depicts paired target species images, 2 an image of their call spectrogram, 3 provides no visual feedback. Each mode has three difficulty levels corresponding to different board dimensions. Total tile count can vary to support input datasets containing more or fewer utterances and target species, depending on project and location. Mode and difficulty define each game iteration played. On my game development analysis framework, these options allow players to move freely along the data-representation and solution-difficulty dimensions, limiting goal-state complexity to familiar and explicit (fig. 5.8, pg. 133).

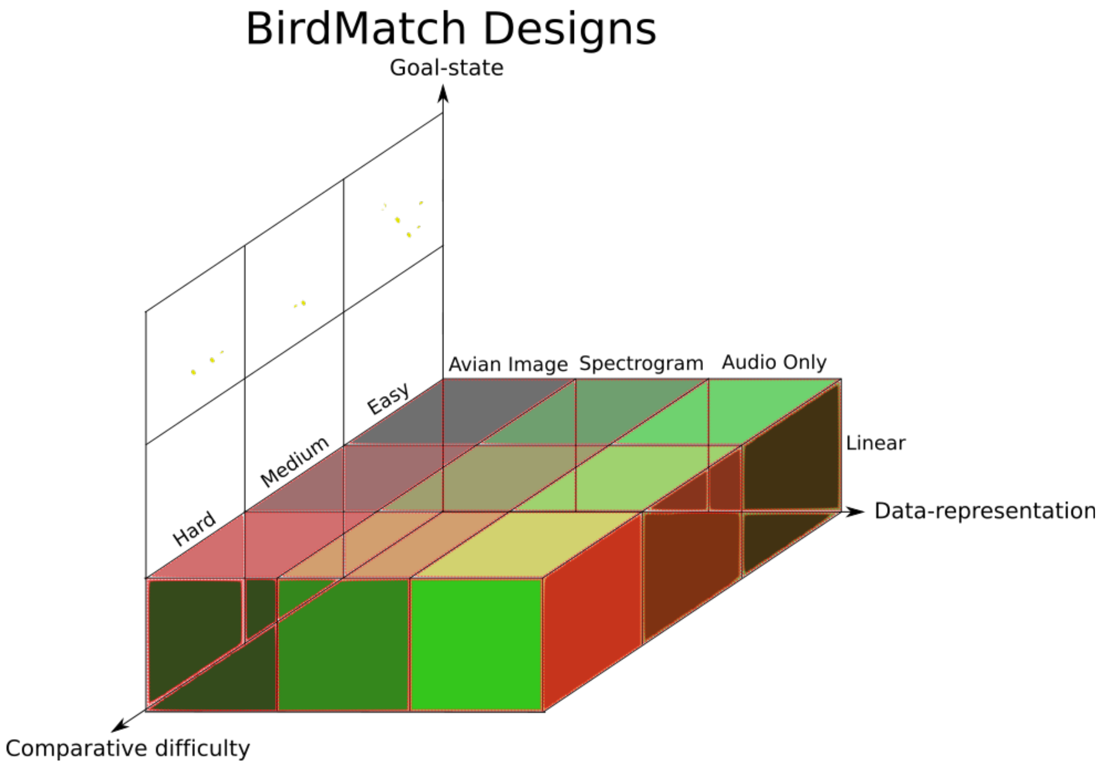

Figure 5.8: Where the variable modes and difficulties of the BirdMatch game lie on my game-design framework; all lie on the linear complexity level, while data-representations and difficulty vary.

5.3.3.1 Interface design & supported interactions

The game interface is partitioned into two regions, a board and playback controls (fig. 5.9, pg. 134). The board region displays a grid of tiles, initially showing question marks, encompassing most of the screen. In all modes, touching a tile animates flipping a metaphorical card and triggers audio playback of the associated sound sample. When two cards are exposed, if matching fails, they flip back. Matching triggers an audible reward sound and users see the species' name, providing informational feedback. The upper right corner of the screen presents a game timer and playback controls: play, pause, and restart. Time limits provide gamified feedback, evident from the outset, which affect final score calculation. Initial timer duration is a function of difficulty, data-representation mode, and a mix preference setting which controls whether concurrent samples may play, or flipping a second card cannot be triggered until playback of the first flipped card's audio has terminated. A boolean flag keeps track of game pause-state; when true, the countdown timer stops and no cards can be flipped. Upon winning, users are presented with a pop-up

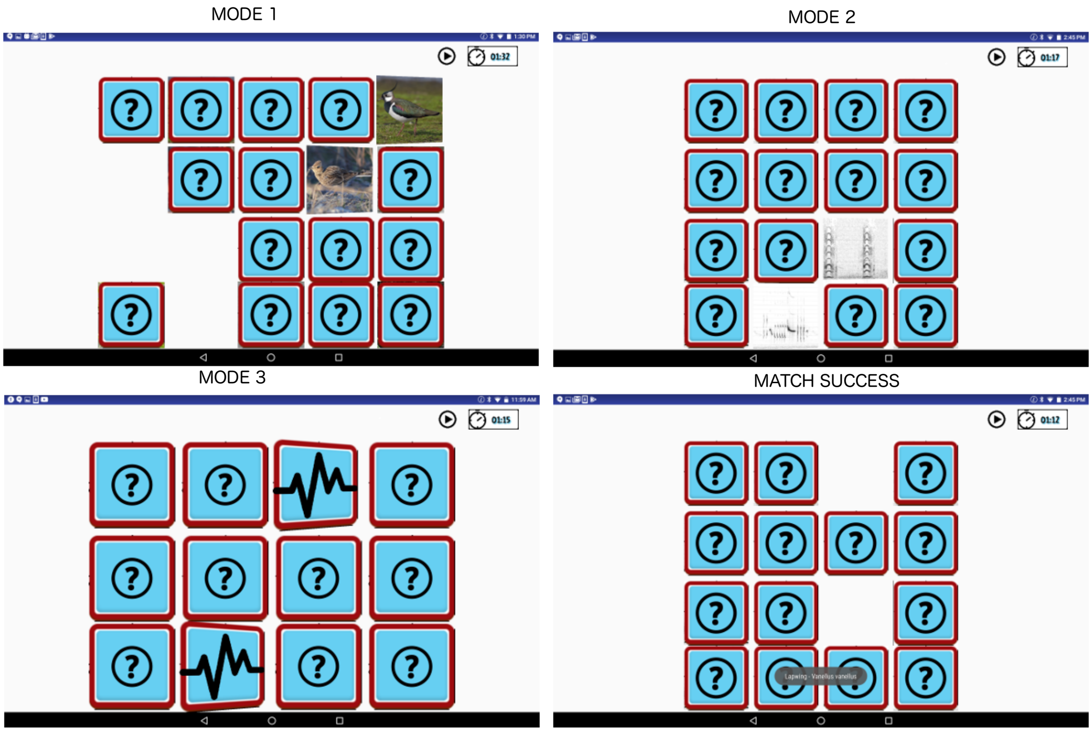

Figure 5.9: The three modes of the first game are shown here at different difficulty levels. Mode 1 is at hardest (5x4 tiles), Mode 2 at intermediate (4x4 tiles), and mode 3 at easy (3x4 tiles). The final screen shows the information provided upon successful match.

reporting elapsed time and a numeric score, calculated as a function of time and difficulty, which maps to stars, a grosser performance measure. The pop-up also contains a series of buttons, drawn as birdcages, which trigger: replaying mode and level; increasing difficulty, if sufficient stars have been achieved; changing subsequent game mode; or exiting to the game selection screen (fig. 5.10, pg. 135). Finally, users may entirely exit the application, presented as *Ready To Survey*, guiding them to apply knowledge acquired through play.

5.3.3.2 Data collected

For each BirdMatch board played, a database is populated with information collected in the matchGameData class (see Appendix A, fig. A.I.2, pg. 251 for complete list). The

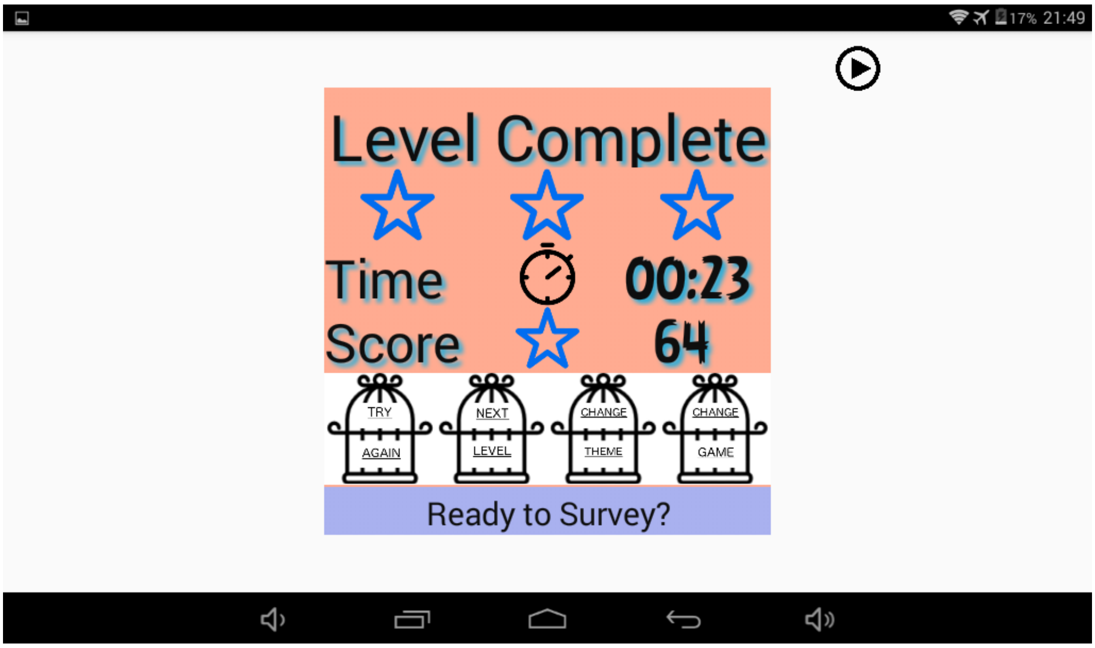

Figure 5.10: The pop-up dialog presented when players complete BirdMatch

primary key to each game in the results database is game-start timestamp, the time when first card flipped. User logins are used as database foreign keys, allowing querying of all games played by a given user. Data points recorded during play include total number of turns taken and turn durations, time elapsed between touches; these can be determined as a card identifier associated with each touch selection is stored for subsequent analysis.

5.3.3.3 Engagement & motivation

This game design enables engagement without training, as clearly defined goals are presented through a familiar game context. Classifying participant motivation in school environments is complex, as underlying motivation for learning may be confounded by interaction context: pupils are extrinsically motivated by the education system. Nonetheless, some participants actively expressed intrinsic motivation for learning about birds. User awareness of the timer provides external regulation, demonstrated to potentially provide extrinsic motivation. Scores provide introjected regulation which modifies control-oriented and autonomy-oriented users' extrinsic motivations differently. Unlocking subsequent levels and games provides intrinsic and extrinsic motivation, depending upon the degree to which participants have internalised the value associated with task success, noted in §2.1.1.1, where competence is introduced in the context of cognitive evaluation theory (CET). The guidance upon game completion, *Ready To Survey*, provides motivation for subsequent citizen science project participation.

5.3.4 SpectroPuzzle

The puzzle prototype was implemented as SpectroPuzzle, *Game 2* in my mobile application. Participants are challenged to reconstruct rows representing multiple species' utterances from shuffled spectrogram fragments, providing species differentiation training from visualised sound. The game offers three difficulty levels — varying numbers of concurrent species' spectrograms depicted by number of board rows — supporting complexity along the solution-difficulty dimension, since rows are unordered. As tiles depict spectrograms, the game does not vary on the data-representation dimension (fig. 5.11, pg. 136). This implementation allows users to select preference for easy or hard modes. In the former, rows are solved when all tiles correspond to spectrograms of a single species, in the latter, within-row syllable sample-order must correctly represent an extended utterance. As there is no intrinsic row order, goal-state complexity is familiar but not explicit and multiple winning states exist.

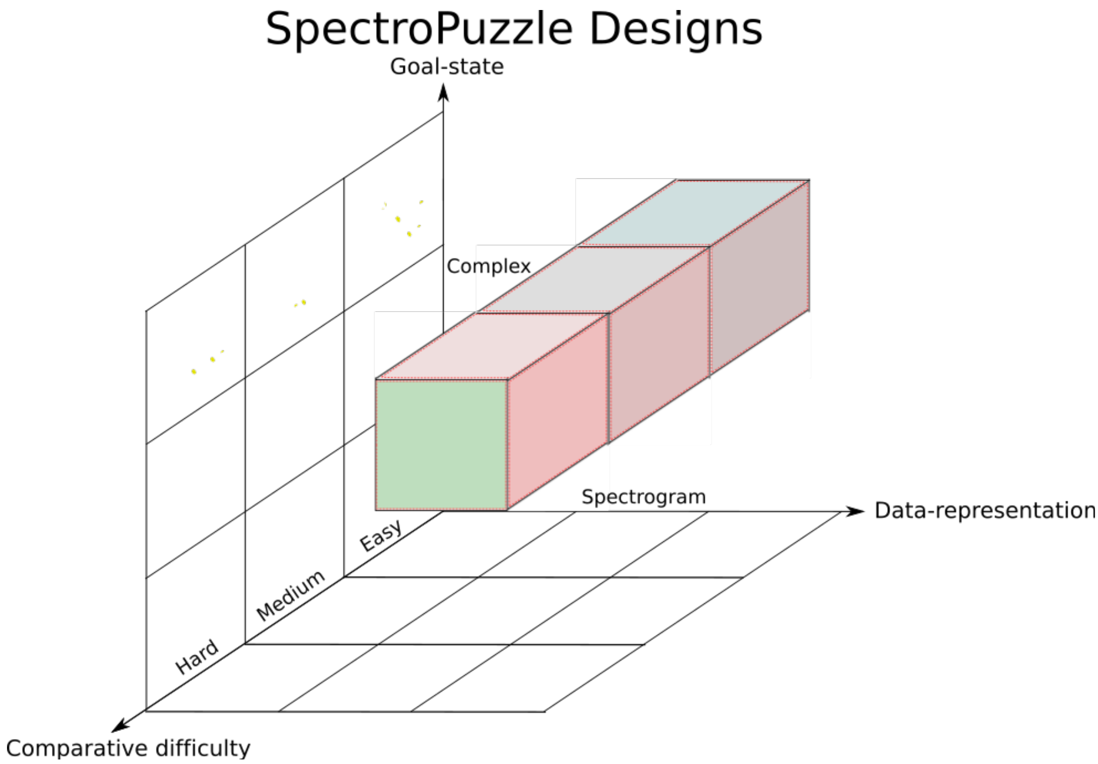

Figure 5.11: A depiction of where variable difficulties of the puzzle game fit in my game-design framework. The goal-state space is complex but tractable, all difficulties are present but only the spectral data-representation is taught.

5.3.4.1 Interface design & interactions enabled

The interface for this game is partitioned into three regions (fig. 5.12, pg. 137). As in BirdMatch, users are presented with a countdown timer and playback controls along the top. The puzzle board inhabits the majority of the screen and is tiled with shuffled spectrogram sample images on a grid with a variable number of rows, depending upon difficulty selected, partitioned into four columns. Along the left, playback controls permit auditioning current sample state for each row. Game play progresses as follows: the

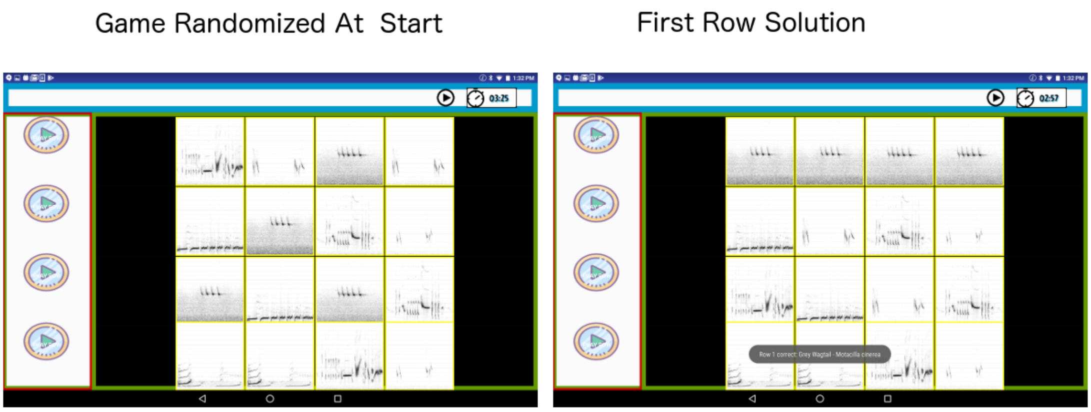

Figure 5.12: The puzzle board, randomised initially (left), and upon solution of the first row (right); this game challenges users to reconstruct rows representing the utterances of multiple species from shuffled spectrogram fragments. This trains them as citizen scientists to differentiate amongst species by visualising utterances.

user touches the first puzzle piece in a pair, highlighting it to denote selection; upon touching a second piece, highlighted in another colour, the two images swap; once swapped, highlighting changes colour to denote success, before disappearing. At any time, a row's spectrograms may be auditioned by pressing the corresponding playback button. Tile swapping continues until each row contains spectrogram images of syllables of a single species' utterance and, in hard mode, the correct initial utterance order. Upon each row reconstruction, users are presented with the name of the species whose utterance is depicted, and samples play, providing informational feedback.. Upon winning, a pop-up presents users with performance feedback, including elapsed time, a score computed as a measure of difficulty, mode and parameterised settings, and a corresponding number of stars (fig. 5.13, pg. 138). The pop-up presents buttons, outlined as birdcages, which allow users to replay the game, increase difficulty given sufficient prior game performance, exit to the game selection screen, or exit the application. A final option, presented as *Ready To Survey*, proposes an application for participant knowledge acquired through play.

5.3.4.2 Data collected

For each solved puzzle, a database is populated from the swapGameData class (see Appendix A, fig. A.I.3, pg. 252 for complete details). The primary key for each swapGame-Data instance written to the database is the game-start timestamp, first tile selection time. User login names are database foreign keys, allowing querying of all games played by a given user. The number of species and corresponding number of rows in the puzzle vary

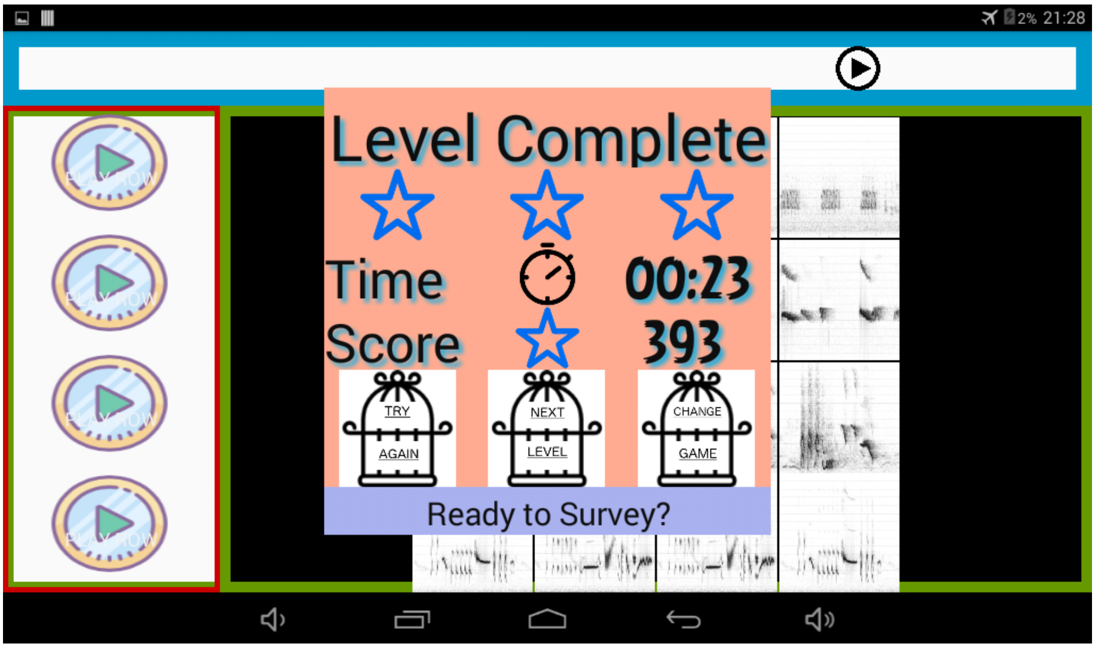

Figure 5.13: The pop-up dialog presented when players complete SpectroPuzzle

with difficulty level. Winning mode determines whether the solution requires pieces be ordered within each row. Combined, these define goal-state complexity and allocated timer duration is calculated as a function of these parameters. A boolean flag keeps track of whether the countdown timer is paused and no pieces can be swapped. Game data collected for subsequent analysis include number of turns taken and turn duration, calculated from time elapsed between paired touches. A board map, encapsulating coordinates of each tile location and current sample, is stored after each swap executes.

5.3.4.3 Engagement & motivation supported

Combination puzzles are presumed to be familiar to the target audience and motivational rewards associated with solving such tasks are myriad. Iwasaki explored motivation for solving puzzles and found participants' competence and playfulness during a game enhance intrinsic motivation[109]. SpectroPuzzle success necessitates prior internalisation of spectrogram data-representations, taught in BirdMatch. Data-representation internalisation promotes competence which provides intrinsic motivations for goal-state achievement. The timer provides external regulation, demonstrated to provide extrinsic motivation for autonomy-oriented users. Scores provide introjected regulation influencing extrinsic motivation, albeit differently for control-oriented and autonomy-oriented users. Unlocking difficulty levels and the final sound toy can provide either intrinsic or extrinsic motivation, depending upon the degree to which participants have internalised value associated with task success. The game completion option of *Ready to Survey* provides motivation to increase ensuing project participation.

5.3.5 ChorusComposer

I implemented the mixed-media tracker prototype, *Game 3* for my mobile application framework, as the ChorusComposer sound toy. The toy provides options for selecting amongst three complexity levels which vary the number of concurrent tracks and number of time-steps prior to looping. Introductory complexity offers 3 tracks and 8 steps-per-loop, intermediate has 4 tracks and 16 steps-per-loop, and complex offers 5 tracks and 32 steps-per-loop. In the context of my game development analysis framework, these options allow players to move along the solution-difficulty dimension (fig. 5.14, pg. 139). Samples are represented with spectrogram images, constraining data-representation dimension variation. With no inherent goal-state, interaction complexity is diffuse.

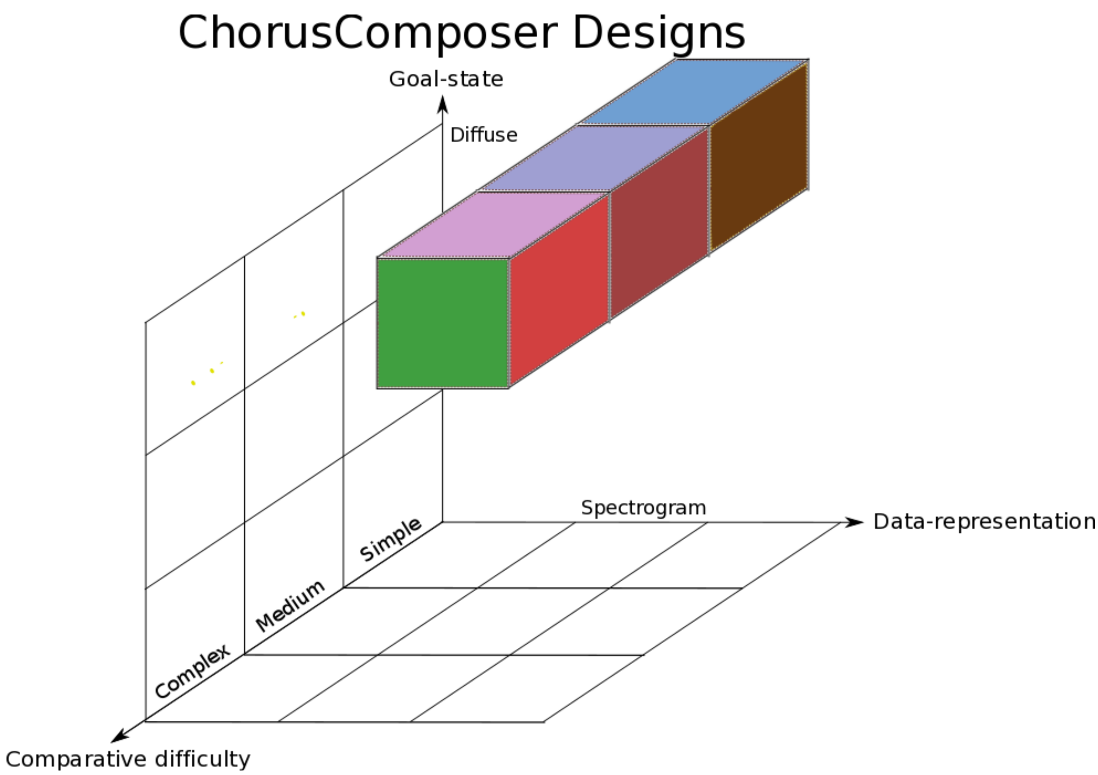

Figure 5.14: A depiction of where the composition toy, with various board complexities, lies on the game-design framework.

5.3.5.1 Interface design

This toy's interface is partitioned into three regions (fig. 5.15, pg. 140). The top of the screen presents users with playback controls and an exit button labelled *Finished*. The tracker board, occupying the screen's central portion, comprises a grid with various numbers of rows and columns, depending upon selected composition complexity. An audio sample library presented as a list of spectrogram images scrolls vertically along the left side. Compositions progress as follows: users touch a tile in the sample library, making it

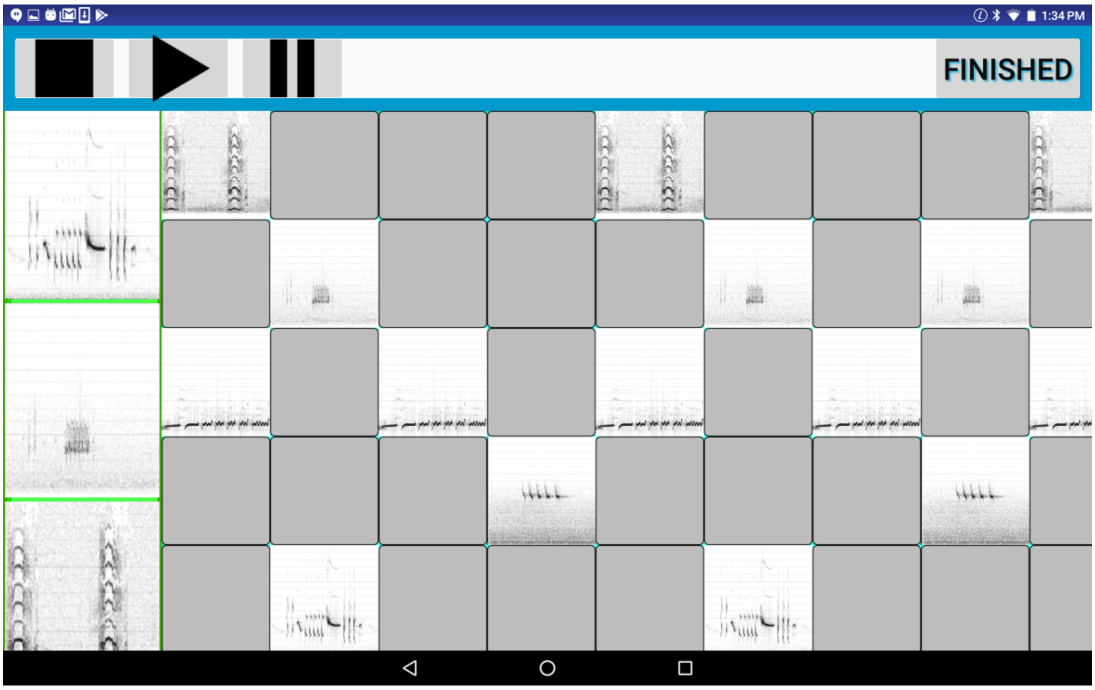

Figure 5.15: An example of the tracker board, the sample library is along the left side and the board is partially filled.

the active sample which can be placed an arbitrary number of times, by tapping, onto the tracker board. Compositions can contain as many instances of as many samples as the user desires within board dimension limitations, with any sample at any location. Long-pressing a tile on the board removes the sample at that location. Samples can be added or removed from the board while playback loops. If the active sample is removed, it will play to completion and be removed for the next loop. Play continues until users lose interest or consider compositions complete.

5.3.5.2 Data collected

When a participant interacts with the ChorusComposer sound toy, a database is populated with information pertaining to the composition (see Appendix A, fig. A.I.4, pg. 253). The database primary key for each composition is the starting timestamp, defined as when the first sample is selected from the library. Login names are passed as database foreign keys, allowing querying of all compositions created by a given user. Within each composition, the number of times players place or remove samples from the board is tracked and elapsed time between such interactions and overall board complexity stored for subsequent analysis.

5.3.5.3 Engagement & motivational support

While music trackers were not assumed to be conceptually familiar to participants prior to interaction with the application, participants were expected to be able to intuit afforded interactions. Magerko et al., exploring student engagement through remixing and looping, albeit with an older target audience, report comprehension and user motivation[143]. ChorusComposer encourages extrinsic engagement with nature through intrinsic engagement with music composition aesthetics. Composing complex mixes of multiple species enables participants to train their ears to distinguish individual species in a natural chorus; this increases participants' capacity to accurately survey in the wild. Building future surveying success motivates participants to engage with nature, increasing autonomy, and thus intrinsic motivation.

5.4 Experimental design, data collection, & evaluation procedures

Research into my interfaces' efficacy for promoting learning and motivating engagement targeted two groups, primary school children and engaged adults. Qualitative and quantitative data pertaining to stakeholder engagement with the Wild Watch were collected. User testing was performed on students across 11 primary schools within the Nidderdale AONB over the course of two week-long sessions, one year apart. Game training-data were updated between the two years to reflect the Wild Watch's evolving needs, reducing the number of species and increasing the variety of calls. In the first year, species image pairs comprised one each of a male and female adult, selected calls were common to both. For the second dataset, images were tied to the specific individuals making calls, be they juvenile, adult male, adult female, or common to adults. I ran three focus groups with adult participants, two in the first year and one in the second. In both years one group was assumed to have prior knowledge of and interest in avian bioacoustics, while in the first year the second group was not. Prior to interacting with the games, student participants completed surveys comprising questions of demographics, prior avian knowledge, and belief about the value of interacting with nature. Survey content evolved between years but maintained similar structure.

In the first year, all participants were instructed to play all games: first memory; as knowledge increased puzzles; and finally open-ended play through composition. The first year dataset comprised the most common utterances of 10 species targeted for HSM within the Nidderdale AONB. After playing, participants completed a post-evaluation survey about prior exposure to educational games, preference for data-representation and goal-state, and belief about the value of games and the potential for games to increase motivation to interact with nature.

In the second year, as the Wild Watch's goals evolved, participants were directed to complete all memory game modes prior to engaging with more complex games as data complexity increased. The dataset comprised known utterances for males, females and juveniles of 4 target owl species in the Nidderdale AONB. Students were tested on their ability to identify species from calls pre- and post-play. Pre- and post-play surveys captured opinions regarding engagement, learning, and whether knowledge increases motivation.

5.4.1 Surveys

Surveys were designed to be answerable by students with reading and writing skills of the target age groups; most students could answer. Questions regarding perceptions of nature were adapted from Royal Society for the Protection of Birds (RSPB) source material for measuring youth perceptions of connections to nature[28]. Additional questions were designed to determine whether perceptions of engagement with nature were influenced by the game interactions, and whether game-play increased perceptions of knowledge of nature.

5.4.1.1 2017 pre-survey

Prior to play, participants were asked their age, their gender, to assess their visual and acoustic avian knowledge, and consider whether both perceptual modes are useful for collecting information relevant to citizen science data-collection surveying (see Appendix A, table A.1, pg. 254 for complete survey). Students ranked agreement with statements pertaining to enjoyment of nature, engagement with wildlife, and desire to increase avian knowledge; results were scored on a 5 point Likert scale, presented as a set of faces. While such a scale has been identified as having the potential for "confusing the emotional continuum of happiness/sadness with the emotional continua of fear/courage and anxiety/confidence and the physical continuum of pain/physical comfort"[37], recent comparative research has shown that when visual representations of faces are treated as a variable augmenting written rankings "response distributions [do] not differ by version ... [and faces] support lower literacy respondents"[212]. For primary school-aged children, the comprehension benefit of a visually-augmented scale justifies its use. Participants were given ∼10 minutes to answer the questions. For those with trouble reading, teaching assistants were present to aid comprehension.

5.4.1.2 2017 post-survey

Having played the games, participants were given ∼15 minutes for another survey (see Appendix A, table A.2, pg. 255 for complete survey). They were asked to quantify prior engagement with educational video games and note their application login name to enable correlation of survey results with performance data. Participants ranked by preference and difficulty each of the three modes of BirdMatch, where data-representation varied, and each of the three games, where goal-state diffuseness increased. Subsequently, participants were given the opportunity to provide open-ended responses to identify what caused confusion during play, and what they liked and disliked about the games. They ranked enjoyment of play and likelihood of continuation with the games, answered questions about whether the games were perceived to contribute to increased visual and bioacoustic avian knowledge, and whether they were motivated to continue learning. Questions about connections to nature were repeated from the pre-survey to analyse whether engagement with games changed these beliefs. Finally, participants judged their confidence in performance for the games with defined goal-states, and identified whether specific goal-states motivated continued play.

5.4.1.3 2018 pre-survey

Second-year surveys were adapted to reflect changes in training content and 2017 survey limitations (see Appendix A, table A.3, pg. 256 for complete survey). Paired questions from pre- and post-surveys were refined, a set of species identification tasks included, and self-reported familiarity numbers constrained to the range of owl species present in the region. As in 2017 surveys, participants provided a baseline for their affinity for nature and identified whether combining senses increased species recognition. Target-species-specific questions were introduced to help determine if game engagement affects learning of specific targets.

5.4.1.4 2018 post-survey

Having played BirdMatch, participants performed shuffled identification tasks from the pre-survey and answered various belief questions (see Appendix A, table A.4, pg. 257 for complete survey). All 7 Likert questions from the pre-survey were replicated with questions pertaining to game enjoyment and perceptions of learning added. For the second deployment, completion of the pre- and post-surveys was allocated $\sim$10 minutes at the beginning and end, the rest of each session was spent exploring the data-representation modes and difficulties of BirdMatch through play.

5.4.2 Primary school experimental procedure

User testing in schools followed similar experimental protocols both years. Each participant was provided with a 7" Android tablet[14] and a pencil for the surveys. Classes were separated into groups of 8 – 15 participants and had ∼50 minutes for the experiment. Upon pre-survey completion, participants were asked to play through a series of games. In the six 2017 schools, where general exposure to the primary calls of 10 target species was desired, protocol allowed participants to progress through BirdMatch data-representation modes and increased goal-state complexity across all games, having satisfied a score threshold. In the five 2018 schools, where the overall project goal was learning detailed call information, a higher score threshold was set for progression through each BirdMatch mode/difficulty combination.

5.4.2.1 Game play protocol

Upon pre-survey completion, the application was projected onto a screen while participants followed along on their devices with the application loaded (fig. 5.16, pg. 144). Participants were instructed to register and open BirdMatch, for experimental purposes labelled *Game 1*. Participants were then instructed to *Choose a Theme*, beginning with *Mode 1* where

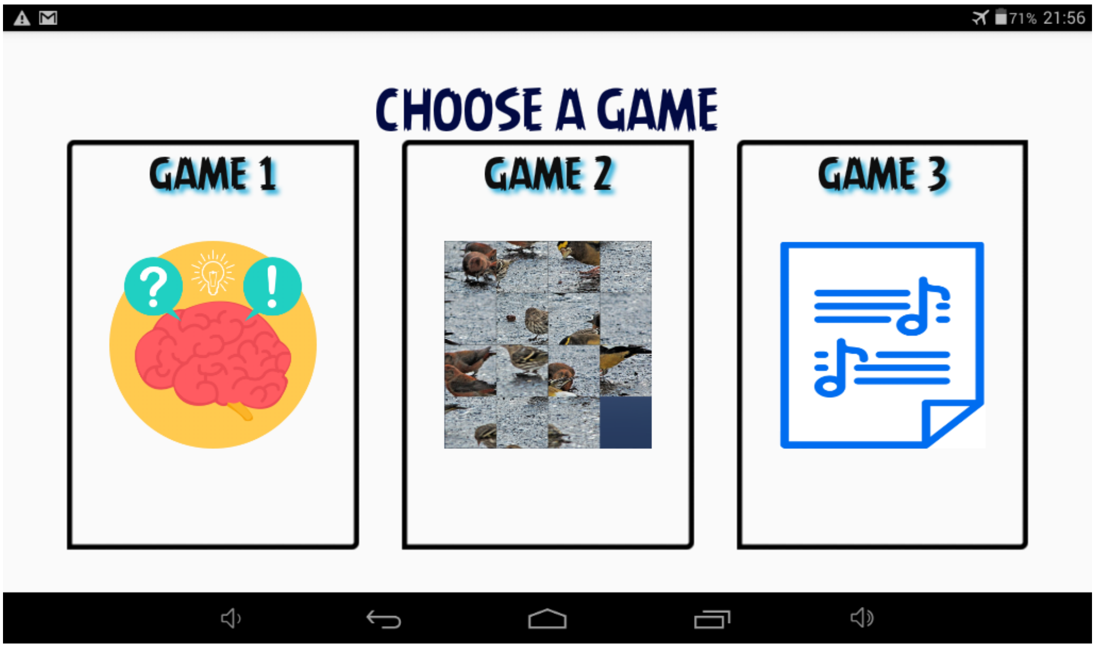

Figure 5.16: The game selection screen; while game icons hint at interactions available, no game names are shown here.

target species images, male and female, were depicted with concurrent auditory feedback (fig. 5.17, pg. 145). Game play began at the easiest level. For the first cohort, 2017,

[14]Following standards for introducing web-enabled technology to classrooms, all applications but mine were blocked.

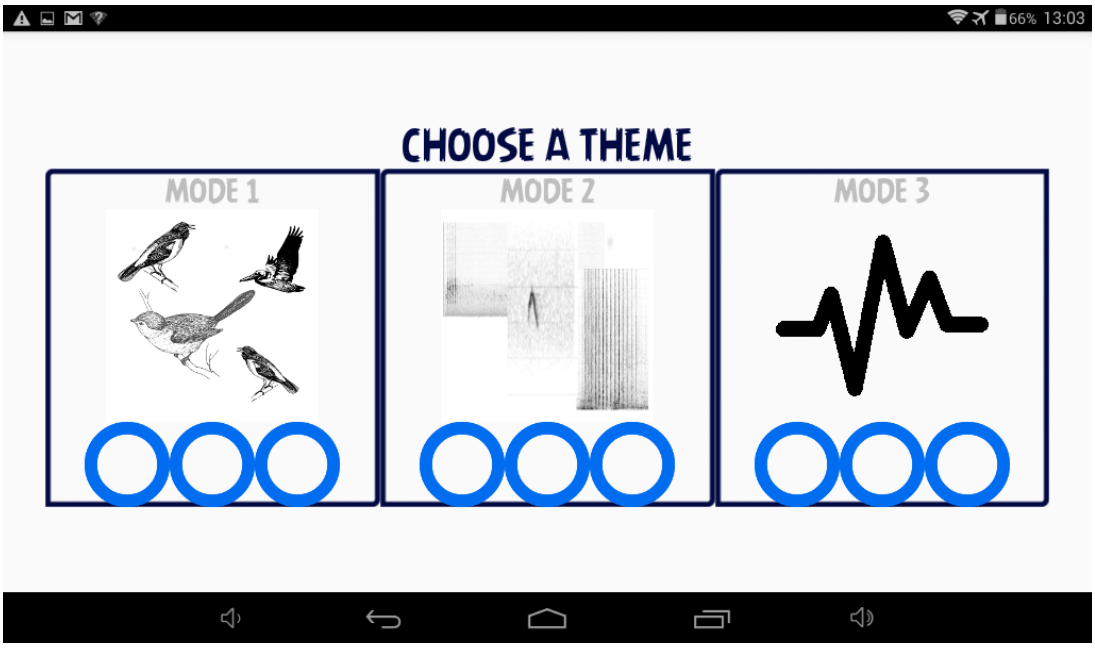

Figure 5.17: The game 1 mode selection screen; while mode icons hint at data-representations presented, no game names are shown here.

the application was designed to prevent players from progressing past difficulty levels or modes until successful completion of each with a score of one star. For the second cohort, 2018, this increased to two stars and participants were constrained when selecting what mode and difficulty to play, ensuring that they started with BirdMatch, as a proportion of the first cohort avoided it entirely. Upon completion to the year's standard of *Mode 1* at any difficulty level, progression to *Mode 2* at an equivalent difficulty was encouraged. In this mode participants matched call spectrogram images with auditory feedback. Upon completion of *Mode 2* to the cohort-dependent standard at each difficulty, *Mode 3*, auditory matching without visual feedback, and the SpectroPuzzle game, labelled *Game 2*, were introduced. Some first-cohort participants started SpectroPuzzle without prior target spectrogram exposure, introduced in BirdMatch *Mode 2*. I provided individuals a demonstration of SpectroPuzzle game-play mechanisms as they progressed. In both years, when participants completed each SpectroPuzzle difficulty to a one-star standard, the ChorusComposer toy, labelled *Game 3*, became available; again, interactions with this toy were demonstrated. Play across all games continued for each group; once a baseline score had been achieved with each game, play was unstructured; play performance data, described in §5.3.3.2, §5.3.4.2, and §5.3.5.2, were recorded and appended to the relevant databases.

5.4.3 Game performance data

While self-perception of increased knowledge resulting from game performance can increase motivation for engagement with nature and further project engagement, validation of survey data quality prior to inclusion in project databases is imperative. In 2018 direct audio-recognition tasks were included in both pre- and post-surveys, analysis provides a measure of short-term learning through play. Long-term learning, for which a web-based follow-up survey — which suffered from complications arising from educational restrictions — was produced, remains to be explored. Participant play performance data were collected for analysis using a touchscreen-suitable extension to the Goals, Operators, Methods, and Selection rules (GOMS) model for interaction performance analysis, Touch-Level Model Goals, Operators, Methods, and Selection rules (TLM-GOMS)[111, 186]. The touch component of my TLM-GOMS model for analysis includes timing information of tactile interactions occurring during game play, given mechanic constraints.

5.4.3.1 Memory game analysis

The memory games, with a single definable goal-state, best lend themselves to TLM-GOMS analysis. During each iteration played, timing of all screen touch actions, enumerated in §5.3.3.2, are recorded for subsequent evaluation. Initially, tile placement on the board is randomised; mode, difficulty, and board-state start data are stored for each iteration played so that optimal solutions can be calculated and performance timing data appropriately scaled. Of initial interest was an upper solution bound, using brute-force depth-first search, for finding matches on a board with n_d tiles (where d is difficulty) and $n_p = \frac{n_d}{2}$ pairs.

$$Expected_{Turns_{max}} = \sum_{x=0}^{x=\frac{n_d}{2}} \frac{n_d - (1 + x)}{2} \tag{5.1}$$

Actual user behaviour being less predictable than brute-force, some participants managed to exceed this limit. Therefore I instead compute success by: considering how rapidly participants found the solution; measuring errors; discounting luck. It remains beyond the scope of this thesis to distinguish whether increased matching success correlates with visual learning, or results from participants using pattern-matching techniques, such as spot-the-difference.

I wrote a Python[15] script to analyse each BirdMatch iteration played from interactions stored to the database upon game completion, as follows:

$$Winning(W) = \frac{Turns_{min}}{Turns_{taken}} \tag{5.2}$$

[15]https://www.python.org

The ratio of minimum number of turns possible (given selected difficulty this is $\frac{Num_{cards}}{2}$) to actual turns taken, is the baseline winning rate.

$$Luck(L) = \frac{Pairs_{random}}{Pairs_{total}} \tag{5.3}$$

Luck diminishes learning outcomes as pairs found without prior exposure to either tile do not reflect player intentionality. The number of pairs found by luck is the number of pairs found without either paired card having been previously seen. If more pairs are found by luck, maximum potential learning success decreases although winning would still occur and gamified results — stars and scores — are unchanged.

$$Learning_{error}(L_e) = \sum_{x>4}^{\infty} X_{seen} \tag{5.4}$$

If the list of selected species IDs is ordered, for pair of cards ID_1, ID_2 depicting a species, an ID that appears only twice will necessarily have been found by luck. If an ID appears 3 or 4 times, then each card in the pair, once found, is matched. If additional instances of an ID occur, a species' card has been forgotten and turned again; I classify this memory error as learning error, L_e.

$$LearningSuccess(L_s) = \left(\frac{(W)(1-L)}{E}\right) \tag{5.5}$$

From W, L, and learning error (L_e), I generate a success metric (L_s) for each iteration played; L_s and L_e allow me to analyse the efficacy of games for participant learning and to map learning success progressions over games played. These analyses support examination of processes by which short-term learning occurs during iterative game play and whether gamified rewards correlate with increased motivation for engagement. As these analyses do not capture data necessary to quantify long-term learning, for the second cohort I included a follow-up section[16] to the project website for participants to again complete the species identification tasks from the pre- and post-surveys; however, poor uptake yielded insufficient responses for analysis.

I subsequently define an exploration quotient metric classifying the degree to which participants exhibit motivation for either repetitive learning or game exploration.

$$Exploration_{quotient}(E_q) = \left(\frac{Variants_{Explored}}{Iterations_{Played}}\right) \tag{5.6}$$

This quotient ranges from 0 to 1, with the caveat that those for whom additional games were blocked due to poor initial performance, who played more iterations than the 9

[16] http://bioacoustic.games/html/owl_survey.html

available data-representation/difficulty combinations, inevitably duplicated some play, their desire for further exploration notwithstanding, thus failing to maximise E_q.

5.4.3.2 Puzzle game analysis

Although multiple solutions exist for the puzzle games, the solution characteristics are sufficiently constrained to warrant collection of touch data for a TLM-GOMS analysis. Timing of each touch on the puzzle board is collected during play, as described in §5.3.4.2. Tile order is initially randomised, a mapping of samples to board coordinates is stored and tracked for each swap; as an expected upper bound based on random swapping is neither relevant nor tractable, I instead calculate the optimal number of steps to solve each board. I wrote scripts analysing database output of participants' game interactions to determine a user goal-state attainment metric, as follows:

$$Turns_{min}(T_{min}) = \left(\frac{Tiles_{board} - \sum_{rows} Tiles_{max}}{2}\right) \tag{5.7}$$

Given the arbitrary board row/species' spectrogram relationship for a valid solution, in this algorithm $Tiles_{max}$ is calculated from the row for which a species is maximally present in a randomly set board. By finding the rows that are closest to completion at outset, the number of swaps to solution is minimised. Comparing the minimum number of swaps to a solution with participant performance provides a winning efficiency metric (W_ϵ) for learning outcomes.

$$Winning_{Efficiency}(W_\epsilon) = \left(\frac{T_{min}}{T_{taken}}\right) \tag{5.8}$$

This performance score permits quantifying the efficacy of puzzle games for enhancing participants' knowledge and ability to comprehend spectral data-representations. While I contend that increased game performance is correlated with learning, this metric does not offer the capacity to distinguish between short- and long-term learning from play.

5.4.3.3 Composition toy analysis

Unlike the previous games, ChorusComposer does not have a goal-state. The primary engagement datum of interest is time users spend playing. Time between additions and deletions supports further analysis of engagement with the interface. An ID for each sample is stored for incorporation into a composition complexity metric. For evaluation purposes, timing data were collected for each insertion and deletion of a sample to and from the board, yielding a board density metric. Additional engagement data is the number of tracker insertions and deletions performed by participants during play, as a function of time; combined these yield composition engagement/complexity metrics as follows:

$$Composition_{Engagement} = Composition_{time} * Turns_{taken} \tag{5.9}$$

Initial composition engagement is computed as a product of time spent in the process of composing and the number of actions performed in this time.

$$Board_{density}(B_d) = \left(\frac{Sample_{insertions} - Sample_{deletions}}{Board_{dimensions}}\right) \tag{5.10}$$

However, composition complexity is also a function of board density, which can be tracked by identifying the number of active samples on the board relative to board dimensions.

$$Composition_{complexity}(C_c) = \frac{B_d}{Samples_{unique}} \tag{5.11}$$

From board density, a more nuanced view of composition complexity requires counting the number of unique library samples used in the composition.

$$Engagement(E) = (C_c) * Time_{scaled} \tag{5.12}$$

Equations (5.9), engagement as a function of time and turns, and (5.12), engagement as a function of complexity and time, support discussion of whether sound toys with spectrogram data-representations effectively motivate participant interaction.

5.4.4 Focus groups

While my research output focusses on student engagement with and learning of avian bioacoustics, the Nidderdale AONB includes multiple stakeholders with vested interests in local avian populations, some of whom have significant local knowledge to contribute to collaborative designs. Both years' focus groups were invited by the Wild Watch organisers to identify desired project contributions and to explore and critique my designs for knowledge development games. In 2017 I guided a focus group of ∼20 participant stakeholders invested in local land management, including farmers, gamekeepers, landowners, and representatives of the Nidderdale AONB and the Game and Wildlife Conservation Trust (GWCT)[17]. I further collaborated on a focus group to introduce my games as tools for engaging musicians with birdsong to ∼12 members of a local community choir. Both groups' participants were asked, in semi-structured open-ended discussions, to consider roles they might play in developing tools for conveying their knowledge and interest in birdsong to the broader public. Audio and video recordings of these discussions were

[17]https://www.gwct.org.uk

made for subsequent transcription and annotation; results are discussed in chapter 6. The questions guiding discussion presupposed that participants anticipated contributing to citizen science and considered mechanisms to mitigate motivational problems associated with a lack of cyclical information flow.

In 2018, I participated in a focus group targeting ∼25 adult citizen scientists who had participated in training and data collection with the Wild Watch during its first year. Through open-ended discussion participants reflected on the games as training tools. These discussions allowed participants to query me and the Wild Watch project coordinator. Participant critiques of project protocols and my games triggered a positive feedback loop in subsequent development.

These adult participants varied widely in age. The local farmers and gamekeepers ranged from 20's to 60's while those who joined as interested musicians were predominantly retirees. Those in the second year, who had previously engaged with the Wild Watch training, were primarily middle-aged, with some retirees.

5.4.4.1 Citizen science interest & familiarity

In all three focus groups participants were asked to clarify their familiarity with the term 'citizen science', and whether they considered crowd-sourcing environmental data a viable means of contributing to conservation projects. Follow-up questions pertained to whether participants had prior knowledge of, and had participated in, such projects. While not all participants in the second year self-identified as citizen scientists, clarification remained relevant. A discussion of how they felt about participating in citizen science data-collection and analysis projects followed. Finally, participants were asked to consider how they might contribute as citizens to science.

5.4.4.2 The value of birdsong

Participants were guided to discuss whether they considered birdsong something that they wanted to preserve in their environment. This covered questions of whether they considered hearing birdsong to be an active or passive activity and whether they identified personal value associated with hearing birds. Participants were encouraged to discuss whether acoustic diversity influenced their engagement with nature. The first group further discussed the degree to which birdsong has static classification ground-truth, or whether it is sufficiently variable that consistent stakeholder knowledge-transfer is difficulty or untenable.

5.4.4.3 Roles for those self-representing as engaged stakeholders

Finally, participants were encouraged to discuss the degree to which they felt that they are stakeholders in AONB preservation. This led to discussions of stakeholder conflict in the region, and what is meant by supporting preservation of a politically-defined entity (an AONB), as opposed to preservation of the region regardless of nomenclature. While the first 2017 focus group comprised stakeholders who self-identified as engaged, that year's second group and the second year group provided insights into issues arising with defining engaged stakeholders when even residents of a region who participate in a project may so self-define. All groups were concerned that mechanisms be in place to extend engagement beyond data collection to analysis and subsequent policy formulation that will result from the Wild Watch project.

5.5 Research explorations: learning, motivation, & collaborative design

My research explores whether play motivates children's engagement with nature through games, whether game performance correlates with learning, and how collaborative design can increase the success of both for citizen science. I introduce games designed to increase avian bioacoustic knowledge and enhance motivation for students and others in an affected geographic community including amateur ornithologists, ecologists, and gamekeepers. Through these games I examine whether data-representation should be novel or familiar, how varying goal-state complexity and rewards during play supports intrinsic and extrinsic motivation, and whether learning and creativity enhance motivation, as posited in §2.5.2. I identify whether sound visualisation provides a useful and viable approach to teaching avian bioacoustics and whether spectral representations increase the scientific potential of generated data.

5.5.1 Questions regarding learning

While motivating participation is prerequisite for engaging citizen scientists, collected and analysed data cannot support broader scientific needs unless there is a measure of confidence in knowledge. I compare students' claimed knowledge and self-reported confidence in knowledge to quantitative outcomes measured against ground-truth. I further examine whether educating participants with previously unfamiliar data-representations increases baseline knowledge. I conclude by examining learning in the context of less constrained goal-states.

5.5.1.1 Research exploration: confidence in claimed knowledge

Are participants confident that my games increase their avian knowledge across data-representations? I investigate participant confidence in prior avian knowledge by sight and sound, whether play increases confidence, and whether confidence is well-founded. These are discussed in light of claimed confidence in learning by sight and sound from the games and a comparison of self-reported avian knowledge pre- and post-play. Results of self-reported claims for the 2018 cohort will be compared with direct call identification results pre- and post-play.

5.5.1.2 Research exploration: effects of data-representation on learning

Are learning effects representation-dependent and will learning be greatest for the most familiar data-representation, presumably bird images? Furthermore, is sound sufficient for knowledge development or do spectral visual representations better augment learning? This will be studied by comparing learning effects across BirdMatch data-representation modes and evaluating transitions between learning metric results with repeated play.

5.5.1.3 Research exploration: learning with relaxed goal-state constraints

Are learning effects present when goal-state constraints are relaxed or do my puzzles only reinforce prior learning? This question will be examined by searching for learning effects through iterative play of SpectroPuzzle. If found, these effects support development of such games to enhance spectrogram understanding.

5.5.2 Questions regarding motivation

Considering my citizen science project participation motivation framework proposed in §2.7.2 (fig. 2.8, pg. 59), I explore how games enhance engagement. Of interest is whether games support both intrinsic and extrinsic motivation through introjected, *i.e.* knowledge development, and external, *i.e.* gamified, rewards. I examine participants' prior exposure to and interest in educational games and desire for knowledge about and engagement with nature. I discuss how varying rewards and data-representations affect participant motivation and whether games contribute intrinsic motivation or primarily provide extrinsically motivating interactions with commensurate limitations. I examine whether my games are sufficient for motivating engagement with citizen science projects, whether games enhance project engagement and engagement with nature, and how game design can encourage participants to segue between stages in the project cycle introduced in §5.1.1. I explore how various rewards resulting from diverse ludic interactions with bioacoustic data enhance motivation as interaction constraints are reduced. Having observed whether games augment engagement, I explore whether engagement increases data quality as a function of learning. Participants who achieved flow in game-play are identified and their divergent motivations discussed.

5.5.2.1 Research exploration: engagement with educational games

I begin by identifying a baseline for prior engagement with educational games and explore whether my games yield further declared interest. Assessment comprises: a discussion of the 2017 cohort's prior exposure to educational games; their reported enjoyment of the games post-play; and their desire to continue playing and learning about birdsong. These results provide insight into whether my educational games enhance engagement.

5.5.2.2 Research exploration: interest in nature and avian bioacoustics

I subsequently explore whether baseline interest in nature and desire for engagement with wildlife is affected by my games. Assessment results from discussing reported interest in nature and whether motivation to interact with and gain knowledge about nature changes with game exposure.

5.5.2.3 Research exploration: how varying data-representation familiarity & goal-state complexity affect motivation

Does engagement with less-constrained goal-state games motivate learning, or is prior data-representation familiarity prerequisite for success? I explore relationships amongst game enjoyment, data-representation familiarity, and degree of perceived challenge. This will be examined from results of the first cohort's game and representation preference responses and declared motivation and confidence in winning.

5.5.2.4 Research exploration: motivating learning through play

Is motivation to play through the data-representation and difficulty variants in BirdMatch age-dependent and does learning success vary as different trajectories are pursued through the variants? If no age dependency exists, then my designs are suitable for motivating students across tested key stages. Do the later game iterations played continue to provide motivation and do participants who find flow still learn? I present results of the degree to which participants explore the game space and of engagement with the final sound toy. I examine divergent ludic behaviours by subsets of participants whose flow states present extremes of exploration and engagement.

5.5.3 Questions regarding collaborative design

I explore how collaboratively designing games promotes the creation of more effective projects and how supporting various contributions enhances diverse stakeholders' engagement. Through focus group conversations with local, regional, non-governmental organisation (NGO), and governmental stakeholders, I identify how my designs support institutional needs as well as individuals' desires for knowledge acquisition. I examine stakeholders' roles as participant collaborators, and the feedback they desire to validate their contributions.

5.5.3.1 Research examination: familiarity with & concerns about avian citizen science

Do focus group participants desire to support avian conservation citizen science and do they believe my designs simultaneously teach effectively and motivate participation? I will summarise and critique participant responses to questions of conceptual familiarity with citizen science, value they associate with avian utterance-recognition training games, and use of participant output by the Wild Watch project.

5.5.3.2 Research examination: does collaborative design yield more engaged stakeholders

Do focus group participants desire to contribute to and think that their contributions can influence project outcomes through games designed to benefit individual and project goals? This will be assessed by summarising and critiquing focus groups' responses regarding perceived roles for citizen scientists in conservation projects, their underlying interest in the preservation of birdsong in the local environment, and their self-representation as engaged stakeholders.

CHAPTER 6

Evaluating games for citizen science

Over the course of two weeks, one year apart, I administered surveys to and collected data from games played by 242[1] primary school children in 11 schools, selected by the Wild Watch coordinators, in the Nidderdale Area of Outstanding Natural Beauty (AONB), following procedures introduced in §5.4.2.

Survey questions are described in §5.4.1.1 – §5.4.1.4 and listed in §A. Results were transcribed to spreadsheets, encoded as .csv, and output for analysis in R[2]. Survey results tabulated for each cohort are presented, when available, with comparisons pre- and post-play. When cohort tasks differed, independent analyses are performed. Game play data are analysed for motivation and learning using metrics introduced in §5.4.3.1, §5.4.3.2, and §5.4.3.3. I perform detailed analyses of participants with unique motivational behaviours selected from each cohort. In addition, I coordinated three focus groups involving ∼60 adult participants who discussed potential citizen science contributions to the Wild Watch project, interest in birdsong, and the AONB, and proposed collaborative design contributions to my games, as introduced in §5.4.4.

6.1 Background results

Prior to play, after an introductory presentation by the Wild Watch project coordinators, each cohort's participants were asked to complete their respective pre-surveys from §5.4.1.1 and §5.4.1.3. Upon pre-survey completion, participants explored semi-structured open-ended play, afterwards they completed the post-surveys presented in §5.4.1.2. and §5.4.1.4. As I iteratively developed the surveys across the two years to reflect project evolution,

[1] 241 produced at least partial survey responses.
[2] https://www.r-project.org

my analyses include, when relevant, a determination of whether the two cohorts can be considered to have been drawn from the same population.

6.1.1 Demographics

Both cohorts' pre-surveys asked participants to provide their ages[3] and gender[4]. First cohort participants ranged in age from age 4 to 11 ($\mu = 9.1, \tilde{x} = 9, \sigma^2 = 2.2, N = 136$), second cohort participants ranged in age from age 5 to 11 ($\mu = 7.8, \tilde{x} = 8, \sigma^2 = 3.0, N = 105$). The two cohorts were selected from the same key stages, as defined by the United Kingdom (UK) national curriculum. As age is categorical and results include ties, a Wilcoxon's rank sum test with continuity correction, assuming sample independence and similar variance, yielded $p = 6.10 * 10^{-9}$, supporting the null hypothesis that age distributions are sufficiently similar to warrant treatment of both cohorts as drawn from the same population (fig. 6.1, pg. 156). When subsequent analyses show divergent participant

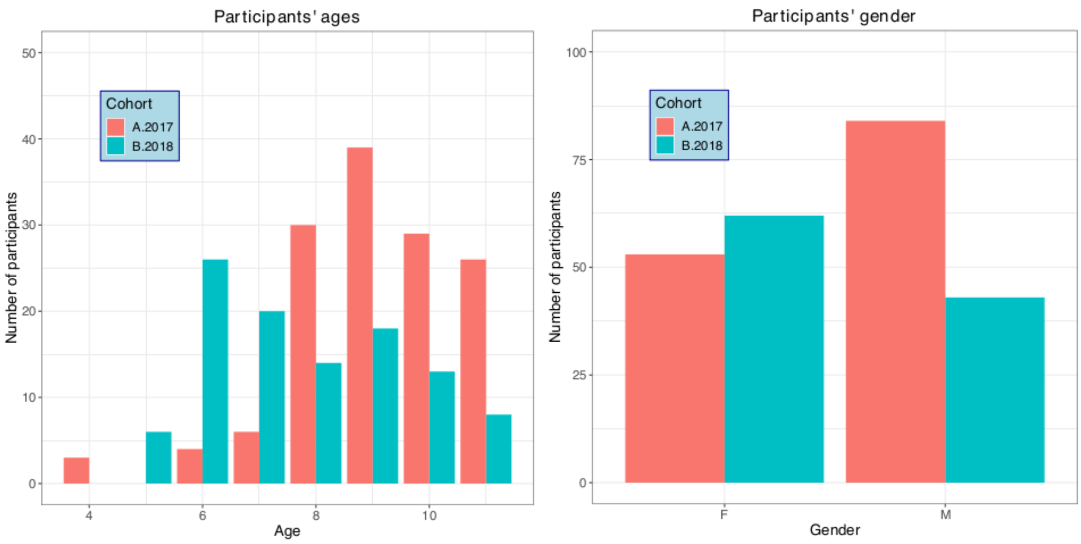

Figure 6.1: Left figure presents participant age distributions for each cohort. Drawn from the same key stages, the two groups can be assumed to come from the same population. The right graph presents participants' gender by cohort. The two groups differ significantly.

behaviour between cohorts, this is independent of age biases. Gender distributions differ significantly: the first cohort was predominantly female (Female=84, Male=52, N=136), the second cohort biased male (Female=43, Male=62, N=105). When my analyses show cross-cohort consistency, this occurs gender biases notwithstanding.

[3] 2017 pre-survey Q.2, 2018 pre-survey Q.2.
[4] 2017 pre-survey Q.3, 2018 pre-survey Q.7.

6.1.2 Baseline affinity for nature

Four identical initial Likert questions pertaining to engagement with nature were asked of both cohorts in their respective pre-surveys, the second cohort was again asked these questions post-play. Responses are compared across cohorts to see whether the two populations, despite gender variation, feel similarly. Likewise, responses between pre- and post-surveys for the second cohort are examined for effects from play.

6.1.2.1 Evaluating: 'I enjoy spending time in nature'

Participants somewhat strongly agreed with the statement 'I enjoy spending time in nature'[5], but for the second cohort this did not change significantly with play (see §B.I for complete results). The second cohort's stated enjoyment of nature insignificantly decreased with play. This result reflects risks associated with mediating engagement with nature through technological interfaces which tautologically motivate interface engagement, to the detriment of environmental engagement, as discussed in §5.1.2. Baseline enjoyment of spending time in nature was not expected to change with a single session of game interaction.

6.1.2.2 Evaluation: 'when I am outdoors, I notice the wildlife around me'

Participants generally agreed with the statement 'when I am outdoors, I notice the wildlife around me'[6]; however, the 2018 cohort showed no significant shift in response with play. Detailed results can be found in §B.II. Cross-cohort comparison provides insight into whether participants' previous engagement with nature, prerequisite for citizen science data collection, was similar. Second cohort changes in response after game play determine whether interface engagement influences predicted outdoor engagement. Both cohorts identified as likely to notice wildlife whilst outdoors. The negligible mean belief drop post-play is insignificant, so no change in baseline belief is associated with play; this is unsurprising as baseline affinity for nature is unlikely to be changed by a single session of game interactions.

6.2 Evaluating learning

In this section I will investigate whether my games enhance knowledge development by querying participants' prior claimed knowledge, their confidence in said knowledge, and whether they believe the games build knowledge. I will present and discuss results from the ∼1400 total BirdMatch iterations played, examining effects of data-representation on

[5] 2017 pre-survey Likert Q.L1, 2018 pre-survey Likert Q.L1 / post-survey Likert Q.L7.
[6] 2017 pre-survey Likert Q.L2, 2018 pre-survey Likert Q.L2 / post-survey Likert Q.L8.

learning and how participants across cohorts engage with the memory training games, given different datasets. For the second cohort, pre- and post-play owl identification tasks provide empirical results of knowledge acquisition; while long-term retention was queried through a web-based follow-up survey, insufficient participants submitted results for analysis. I will conclude by presenting SpectroPuzzle play results, examining learning when goal-state constraints are relaxed. Across all three SpectroPuzzle difficulty variants, participants played ∼200 game iterations for analysis. Unlike BirdMatch, for which a single solution exists, SpectroPuzzle requires abstract comprehension of the types of target boards which satisfy goal-state requirements, various winning paths exist.

In response to the research question from §5.5.1.1 I will present results showing participants were confident that my games developed visual and audial knowledge across data-representations. In response to the research question from §5.5.1.2 I will show that overall BirdMatch learning metrics increased across the first three iterations of play regardless of data-representation although effect strength declined after initial exposure. In response to the research question from §5.5.1.3 I will show that SpectroPuzzle learning was evident across all three difficulties. This last result is attributable, at easy difficulty, to learning game-mechanics rather than increased data-representation comprehension, but it implies learning at harder difficulties.

6.2.1 Research exploration: confidence in claimed knowledge

I investigate whether participants are confident about their prior acoustic and visual avian knowledge, whether play increases confidence, and whether confidence is well-founded. Participants across both cohorts were asked questions pertaining to prior avian knowledge by sight and sound and whether perception by either supported species identification. For the 2018 cohort an auditory identification task was presented pre- and post-play, results inform whether the games aid short-term learning and recall. Participants are generally confident that my games increase their avian knowledge across data-representations. Participants consistently agreed, slightly strongly, that the games helped visual and audial learning of birds. Observed call identification accuracy increases indicate that my games have training value for citizen science projects.

6.2.1.1 Evaluating: 'When identifying birds I use both sight & sound'

Answers to pre-survey questions as to whether participants considered using sound and sight for avian identification[7] showed consistent bias for combining visual and acoustic identification (fig. 6.2, pg. 159): 63.4% of the first cohort (N=124) and 65.7% of the second (N=102) responded affirmatively. These baseline results support my contention that sound

[7] 2017 pre-survey Q.7, 2018 pre-survey Q.10.

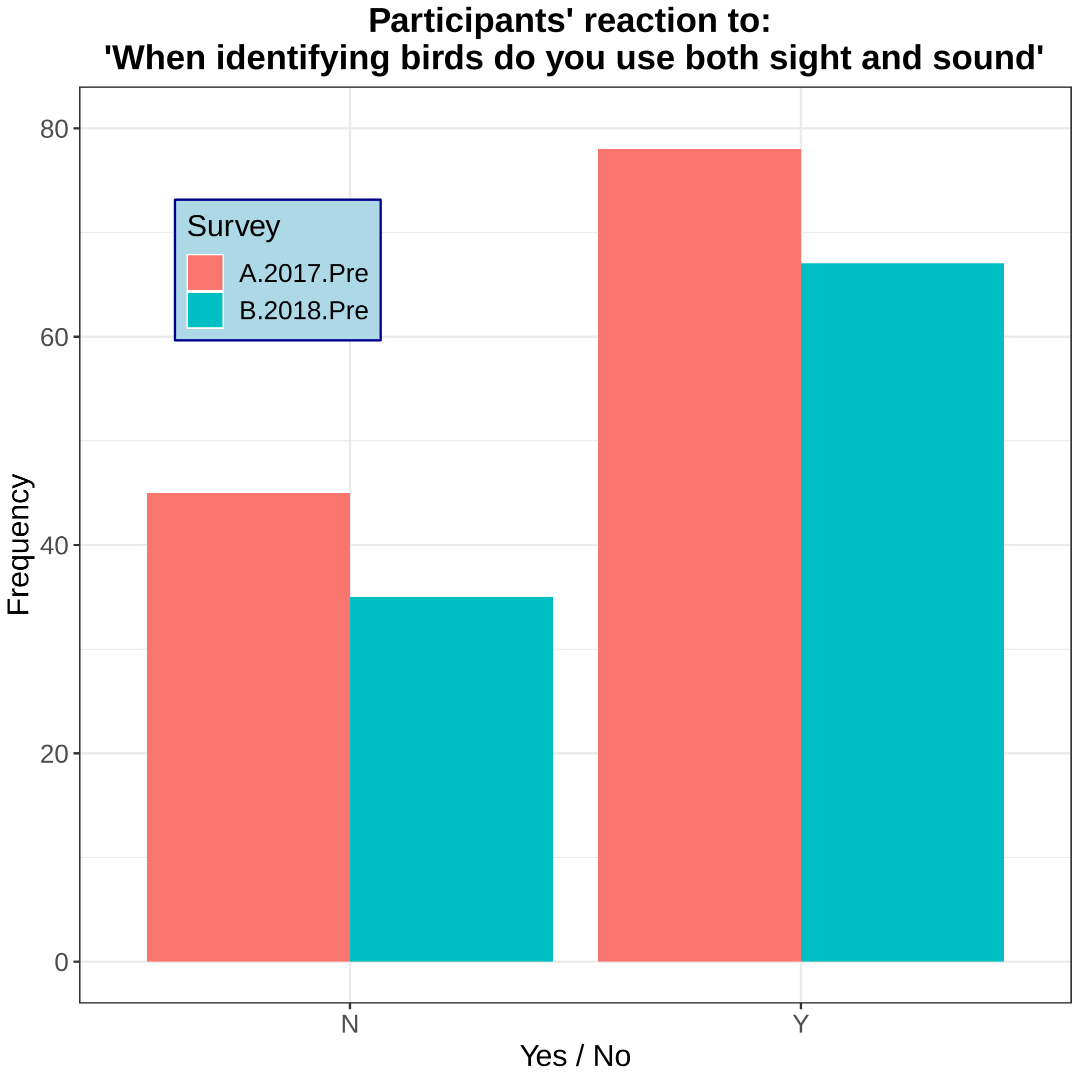

Figure 6.2: Participants' use of both sight and sound for avian identification. In both cohorts a preponderance replied in the affirmative.

is valuable when training citizen scientists. My games introduce content knowledge to participants who had not previously thought to use both senses for identification.

6.2.1.2 Evaluating: 'the games helped me to learn birds/owls by sight/sound'

Participants consistently agreed, slightly strongly, that the games helped visual and audial learning of target avian species. Both cohorts were drawn from the same population, and an increased perception of audial learning was not significant. Post-play, participants provided Likert responses to the statement 'the games helped me to learn birds/owls by sight'[8]. While my games were designed to explore whether sound and visual representations of sound enhance avian identification training for citizen scientists, participants' visual identification comfort provides a basis for comparison. The median response for the 2017

[8] 2017 post-survey Q.L3, 2018 post-survey Q.L2.

cohort was $\tilde{x} = 4$, for the 2018 cohort $\tilde{x} = 5$. Mean responses were similar, with summary statistics for the first cohort ($\mu = 4.10$, $\sigma^2 = 0.75$, N=97), the second showed marginally stronger belief that my games aided visual learning, albeit with greater variance ($\mu = 4.13$, $\sigma^2 = 1.34$, N=68). For comparison between cohorts, a one-tailed Wilcoxon's signed rank

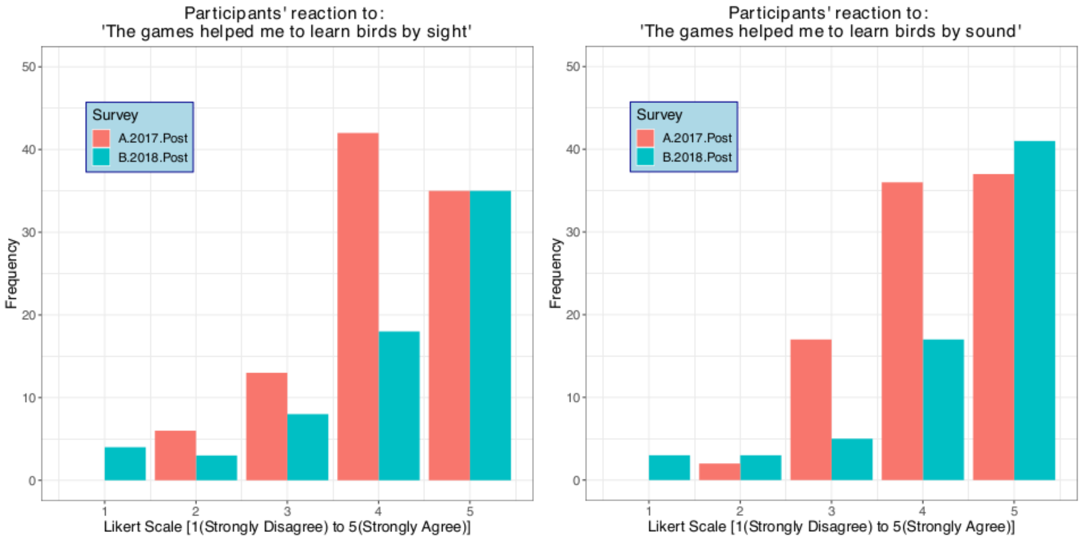

Figure 6.3: Histogram of Likert-scaled responses as to whether the games aided visual and auditory learning. Results support the use of both sight and sound for training.

test determined that both cohorts can be assumed to have been randomly selected from the same representative population, $p = 0.98$.

Participants in both cohorts were asked to respond to the statement 'the games helped me to learn birds/owls by sound'[9]. When compared with sight results, sound results aid in deciding whether games provide a basis for training participants to recognise birds exclusively acoustically. The median response for both cohorts was $\tilde{x} = 4$. The second cohort believed slightly more strongly ($\mu = 4.30$, $\sigma^2 = 1.16$, N=69) that the games helped auditory learning, albeit with greater variance, than the first cohort ($\mu = 4.17$, $\sigma^2 = 0.65$, N=93). Results of a one-tailed Wilcoxon's signed rank test indicate that the cohorts have been randomly selected from the same population, $p = 0.99$. While across both cohorts, the games were perceived to aid mean sound learning somewhat more that visual, for neither was this significant, $p_{2017} = 0.58, p_{2018} = 0.82$. As the games were perceived to enhance both visual and audial knowledge, I posit that learning success will be greatest when a combination of senses is used. These results support using visual depictions of sound, *e.g.* spectrograms, for training.

[9]2017 post-survey Q.L4, 2018 post-survey Q.L3.

6.2.1.3 Evaluating participants' prior avian knowledge by sight & sound

Both cohorts were asked questions of visual and audial avian familiarity, the second cohort responded pre- and post-play[10]. While prior visual familiarity is consistently higher than audial familiarity, in neither cohort was this difference significant. Both cohorts can be considered to have been drawn from the same population, and in no instance was knowledge through one sense significantly different than through the other. The first cohort's avian familiarity question encompassed all birds, with insignificantly stronger belief in prior visual knowledge (fig. 6.4, pg. 161). Answers were smoothed to range [1 –

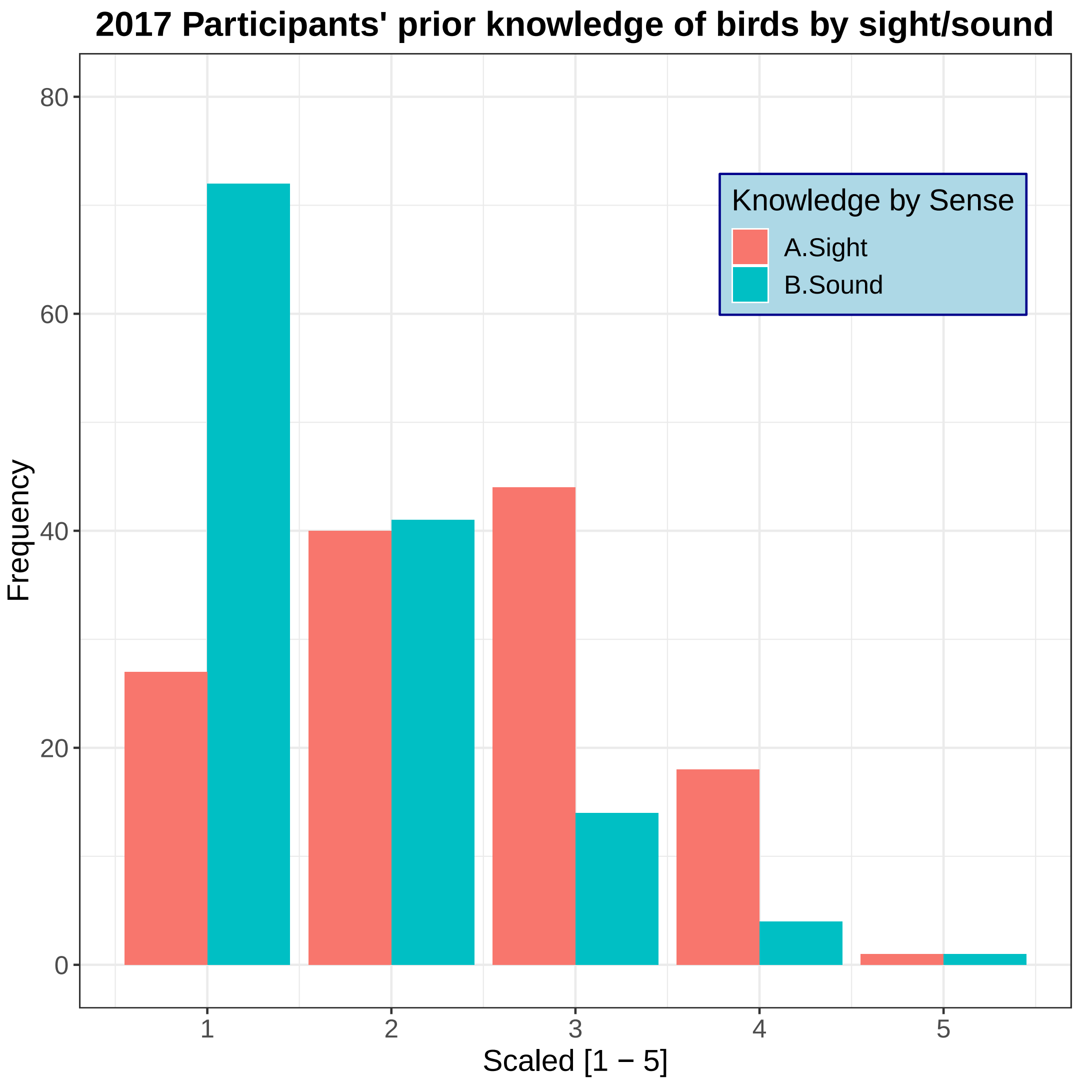

Figure 6.4: A comparison of the first cohort's self-reported knowledge of birds by sight and sound. Scores (1-5) are either mapped from text responses or natural logarithm scaled from numeric responses, as described in the text.

5] by computing the natural logarithm of the stated number of birds known, rounded to the nearest whole number. Those who knew 3 or fewer birds scored 1, up to 7 inclusive

[10] 2017 pre-survey Q.5/Q.6, 2018 pre-survey Q.8/Q.9 and post-survey Q.7/Q.8.

scored 2, up to 20 scored 3, under 55 scored 4, and 55 or more, approximately 10% of birds found in the UK, scored 5. For those who gave written answers, *few* scored 1, *not many* scored 2, *some* scored 3, *a lot* & *many* scored 4, and *most in the UK*[11] scored 5.

For the 2018 cohort, the range of answers was 0 - 4+; 4 target species of owl are taught through the games while 6 owl species may be familiar to students in the UK. To eliminate written answers, participants were instructed to circle an appropriate number. Results were similarly distributed for both cohorts, with insignificantly stronger prior belief in visual knowledge. Pre-play the mean number of owl species claimed known by sight was $\mu = 2.15$ ($N = 104$) and by sound, $\mu = 1.54(N = 98)$; these rose post-play to mean claimed sight recognition $\mu = 2.67$ ($N = 49$) and sound recognition $\mu = 2.31$ ($N = 48$) (fig. 6.5, pg. 162). Wilcoxon's test results examining second cohort changes over play in

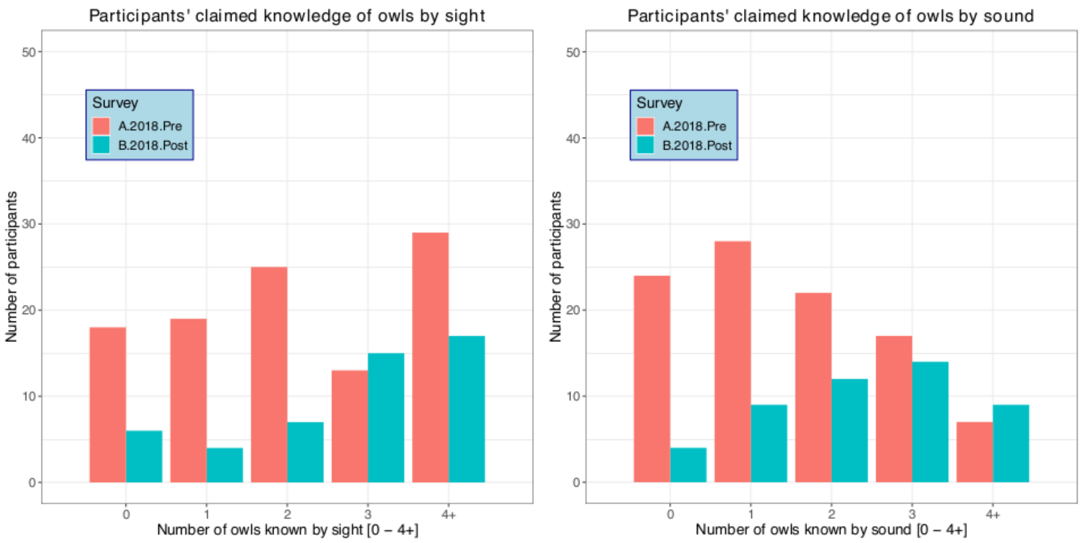

Figure 6.5: A comparison of the second cohort's self-reported knowledge of owls by sight and sound. Neither median shifts are statistically significant despite a notable rise in mean results.

perceived knowledge by sight or sound do not support that play elicits a significant shift in perception ($p_{sight} = 0.99, p_{sound} = 0.999$); for neither cohort was one sense significantly more familiar than the other (2017 p=1, 2018 pre-play, p=0.99, post-play p=0.999).

6.2.1.4 Evaluating: 'I am confident that I could recognise some of the birds/owls from the games by sight/sound'

Both cohorts marginally believed that they could recognise the taught species by sound; they significantly more strongly believed in their future visual species recognition. Participants in both cohorts were asked their responses to the statement 'I am confident that I could recognise some of the birds/owls from the games by sight'[12]. Both cohorts generally

[11] Just one student stated this, in later discussion with the participant and their teacher the participant identified familiarity with over 100 avian species.

[12] 2017 post-survey Q.L11, 2018 post-survey Q.L11.

agreed that the games contributed to avian visual knowledge. Median response were $\tilde{x} = 4$ for the 2017 cohort and $\tilde{x} = 5$ for the 2018 cohort. Aligned with results from §6.2.1.2, these responses qualify the validity of the games for training participants to recognise birds without device mediation. Summary statistics for first cohort responses were $\mu = 4.17, \sigma^2 = 0.87$, N=90, the second cohort showed insignificantly stronger belief that games provided training for future recognition, albeit with greater variance: $\mu = 4.26, \sigma^2 = 1.43$, N=69.

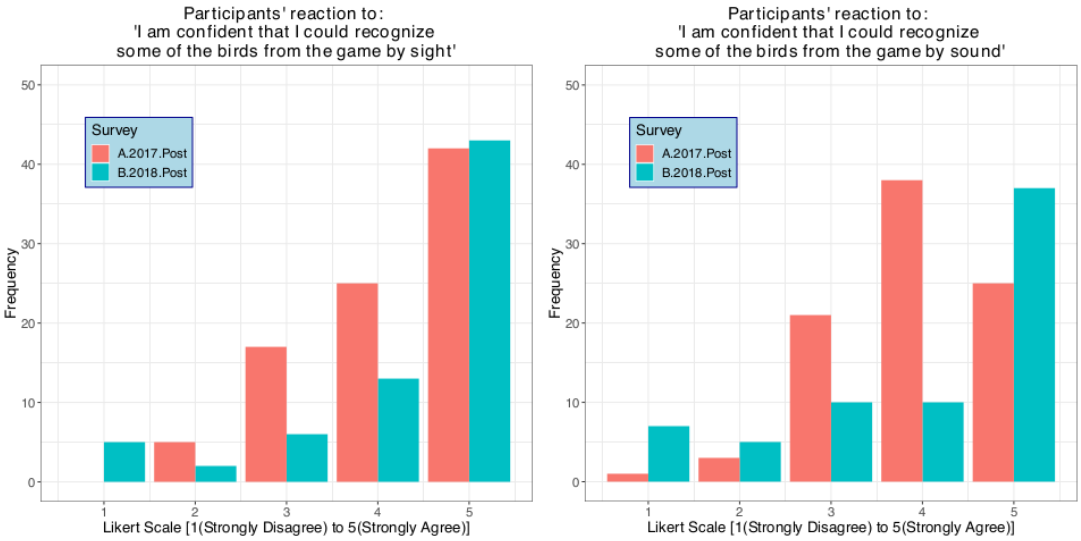

Figure 6.6: Histogram of Likert-scaled responses as to whether game training is perceived to transfer knowledge to the field for visual or auditory identification. Generally, participants are more likely to trust visual learning but do agree with my contention that the games provide useful auditory training.

Participants in both cohorts were asked their responses to the statement 'I am confident that I could recognise some of the birds/owls from the games by sound'[13]. Both cohorts agreed that the games supported acoustic learning. These results, considered in light of results from §6.2.1.2, show that my games provide perceived auditory training for participants to recognise birds without device mediation. The median response was $\tilde{x} = 4$ for the 2017 cohort and $\tilde{x} = 5$ for the 2018 cohort. First cohort summarised responses were $\mu = 3.94, \sigma^2 = 0.77$, N=85; identical mean responses for the second cohort varied more: $\mu = 3.94, \sigma^2 = 1.91$, N=69. Both cohorts' participants are significantly more likely to trust visual learning (2017: $p = 1.45 * 10^{-5}, 2018 : p = 8.66 * 10^{-4}$), and agree that the games provide valuable audio training through spectral data-representations.

6.2.1.5 Evaluating changes in identification accuracy with play

To test empirically whether self-reported knowledge claims were exaggerated, participants in the 2018 cohort were asked to identify the sound in a set of 4 recordings used in the

[13] 2017 post-survey Q.L12, 2018 post-survey Q.L12.

games before and after play (fig. 6.7, pg. 164). In neither pre- nor post-play conditions

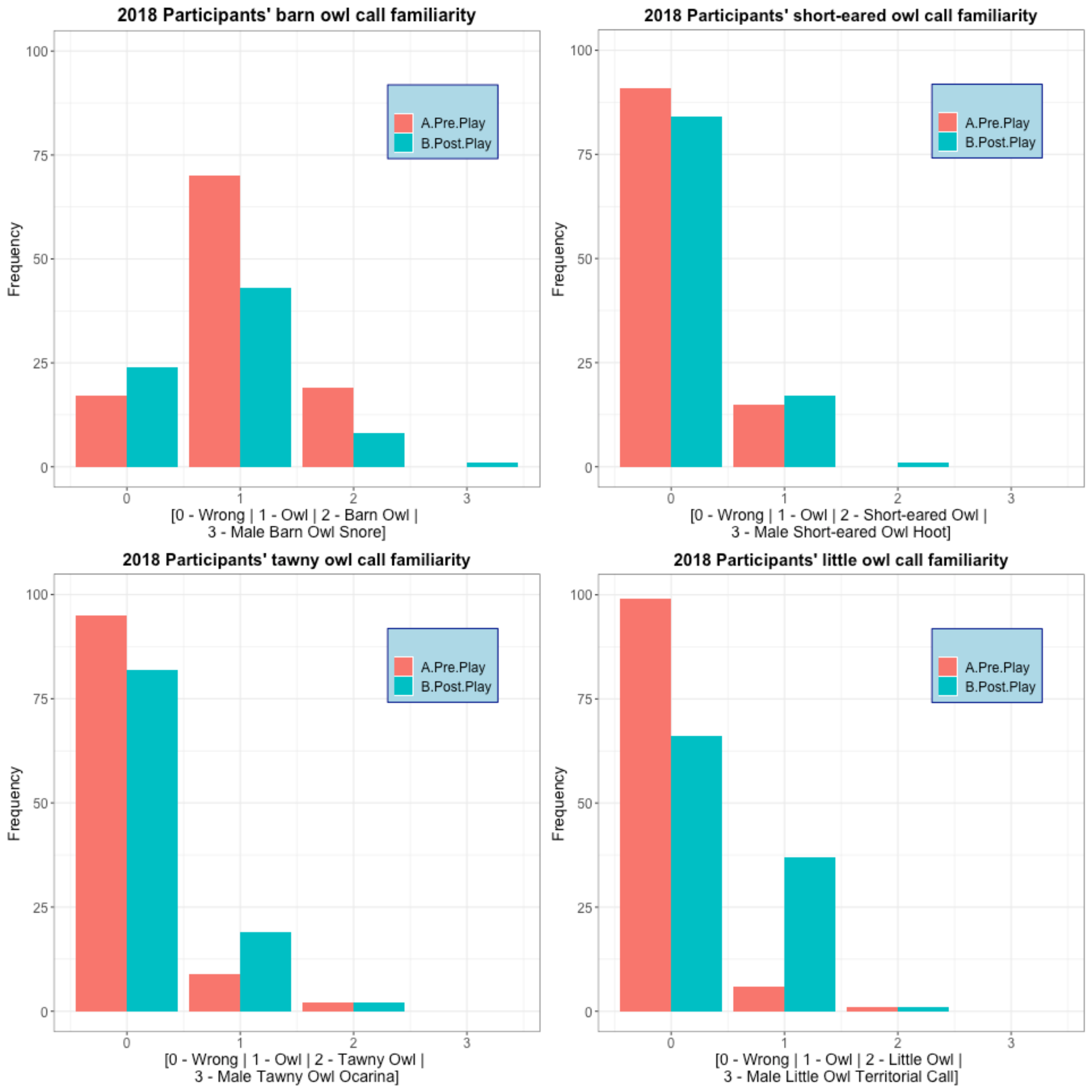

Figure 6.7: A comparison of the second cohort's attempts to identify recordings of calls from the target owl species pre- and post-play.

were participants informed that the sounds were of owls. Between the first and second auditioning of recordings participants spent one session engaging with my owl games, and one with Planet Birdsong (PBS) discussing a broader set of species and their spectrograms. Results were scored as follows: an incorrect response or 'unknown' scored 0, identifying an owl scored 1, naming the owl species scored 2, and identifying the owl species and sex or call type scored 3. The barn owl pre-play mean familiarity score ($\mu = 1.02, N = 106$) fell significantly post-play ($\mu = 0.82, N = 76, p = 0.025$) indicating that barn owl call familiarity confidence decreased with play and exposure to the broader teaching of the Planet Birdsong programme. This anomalous fall likely results from confusing participants as they transitioned from expecting owls hooting to a more nuanced understanding of utterance variety, combined with exposure to numerous other avian utterances. The short-eared owl pre-play mean familiarity score ($\mu = 0.14, N = 106$) rose slightly but insignificantly post-play ($\mu = 0.19, N = 102, p = 0.83$). Mean tawny owl pre-play familiarity scores ($\mu = 0.12, N = 106$) nearly doubled, albeit insignificantly ($p = 0.97$), post-play ($\mu = 0.22, N = 103$), overall utterance familiarity remained marginal. The little owl pre-play mean familiarity score ($\mu = 0.08, N = 106$) rose post-play ($\mu = 0.38, N = 104$) but the change did not reflect significant learning in the student population ($p = 0.99$). For all but the barn owl, participant familiarity increased, albeit insignificantly. Learning results from a single session of play are below a useful threshold ($\mu = 2$) to identify species. However, observed call identification accuracy increases indicate that my games have training value for citizen science projects. Differences between claimed owl call familiarity (§6.2.1.3) and tested call identification ability may result from brief training on unfamiliar calls.

6.2.1.6 Discussion & conclusions: confidence in claimed knowledge

Results from §6.2.1.1 support my contention that sound adds value to training for citizen science data collection. My games introduce this knowledge acquisition approach to the roughly one third of participants who had not previously thought to use both senses for identification. While prior visual familiarity is consistently higher than audible familiarity, in neither cohort was this difference significant. Continued audio training though play helps with learning as while both cohorts marginally believed that they could recognise the taught species by sound post-play they significantly more strongly believed in their future visual species recognition. The games were perceived to enhance visual and audial learning, therefore I posit that learning success will be greatest when a combination of senses is used. This supports the use of visual depictions of sound, *e.g.* spectrograms, for training. Results do not show that a single session of play elicits a significant shift in perception of knowledge by sight and sound. Generally, participants are more likely to trust visual learning but do agree that the games provide useful auditory training when spectrograms are introduced. Changes in call identification accuracy, positive for all but the barn owl, indicate that my games have training value for citizen science projects. Differences between claimed owl call familiarity and tested call identification ability likely result from the use of previously unfamiliar utterances. Both sight and sound should be incorporated into training materials as there is more room for learning with sound since participants brought less prior knowledge.

6.2.2 Research exploration: effects of data-representation on learning

Are learning effects representation-dependent and will learning be greatest for the data-representation presumed most familiar, an avian image? My analyses consider metrics introduced in §5.4.3.1, specifically for success, (5.5), and error, (5.4). Aggregate results for each cohort across the first three iterations of play are presented for each data-representation mode at level easy; complete representation/difficulty results are presented in §C.I. How data-representation affects learning effects will be analysed from trajectories of aggregate interstitial metrics for the two transitions between the first three iterations of play. I find that, while all three data-representations augment knowledge development, the first cohort, with the simple data-set, learns most with the first representation, while the second cohort, exposed to more complex auditory content, learns better from spectral data-representations.

6.2.2.1 Generating analysis metrics for identifying learning from BirdMatch

Databases populated on devices during play, introduced in §5.3.3.2 and enumerated on A.I.2, pg. 251, having been updated to incorporate a per-user UID field, were queried via sqlite3. Card selection order for each game played, ordered by game-start timestamp, by a user for a given data-representation/difficulty combination, was output to a .csv file (fig. 6.8, pg. 166). A python script parses output files for each UID/mode/difficulty

```
16:32:36$ sqlite3 ./Kirkby_Malzear_device_7_bridgeowls.db
SQLite version 3.24.0 2018-06-04 14:10:15
Enter ".help" for usage hints.
sqlite> .mode csv
sqlite> .output ./Analysis_Match_Game/UID_001/001_MM1D1.csv
sqlite> SELECT cardSelectedOrder FROM match_game_data WHERE UID="1" AND themeID="1" AND difficultyLevel="1"
ORDER BY gameStartTimestamp ASC;
sqlite> .quit
```

Figure 6.8: Extracting ordered turn data from the database for each iteration of a data-representation/difficulty combination played by a given participant.

combination, computing metrics proposed in §5.4.3.1 (fig. 6.9, pg. 167). Success and error metrics for each iteration of play for every representation/difficulty combination were computed for each UID, when existent, inserted into a spreadsheet encompassing metrics for all games played, output by iteration played to .csv, and imported into R for analysis and visualisation. These metrics provide a basis for discussing learning through play. Data summarising the number of iterations of each combination played as participants followed various trajectories through the game space support subsequent discussion of engagement through play.

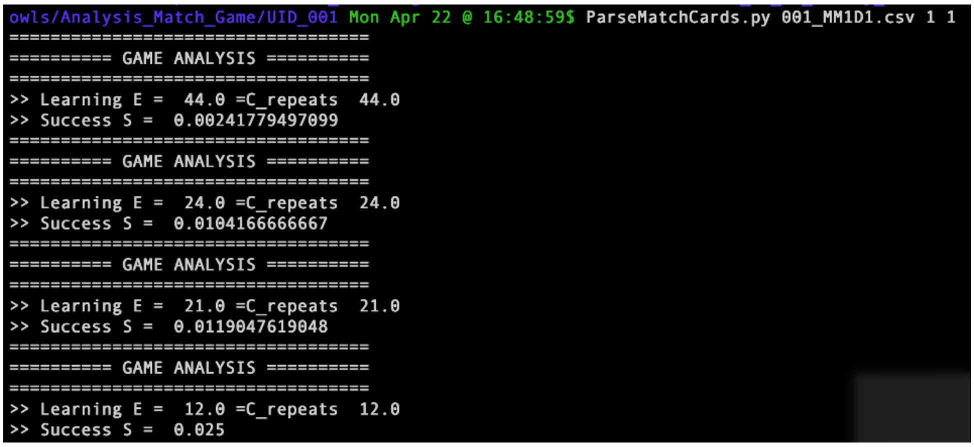

Figure 6.9: Script output reporting results of time-ordered analysis metrics for repeated games played by a given participant at a single data-representation/difficulty combination.

6.2.2.2 Learning metric results

The number of times a given participant played each data-representation/difficulty combination varied widely with divergent trajectories of play. Learning through play is observed when success scores rise and error-rates correspondingly fall between sequential iterations of data-representation/difficulty combinations. Play success/error results of all data-representation/difficulty variants wherein multiple participants played are in §C.I. Here I compare success- and error-rates across the initial three iterations of play for each data-representation at level easy. Results expose the potential for gameful interactions to enhance citizen science learning.

Both the first cohort, trained on single utterances of 10 target species, and the second, trained on detailed utterances of 4 target owl species, saw broadly decreasing mean error and correspondingly increasing mean success (fig. 6.10, pg. 168) over the first three iterations of play across data-representations. Success (5.5) is not solely a function of error (5.4), as luck (5.3) can play a confounding role. Validating whether increased success reflects decreased error is relevant (fig. 6.11, pg. 169). While overall learning metrics increased, across the first three iterations of play, regardless of data-representation, declining learning effects after initial exposure warrant discussion.

Summary statistics for the first cohort's success scores for the first three iterations of play ($I_1 \dots I_3$) with each data-representation are presented in table 6.1, pg. 170, no transitions are significant. The 2017 cohort played a total of 100 first, 47 second, and 19 third iteration games at the easy level for the first data-representation mode. For the first data-representation, neither change in L_s between $I_1 \rightarrow I_2$ (Wilcoxon's $p = 0.96$), nor $I_2 \rightarrow I_3$ ($p = 0.44$) is significant. First representation mean L_s increases across iterations while the rate of increase falls and median L_s falls marginally once game-mechanic familiarisation occurs. The first cohort played a total of 46, 20, and 5 repetitions

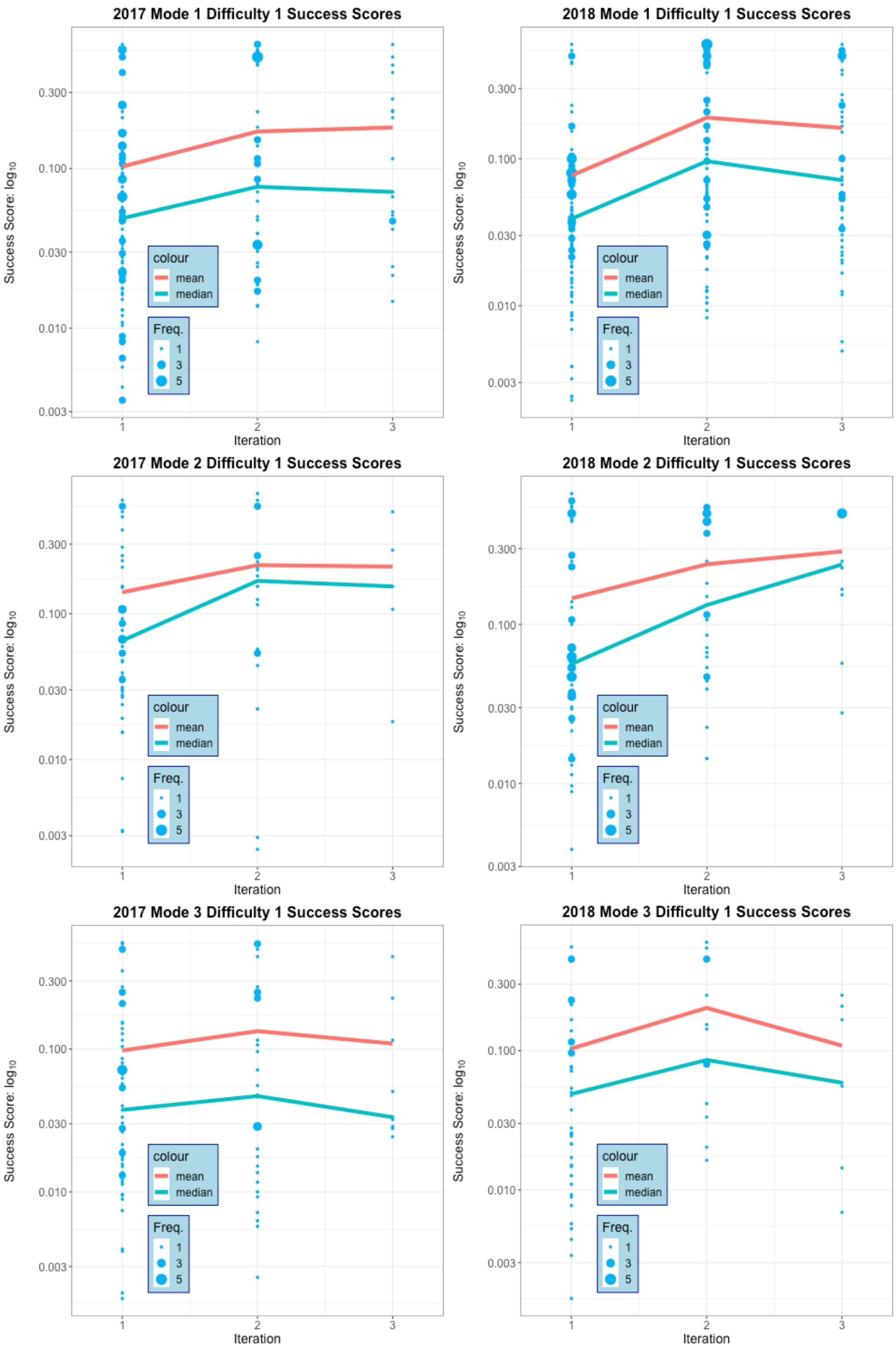

Figure 6.10: Comparison of both cohorts' success scores across the first three iterations played with each data-representation at level easy. The viable success metric range is from 0 to 1, results are presented log_{10}-scaled.

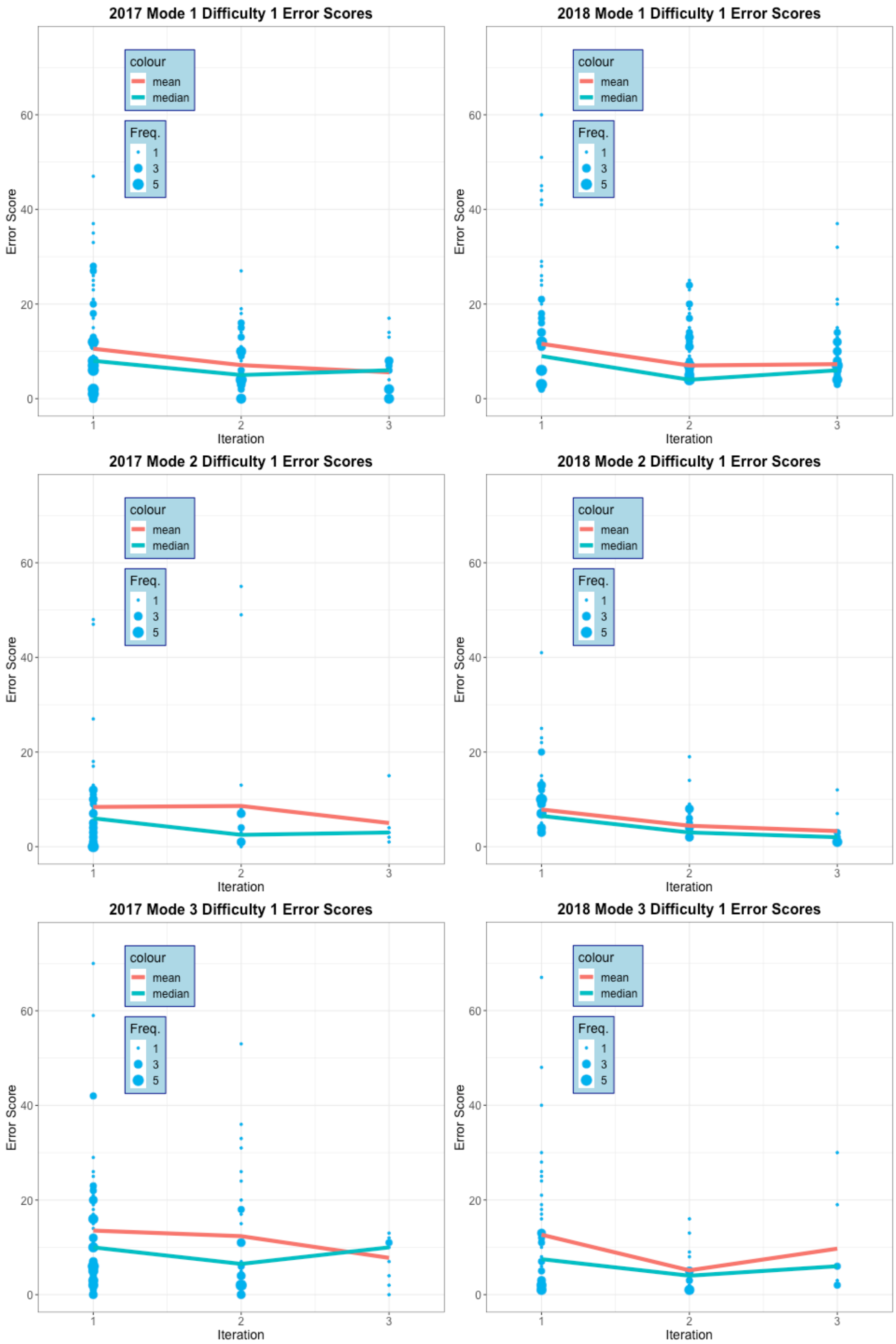

Figure 6.11: Comparison of both cohorts' error scores between the first three games played in the easy mode for each of the three data-representations. There is no fixed upper bound on possible participant errors. With repetitive play error-rate variance fell overall, albeit least for the third data-representation.

2017 L_s summary statistics, level easy, first three iterations of play									
Representation	1^{st} Iteration			2^{nd} Iteration			3^{rd} Iteration		
	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2
Audio & Avian Image	0.103	0.049	0.020	0.171	0.077	0.040	0.181	0.071	0.034
Audio & Spectrogram	0.140	0.065	0.030	0.215	0.168	0.044	0.210	0.154	0.035
Audio Only	0.098	0.038	0.019	0.134	0.047	0.030	0.109	0.033	0.020

Table 6.1: L_s summary statistics for the 2017 cohort playing the first three iterations of each data-representation mode at level easy.

across the first through third iterations with the second data-representation's easy level. For the second data-representation neither change in L_s between $I_1 \rightarrow I_2$ ($p = 0.98$), nor $I_2 \rightarrow I_3$ ($p = 0.58$) is significant. With this data-representation, initial rising mean and median L_s rates leading into the second iteration of play were followed by a slight $\tilde{x}_{L_s}$ decrease, while μ_{L_s} plateaued in the $I_2 \rightarrow I_3$ transition. For the third data-representation, level easy, 59 first iteration, 30 second, and 9 third iteration games were played. For this data-representation neither change in L_s between $I_1 \rightarrow I_2$ ($p = 0.75$), nor $I_2 \rightarrow I_3 (p = 0.36)$ is significant. L_s metric drops after the second iteration are likely a result of small sample size in the third iteration, as the few participants who played this iteration performed below cohort averages in earlier iterations.

Second cohort L_s summary statistics for the first three iterations of play with each data-representation are presented in table 6.2, pg. 170, no transitions are significant. This

2018 L_s summary statistics, level easy, first three iterations of play									
Representation	1^{st} Iteration			2^{nd} Iteration			3^{rd} Iteration		
	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2
Audio & Avian Image	0.077	0.039	0.014	0.190	0.096	0.041	0.162	0.071	0.032
Audio & Spectrogram	0.146	0.057	0.035	0.239	0.133	0.041	0.288	0.239	0.038
Audio Only	0.103	0.049	0.019	0.203	0.086	0.042	0.109	0.059	0.009

Table 6.2: L_s summary statistics for the 2018 cohort playing the first three iterations of each data-representation mode at level easy.

cohort played 99, 75, and 53 repetitions of the first three iterations, respectively, with the first data-representation. For the first data-representation, neither change in L_s between $I_1 \rightarrow I_2$ ($p = 1.00$), nor $I_2 \rightarrow I_3$ ($p = 0.13$) is significant. With this data-representation, a plateau in mean and fall in median L_s after initial learning from exposure suggest either that this representation is insufficiently interesting to maintain participant focus or that repeated images of the same birds, corresponding to different utterances, are overly challenging. This cohort played 66 first iteration games, 28 second, and 10 third with the second data-representation. For this data-representation neither change in L_s between $I_1 \rightarrow I_2$ ($p = 0.99$), nor $I_2 \rightarrow I_3$ ($p = 0.76$) is significant. The spectral data-representation provides consistent increases in both mean and median L_s, supporting my contention that the spectral representation, being visually unique for each utterance, provides additional dimensions of information relevant and of interest to citizen science trainees. In the third mode, 44 repetitions of the first iteration, 15 of the second, and 7 of the third were played. The few participants who played a third iteration performed below cohort averages in earlier iterations. Neither third data-representation transition is significant, with Wilcoxon's test statistic for the population shift between $I_1 \to I_2$, $p = 0.99$, falling to $p = 0.07$ for $I_2 \to I_3$. An alternative interpretation is that the first transition rises as a result of focus, while the second falls as a result of frustration because this data-representation provides no visual aid to auditory learning.

First cohort L_e summary statistics for the first three iterations of play with each data-representation are presented in table 6.3, pg. 171, the first data-representation, α-transition shows significant error reduction. The first cohort saw decreased L_es across all data-representations for the easy mode between the first three iterations played, correlating with previously reported rising success scores. For the first data-representation the first

2017 L_e summary statistics, level easy, first three iterations of play									
Representation	1^{st} Iteration			2^{nd} Iteration			3^{rd} Iteration		
	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2
Audio & Avian Image	10.55	8	87.46	7.09	5	40.41	5.58	6	25.37
Audio & Spectrogram	8.39	6	100.3	8.6	2.5	231.3	5	3	32.5
Audio Only	13.54	10	185.9	12.37	6.5	171.0	7.78	10	22.44

Table 6.3: L_e summary statistics for the 2017 cohort playing the first three iterations of each data-representation mode at level easy.

transition is significant, with Wilcoxon's test statistic for the population shift between $I_1 \to I_2$ ($p = 0.02$), falling to insignificance ($p = 0.27$) for $I_2 \to I_3$. In the third iteration of the first data-representation, the median rise is offset by falling standard deviation, mean error continues to decrease. Unlike for corresponding success scores, L_e variance falls across both iterations as outliers continue to learn. For the second data-representation, neither change in L_e between $I_1 \to I_2$ ($p = 0.17$) nor $I_2 \to I_3$ ($p = 0.63$), is significant. As $\tilde{x}_{L_e}$ fell from 6 to 2.5 between the first two iterations, the rise in μ_{L_e} is disproportionately affected by high-error outliers in the second iteration, indicated by increased variance; this stabilises with the second transition as variance drops. For the third data-representation, neither change in L_e between $I_1 \to I_2$ ($p = 0.41$) nor $I_2 \to I_3$ ($p = 0.13$), is significant. Overall, L_e variance decreases as participants learn. Falling L_e closely tracks previously presented L_s increases, irrespective of data-representation.

Second cohort L_e summary statistics for the first three iterations of play with each data-representation are presented in table 6.4, pg. 172; for all data-representations, α-transitions show significant error reduction. The first cohort generally produced lower mean L_e across all data-representations for the easy mode between the first three iterations. Investigation

2018 L_e summary statistics, level easy, first three iterations of play									
Representation	1^{st} Iteration			2^{nd} Iteration			3^{rd} Iteration		
	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2	μ	$\tilde{x}$	σ^2
Audio & Avian Image	11.6	9	122.6	7.01	4	45.93	7.30	6	54.75
Audio & Spectrogram	7.85	6.5	49.79	4.43	3	19.37	3.3	2	12.68
Audio Only	12.66	7.5	188.4	5.13	4	20.84	9.71	6	114.9

Table 6.4: L_e summary statistics for the 2018 cohort playing the first three iterations of each data-representation mode at level easy.

into whether similar results appear when the first data-representation is less strongly tied to the underlying content, *e.g.* the second cohort's owl call dataset, follows. For the first data-representation, the $I_1 \rightarrow I_2$ L_e reduction is significant, Wilcoxon's $p = 1.0 * 10^{-3}$, showing learning between the first two iterations of play; this falls to insignificance, $p = 0.70$, for $I_2 \rightarrow I_3$. Despite overall falling L_e with this data-representation, the L_s issue, attributed to confusion surrounding repeated bird images corresponding with different utterances, continues here with initial learning effects partially reversing in the third iteration of play. For the second data-representation, the first transition shift in L_e is significant, Wilcoxon's test statistic for the population shift between $I_1 \rightarrow I_2$ is $p = 1.0 * 10^{-4}$; the $I_2 \rightarrow I_3$ transition is insignificant, $p = 0.27$. Significant $I_1 \rightarrow I_2$ L_e reduction for the spectral data-representation supports my contention that when each bird is correlated with a single utterance the bird image supports superior knowledge development, when bird images are associated with multiple utterances, spectral data-representations become powerful learning tools. For the third data-representation the $I_1 \rightarrow I_2$ L_e shift again is significant ($p = 2.0 * 10^{-3}$); this falls to insignificance ($p = 0.93$) for $I_2 \rightarrow I_3$. With this data-representation, initial learning from primary exposure is subsequently confounded by the intrinsic difficulty of ear-training without prior knowledge. The games do support an initial reduction in auditory identification errors. The spectral data-representation is best suited to the owl dataset, as both visual and audio learning are supported. Issues arising when teaching multiple calls associated with specific bird images are avoided.

6.2.2.3 Evaluating learning transitions

Rising L_s-transitions between iterations of play demonstrate learning effects between iterations of the same data-representation/difficulty variant. I propose an overall learning success trajectory metric (σ_τ), computed from L_s transitions over the first three iterations where sufficient play occurred for each data-representation, averaged across difficulty; success-score medians, less affected by outliers than means, best represent population learning shifts. For each game iteration, ($I_n; n \in [1, 3]$) played at each difficulty for each data-representation, the median L_s, ($\tilde{x}_{I_n}$) is determined; α and β success transitions are then computed for each data-representation/difficulty variant as ratios of medians between

$\tilde{x}_{I_{1\to2}}$ and $\tilde{x}_{I_{2\to3}}$. The arithmetic mean of difficulties, for which enough games were played with each data-representation, produces $\mu_{\tilde{x}_\alpha}, \mu_{\tilde{x}_\beta}$; these represent instantaneous learning trajectory scores across the first and second transitions for each data-representation and, when greater than 1, imply learning. The overall L_s trajectory metric (σ_τ) is then computed across the first two instantaneous learning trajectories as:

$$\sigma_\tau = \left(\frac{\mu_{\tilde{x}_\beta}}{\mu_{\tilde{x}_\alpha}}\right) \tag{6.1}$$

A $\sigma_\tau > 1$ results from median play L_s increasing with repeated play and demonstrates a learning effect across iterations for a given data-representation.

Detailed results are presented in §C.I.4, showing mean and median transition values for each data-representation/difficulty variant as well as computed success evolution scores. I present instantaneous learning trajectory scores ($\mu_{\tilde{x}_\alpha}$, $\mu_{\tilde{x}_\beta}$) as well as overall learning success trajectory metrics (σ_τ) for each cohort and data-representation (see table 6.5, pg. 173). The mean of the 2017 cohort's $\mu_{\tilde{x}_\alpha}$ across all data-representations was 2.05, the

Instantaneous and overall learning scores with each data-representation by year						
Representation	2017			2018		
	$\mu_{\tilde{x}_\alpha}$	$\mu_{\tilde{x}_\beta}$	σ_τ	$\mu_{\tilde{x}_\alpha}$	$\mu_{\tilde{x}_\beta}$	σ_τ
Audio & Avian Image	1.72	1.06	0.62	1.70	1.29	0.76
Audio & Spectrogram	2.54	0.91	0.36	1.52	1.80	1.18
Audio Only	1.89	0.71	0.38	3.59	0.69	0.19

Table 6.5: Metrics reflecting learning are blue. The first data-representation presents the most consistent path to continued learning across the first three iterations of play for both cohorts. However, when presented with the more nuanced owl dataset, the second cohort learned more strongly in the second transition with the spectral data-representation than with any other. Generally, continued play correlates with increased knowledge development. Except for the second cohort's interactions with the spectral representation, learning rates dropped after initial exposure; this result supports my contention that spectrograms provide both ease-of-learning and relevant additional content for training on complex utterances.

mean of $\mu_{\tilde{x}_\beta}$ was 0.89. While first instantaneous learning, expected from initial exposure, is positive, this is not immediately continued as a baseline for subsequent knowledge acquisition has been reached when data are simple and mechanics familiar. In the second cohort, the mean of $\mu_{\tilde{x}_\alpha}$ was 2.27, the mean of $\mu_{\tilde{x}_\beta}$ was 1.26. This demonstrates that when taught data increases in complexity instantaneous learning continues to rise over multiple iterations of play. The first cohort, learning a simplified dataset, showed most consistent learning with the data-representation presumed to be most familiar, the avian image. The second cohort likewise continued to learn with this data-representation. First cohort learning with the spectral representation continued, albeit at a reduced rate, in the second transition, supporting my contention that this representation provides relevant training content even with simple data. For the second cohort, instantaneous ($\mu_{\tilde{x}_{\alpha,\beta}}$) and overall ($\sigma_\tau$) learning success trajectory metrics were stronger with the less familiar spectral data-representation, supporting the contention that spectrograms are suitable for augmenting visual knowledge of ambiguous sounds. In both cohorts, the α learning transition is strong with the third data-representation, presumably due to initial data-representation unfamiliarity. This does not translate to immediate subsequent learning, as "repetition *per se* does not provide a basis for the improvement of performance" in ear-training[147].

6.2.2.4 Discussing effects of data-representation on learning

Each of the data-representations in both years produced some significant α-transition instantaneous learning, showing that no representation is overly complex for the target age group (see table in §C.I.4.5 for complete results) For the first cohort, α-transitions for all L_e changes at all difficulties were significant for the first data-representation ($p_{\alpha_e} = 0.02, p_{\alpha_m} = 2.37 * 10^{-3}, p_{\alpha_h} = 3.25 * 10^{-3}$), thus for the simple data-set images sufficed, regardless of set size. While for the second cohort, these transitions remained significant for the easy ($p_{\alpha_e} = 7.45 * 10^{-4}$) and medium ($p_{\alpha_m} = 0.03$) levels, transition L_e reduction was not significant at the hardest level, likely a result of high data-set image repetition with multiple calls. For the first cohort, the second data-representation α-transition L_e reduction was not significant at the easy level, perhaps because initial L_e was lower as participants had prior exposure to the audio from the first data-representation. Both transitions were significant ($p_{\alpha_m} = 0.02, p_{\alpha_h} = 2.89*10^{-4}$) in the harder levels. For the second cohort only the easy level α-transition L_e reduction was significant ($p_{\alpha_e} = 1.4*10^{-4}$). The first cohort saw significant α-transition L_e reduction for the third data-representation only at medium difficulty ($p_{\alpha_m} = 0.03$), likely because with first exposure at easy level learning by sound remained too difficult, and at the hard level the data-set may have comprised too many utterances. For the second cohort, which was more attuned to the audio from the outset as images did not provide as much meaning in the first data-representation, third data-representation α-transition L_e fell significantly at both easy ($p_{\alpha_e} = 1.88 * 10^{-3}$) and medium ($p_{\alpha_m} = 2.36 * 10^{-4}$) difficulties. The hard level likely remained inordinately complex. While these L_e reductions across α-transitions may be attributed in the first instance of play to increasing game-mechanic familiarity, in all other instances, they likely result from data-representation comprehension and knowledge acquisition. Few β-transition results remained statistically significant. The first cohort saw first data-representation L_s at medium difficulty increase ($p_{\beta_m} = 0.01$) and the second cohort saw first data-representation L_e fall at hard difficulty ($p_{\beta_h} = 0.02$). Lack of significant β-transition metric shifts notwithstanding, general learning continued in both cohorts. For the first cohort, the strongest learning effects were found with the first data-representation, likely due to lack of ambiguity between images and calls, while the second cohort saw the strongest continued learning effect with the second data-representation supporting the contention that for more complex relationships between species and their calls, spectrogram images provide meaningful benefit to utterance comprehension.

6.2.3 Research exploration: learning from relaxed goal-state constraints

Are learning effects still present with the relaxation of goal-state constraints? I will present my search for learning effects through iterative play of SpectroPuzzle. If found they support my development and introduction of such games to further spectrogram familiarity and enable discussion of optimal goal-state complexity.

6.2.3.1 Results: learning through play with SpectroPuzzle

Databases populated on-device during play, introduced in §5.3.4.2 and enumerated in fig. A.I.3, pg. 252, having been updated to incorporate a UID field for each user, were queried via sqlite3. For each iteration played, difficulty, number of turns taken, and a board map at each move were stored for analysis with a python script for computing the (W_ϵ) metric, introduced in §5.4.3.2, equation (5.8)(fig. 6.12, pg. 175). Observed increases in W_ϵ with continued interaction support my contention that learning occurs through SpectroPuzzle play. I examine whether repeating SpectroPuzzle play increases participant success. If so,

```
Thu Apr 25 @ 15:55:31$ ParsePuzzleGameV2.py 124_D3.csv
==================================
========== GAME ANALYSIS ==========
==================================
Difficulty Level Is:  3
Num Turns Taken:  6
Clean Map:  <2.3=9.3>;<3.3=7.0>;<1.3=10.3>;<0.0=9.1>;<0.3=6.1>;<3.1=10.2>;<0.2=6.0>;<2.1=7.2>;<1.0=10.1>;<1.2=7.1>;<3.0=6.2>;<3.2=
10.0>;<0.1=6.3>;<1.1=9.2>;<2.0=9.0>;<2.2=7.3>;] | type(initBoardmap):  <type 'str'>
countMatrix Final:  [[1, 1, 2, 0], [0, 1, 2, 1], [0, 2, 0, 2], [3, 0, 0, 1]]
sum correctly placed tiles:  9
turnsMin:  4.0 type(turnsMin):  <type 'numpy.float64'>  | turnsTaken:  6.0  | type(turnsTaken):  <type 'float'>
>>> gameEfficiency:  0.666666666667
```

Figure 6.12: Script output reporting (W_ϵ) results from one iteration of SpectroPuzzle.

I contend that my puzzle games increase spectrogram familiarity, and enable learning of use for further citizen science involvement.

Mean and median puzzle success metric results for each difficulty level over the first three iterations of play are presented (fig. 6.13, pg. 176). Results exclude games in which the random initial tile placement built a solved board. In the easy mode, which incorporates only 2 species, this occurred frequently and most third iteration easy games are elided from the analysis set. This did not occur within harder games played but it potentially could. Across all three difficulties a learning effect was evident in the α-transition between first and second iterations played. This is attributable, at easy level, to exposure to game-mechanics rather than increased underlying data-representation familiarity, but is likely related with learning at harder levels. Easy level W_ϵ increased and narrowed from

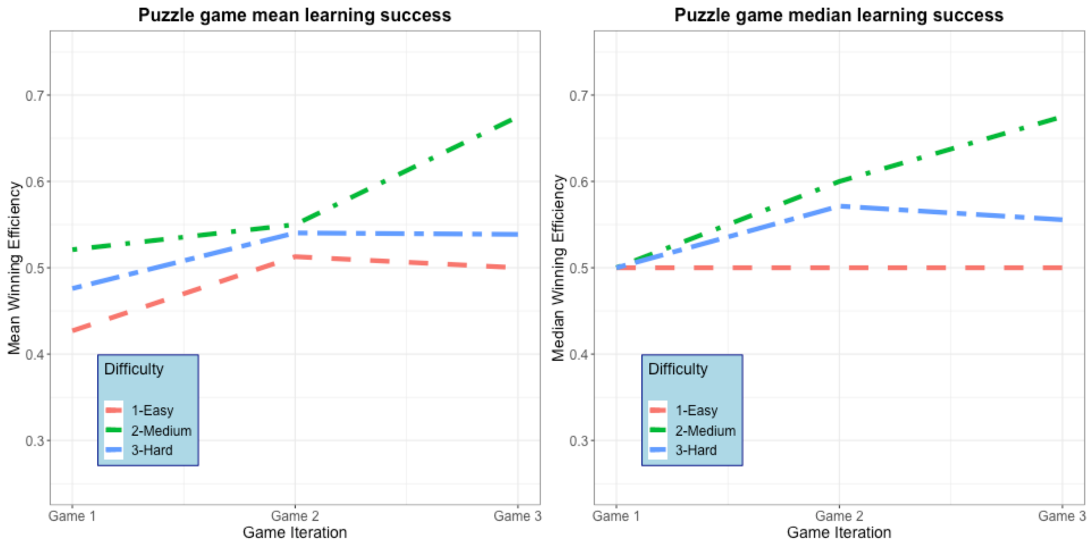

Figure 6.13: SpectroPuzzle learning success is visualised as a shift in W_ϵ across all three difficulties. The easy level functions primarily as training for game mechanics whilst the latter modes require further understanding of the represented data.

$\mu = 0.43, \tilde{x} = 0.5, \sigma^2 = 0.02$ in the first iteration of play to $\mu = 0.51, \tilde{x} = 0.5, \sigma^2 = 0.01$ in the second. As the easy level was designed to introduce game mechanics, few participants played a third iteration of the game, opting instead for more challenging interactions afforded by a more complex board.

With the expanded medium level board learning was evident, as W_ϵ increased in both α- and β-transitions. The first iteration of play (I_1) produced $\mu = 0.52, \tilde{x} = 0.5, \sigma^2 = 0.03$, I_2 rose to $\mu = 0.55, \tilde{x} = 0.5, \sigma^2 = 0.02$, and I_3 to $\mu = 0.68, \tilde{x} = 0.68, \sigma^2 = 0.01$. This complexity level reinforced spectral knowledge taught through BirdMatch play, while reducing confounding results which occurred with spectrally similar calls such as those of *Streptopelia decaocto* and *Columba palumbus*. Knowledge gained through playing the first two difficulties remained useful at the hardest level, however increased game complexity led to a slight overall decrease in W_ϵ from the intermediate level as a result of decreased W_ϵ in games that required sorting spectrally similar calls. W_ϵ rose from $\mu = 0.48, \tilde{x} = 0.5, \sigma^2 = 0.02$ in I_1 to $\mu = 0.54, \tilde{x} = 0.57, \sigma^2 = 0.04$ in I_2, before dropping slightly, albeit remaining above the I_1 baseline, to $\mu = 0.54, \tilde{x} = 0.56, \sigma^2 = 0.03$ in I_3.

6.2.3.2 Discussion: learning through play with SpectroPuzzle

I observe that undirected play with SpectroPuzzle supports continued spectrogram learning. However, data-set curation must consider circumstances where utterances with similar spectral representations cause confusion. In these cases, avian images might be more useful than utterance spectrograms. With the 2-row version, shuffled tiles must be checked for board completion before presentation to the user and, if in a goal-state, reshuffled. As the 2-row version provides little more than training in game-mechanics, and subsequent fieldwork feedback has requested additional game-mechanic explication, I could offer this as a stand-alone training mode. Future designs may be adapted to offer 3-5 species rather than the current 2-4. Testing with 5 rows will be needed to identify whether solutions remain tractable for the target age group and visible on smaller screens.

6.2.4 Summary: questions regarding learning

In summary, I found that participants were confident that my games developed visual and audial knowledge across data-representations (§5.5.1.1). The second cohort's pre- and post-play owl identification task performance changes provide empirical results of knowledge acquisition. These were positive for 3 of 4 species, indicating that my games have training value. Differences between claimed owl call familiarity and tested call identification ability likely result from testing on less common utterances.

For BirdMatch I generated L_s and L_e metrics for analysing participant learning progressions. These metrics support examination of short-term learning processes during iterative game play and may predict future survey data-quality. Both cohorts saw broadly decreasing mean L_e and correspondingly increasing mean L_s over the first three iterations of play, independent of data-representation. While overall learning metrics increased (§5.5.1.2), across the first three iterations of play (α- and β-transitions), regardless of data-representation, learning effect strength declined after initial exposure. At easy difficulty, for the first cohort only first data-representation α-transition error decline was significant, while second cohort α-transition falling error was significant across data-representations. BirdMatch mode difficulty ranking results showed the first overwhelmingly considered easiest and the third hardest, supporting my contention that visual data-representations enhance and ease learning. Examining BirdMatch performance of highly motivated participants, I found subset participants' mean L_s scores generally increased. With increasingly complex second data-set utterances, repetitive play better supported learning across difficulties and data-representations.

Examining SpectroPuzzle learning (§5.5.1.3), I observed an α-transition learning effect across all three difficulties. This is attributable, at easy difficulty, to learning game-mechanics rather than increased data-representation comprehension, but it implies learning at harder difficulties.

6.3 Motivation results and discussion

Educational games have long existed both in physical and virtual forms. In this section I will examine the roles of data-representation familiarity and goal-state complexity, introduced in Chapter 2 (fig. 2.7, pg. 57), in supporting engagement through play. Each of my games explores a different region of my play engagement framework (see figures 5.8, 5.11, and 5.14). I will examine whether my virtual games motivate engagement with the physical world through birdsong education. I will identify the degree of participant game engagement, which modes of play are most familiar and which most engaging, and whether these correlate. I will follow with an investigation into both cohorts' prior interest in nature and avian bioacoustics and whether play increases this interest. I will conclude with an analysis of participants' motivation to follow diverse paths available through the games presented in the previous chapter.

In response to the research question from §5.5.2.1 I will present findings that game-mechanics motivate engagement. In response to the research question from §5.5.2.2 I will present results showing that play does not significantly change reported mean interest in learning about wildlife or desire for engagement with nature. In response to the research question from §5.5.2.3 I will explore whether varying data-representation familiarity and goal-state complexity affects motivation and will present findings that BirdMatch, with simple game-mechanics and clearly defined goal-state, was strongly favoured as easiest, while ChorusComposer was considered most challenging. I will show that for interactions designed to teach a single dominant call per species, motivating engagement with as many variants as possible produces valuable learning trajectories. In response to the research question from §5.5.2.4 I will present results for interactions with more complex training data and will contend that gamified mechanics should predispose participants to focus longer on each variant before continuing.

6.3.1 Research exploration: engagement with educational games

Answers from the first cohort identified 44 participants without prior exposure to educational games and 72 with. While educational games were novel to some, most participants were familiar with their premises. As I treat engagement and game enjoyment interchangeably, I asked both cohorts upon completion of play to rank their enjoyment of the games and their desire for continued engagement with the games. Independent of data-set, participants strongly agreed that game-play provided enjoyment and positively desired to continue play.

6.3.1.1 Evaluating: 'I enjoyed playing the games'

All participants agreed, rather strongly, with the statement 'I enjoyed playing the games'[14]. Their answers provide a measure for participant enjoyment of the games (fig. 6.14, pg. 179). Both cohorts' median responses on a 5-point Likert scale were $\tilde{x} = 5$. However, the

[14] 2017 post-survey Likert Q.L1, 2018 post-survey Likert Q.L1.

2017 results ($\mu = 4.59$, $\sigma^2 = 0.35$, N=98) are slightly lower, albeit less varied than the 2018 results ($\mu = 4.61$, $\sigma^2 = 0.82$, N=70). As hoped, changing data-sets did not significantly

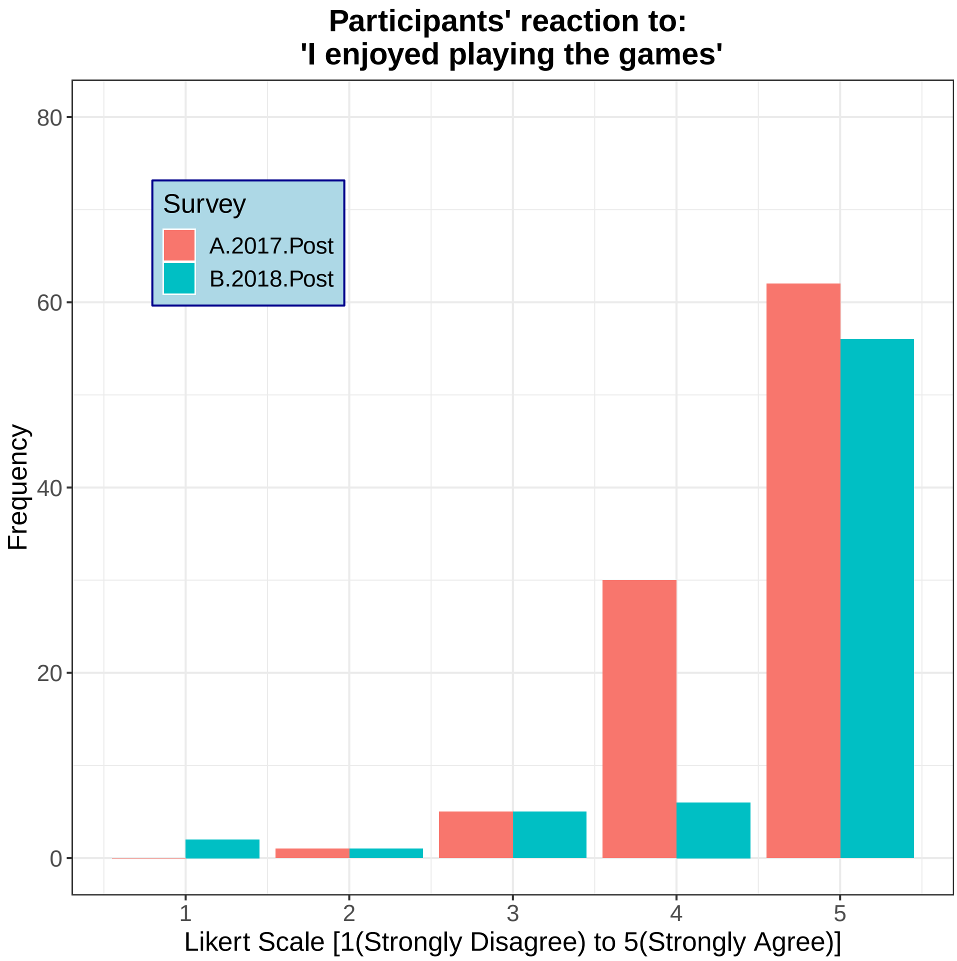

Figure 6.14: Histogram of Likert-scaled responses to questions of whether game play was enjoyable. Both cohorts strongly agreed that it was.

affect game engagement ($p = 0.93$). Both cohorts strongly believed that my games were enjoyable, which supports my contention that these game-mechanics engage citizen science participants. Adapting regional and local datasets for these games is unlikely to diminish engagement.

6.3.1.2 Evaluating: 'I want to continue playing such games/learn about birdsong'

First cohort participants generally agreed with the statements 'I want to continue playing such games' and 'I want to learn about birdsong'[15]. Responses provide a basis for discussing whether games engage participants with data, and whether such interactions are sufficient to motivate further birdsong engagement without ludic intervention (fig. 6.15, pg. 180).

[15] 2017 post-survey Likert Q.L2 & Q.L5.

Median responses from both cohorts were $\tilde{x} = 4$. Participants strongly agreed with the first statement ($\mu = 4.26$, $\sigma^2 = 0.70$, N=98) and somewhat with the second, ($\mu = 3.92$, $\sigma^2 = 0.83$, N=100). These results suggest that although game mechanics increase desire

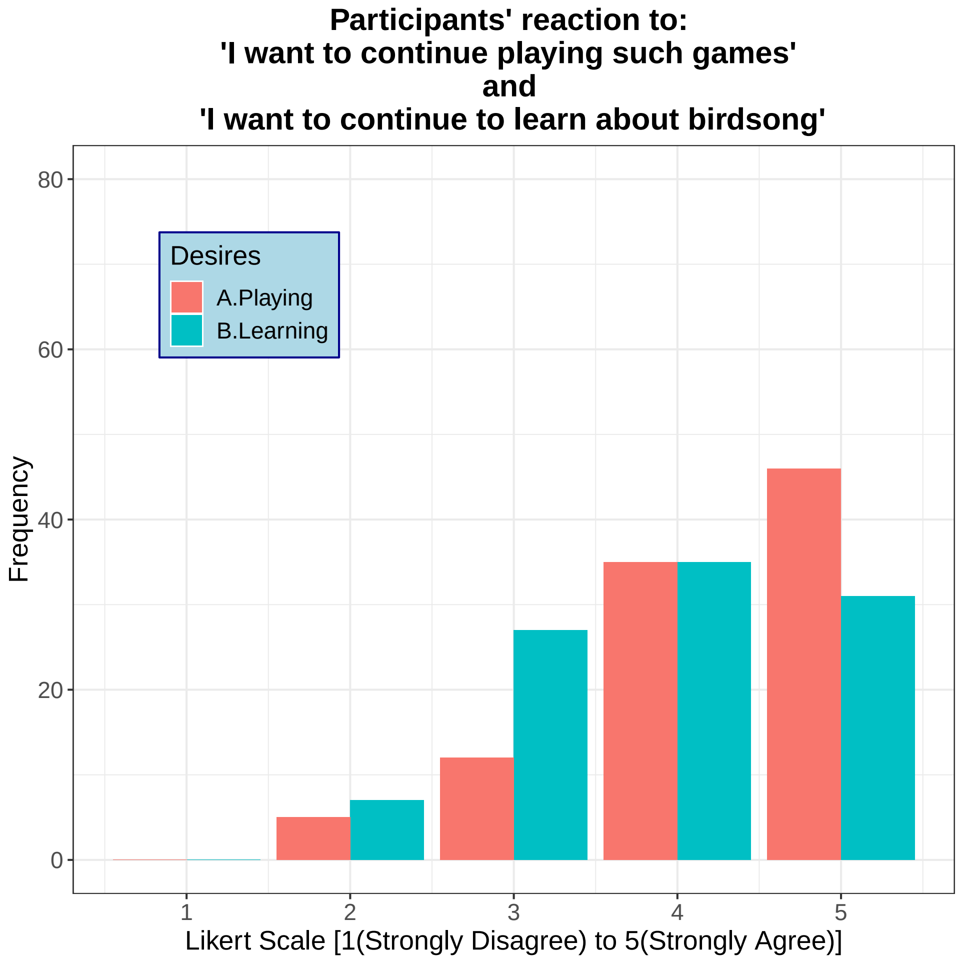

Figure 6.15: Histogram of Likert-scaled responses to questions of whether participants were motivated to continue playing the games and learning about birdsong. These results suggest that while game-mechanics support desire for play, they provide less motivation for abstract engagement with avian utterances.

for play, they provide less motivation for abstract engagement with avian utterances when more nuanced knowledge is taught.

6.3.1.3 Discussing engagement with educational games

Strong agreement by both cohorts that my games were enjoyable supports my contention that game-mechanics engage citizen science participants. Subsequent feedback included a description of the games as 'satisfyingly addictive'. Implementing regional and local datasets for these games is unlikely to diminish engagement, as mechanics remain unchanged. However, while mechanics extrinsically motivate desire for play, they may amotivate non-gamified engagement with avian utterances, which can decrease engagement with nature, when nuanced content, such as multiple calls per species, is taught.

6.3.2 Research exploration: interest in nature & avian bioacoustics

I identify baseline interest in nature and desire for engagement with wildlife outdoors prior to game interaction and examine whether play influences these. I follow with questions relating knowledge of nature with enjoyment and ask whether participants desire further learning. Responses are analysed across cohorts to see whether both feel similarly, despite demographic variation. Responses between pre- and post-surveys for the second cohort are examined to see whether playing the games causes an effect.

6.3.2.1 Evaluating: 'I am interested in learning more about wildlife'

Participants from both cohorts generally agreed with the statement 'I am interested in learning more about wildlife'[16], although this was not significantly affected by play (see §B.III for detailed results). Responses gauging intrinsic motivation for learning about wildlife and whether this is affected by play (fig. B.III.3, pg. 262) show no significant shifts in belief were associated with having played the games in the second cohort. Second cohort mean interest in learning about owls rose post-play, but some strong disagreement remained with use of the games as tools for increasing interest in learning.

6.3.2.2 Evaluating: 'I would like to spend more time outdoors listening to birds/owls'

All participants only slightly agreed with the statements 'I would like to spend more time outdoors listening to birds/owls'[17]. My game interactions guide learning these vocalisations not *in situ*. For detailed results quantifying participant desire for engagement with avian vocalisations in nature see §B.IV (fig. B.IV.4, pg. 263). While the two cohorts can be assumed to have been drawn from the same population, desire for engagement does not significantly shift with play but does not significantly reinforce previously identified risks associated with technologically mediating nature engagement.

6.3.2.3 Evaluating: 'if I knew more about wildlife I would enjoy nature more'

Participants in both cohorts generally agreed with the statement 'if I knew more about wildlife I would enjoy nature more'[18]. Results quantify the extent to which participants

[16] 2017 pre-survey Likert Q.L3, 2018 pre-survey Likert Q.L3 / post-survey Likert Q.L9.
[17] 2017 post-survey Likert Q.L7, 2018 pre-survey Likert Q.L7/post-survey Likert Q.L6.
[18] 2017 pre-survey Likert Q.L4, 2018 pre-survey Likert Q.L4/post-survey Likert Q.L10.

correlate increased knowledge about with increased enjoyment of nature (fig. 6.16, pg. 182). Median responses for both cohorts prior to play were $\tilde{x} = 4$; this remained unchanged post-play for the 2018 participants. Mean 2017 pre-survey results ($\mu = 4.18$, $\sigma^2 = 0.73$, N=132) are slightly higher and less varied, respectively, than the 2018 pre-survey results ($\mu = 4.04$, $\sigma^2 = 1.40$, N=102). Both cohorts believed, albeit not strongly, that learning about wildlife could increase enjoyment of nature prior to play. Post-play Likert-scored 2018 participants' responses regarding interest in nature fell ($\mu = 3.80$, $\sigma^2 = 2.16$, N=65). The two cohorts

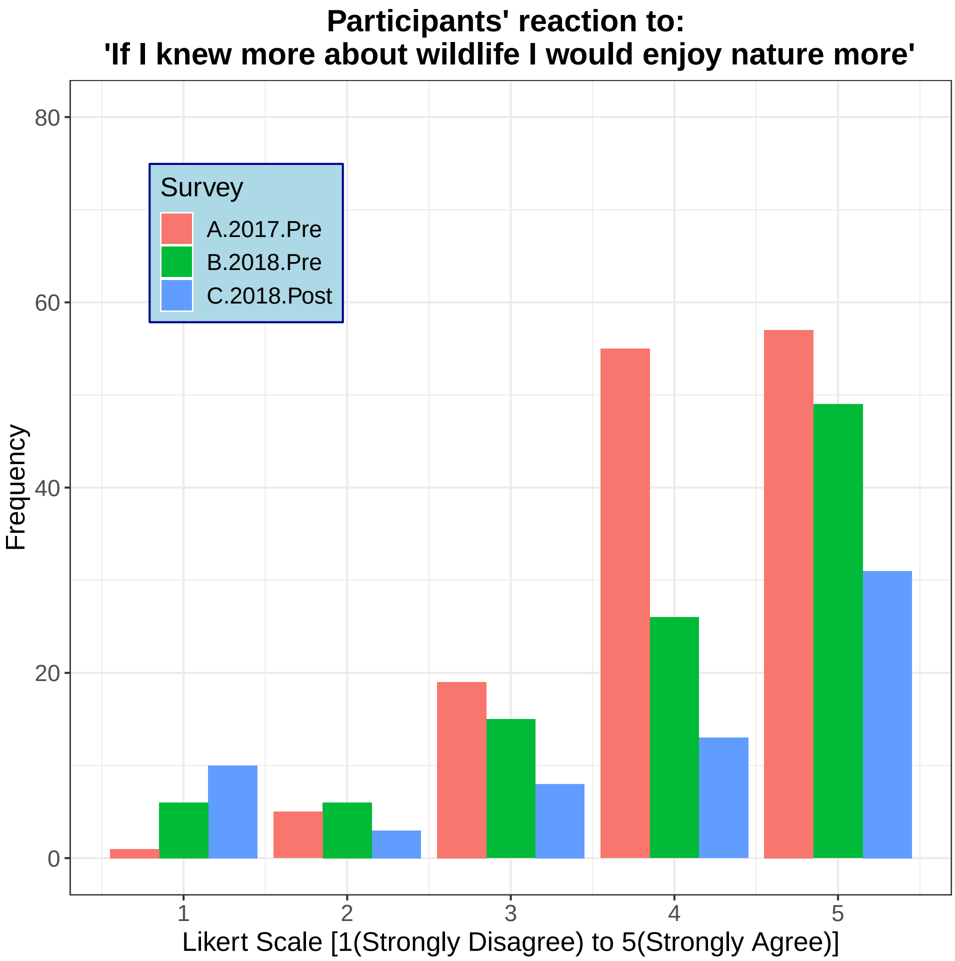

Figure 6.16: Histogram of Likert-scaled responses to the premise that participants think that increased knowledge of wildlife will lead to increased enjoyment of nature. Both cohorts believed that learning about wildlife could increase enjoyment of nature, although this again diverges post-play.

cannot be assumed to have been drawn from distinct populations ($p = 0.14$), although they are not particularly similar prior to play. Second cohort's beliefs did not significantly fall with play ($p = 0.19$). This result may reflect the risk of providing a motivational interface for avian knowledge acquisition which does not require *in situ* engagement with nature.

6.3.2.4 Evaluating: 'being able to recognise bird/owl call/song makes being outdoors more enjoyable'

All participants generally agreed both pre- and post-play with the statement 'being able to recognise bird/owl calls/songs makes being outdoors more enjoyable'[19]. Despite generally positive results, changes with play were inconclusive (for complete results see §B.V, pg. 262). Results quantify whether hearing avian utterances provides intrinsically motivating enjoyment and whether knowledge, gained through play, increases expected future enjoyment when listening to nature (fig. B.V.5, pg. 264). Participants in both cohorts were comfortable associating avian utterance familiarity with enjoyment of nature, but this did not change significantly with play and commensurate knowledge acquisition, possibly because the games enable exposure to birdsong without the need to go outside.

6.3.2.5 Evaluating: 'I want to learn more about owls'

Participants in 2018 generally agreed both pre- and post-play with the statement 'I want to learn more about owls'[20]. Results quantify a baseline for the second cohort's interest in their games' target species (fig. 6.17, pg. 184). Median responses before and after play were $\tilde{x} = 5$. 2018 pre-survey results ($\mu = 4.27$, $\sigma^2 = 1.34$, N=98) are slightly higher and less varied, respectively, than post-survey results ($\mu = 4.10$, $\sigma^2 = 1.56$, N=69). While the somewhat strongly agreed desire for knowledge validates introducing students to owl calls through games, these results are insufficient to claim that the adverse fall in cohort belief changed significantly with play ($p = 0.11$).

6.3.2.6 Discussing changes to interest in nature with play

Mean interest in learning about wildlife rose insignificantly with play. For the second cohort, more strong than general disagreement with the statement remained after engaging with the games. While mean desire for engagement with nature fell, this is not significant and fails to reinforce previously identified risks associated with technologically mediating nature engagement. Both cohorts believed that learning about wildlife could increase enjoyment of nature. Marginal decrease in second cohort participant belief that knowledge of nature increases enjoyment of nature may reflect the risk of providing a motivational interface for avian knowledge acquisition which does not require *in situ* engagement with nature. I cannot claim that associating avian recognition with enjoying the outdoors significantly changes with play, perhaps because the games enable birdsong exposure without the need to go outside. However the generally expressed desire for knowledge validates introducing students to owl calls through games.

[19] 2017 pre-survey Likert Q.L5/post-survey Likert Q.L6, 2018 pre-survey Likert Q.L6/post-survey Likert Q.L5.

[20] 2018 pre-survey Likert Q.L5 / post-survey Likert Q.L4.

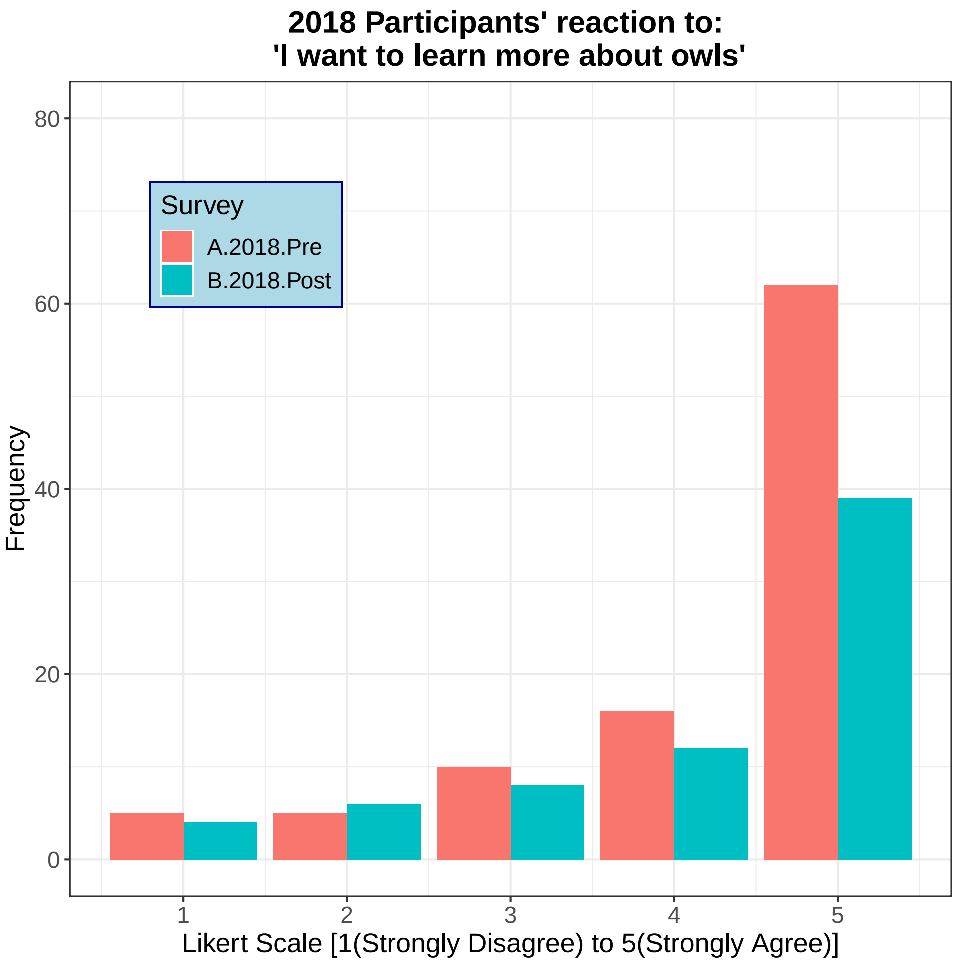

Figure 6.17: Histogram of Likert-scaled responses to the question of whether participants were interested in acquiring knowledge about owls. Despite strong agreement with the premise, this result is insufficient to claim that play positively influenced belief.

6.3.3 Research exploration: how data-representation familiarity & goal-state complexity affect motivation

Following previous results I investigate first cohort feelings about game difficulty and preferred class of ludic interaction. I explore relationships amongst game enjoyment, data-representation familiarity, and degree of perceived challenge. Results of the first cohort's game and representation preference responses guide discussion to questions of confidence in and motivation for finding solutions to the games, as applicable.

6.3.3.1 Evaluating game preference & difficulty ranking

The 2017 participants were asked to identify which of the three games they found hardest and easiest[21]. Results were roughly symmetrical and BirdMatch, with its simple game-mechanics and clearly defined goal-state, was strongly favoured as easiest, while the

[21] 2017 post-survey Q.5/Q.6.

composition game was considered most challenging. (fig. 6.18, pg. 185.) This result may

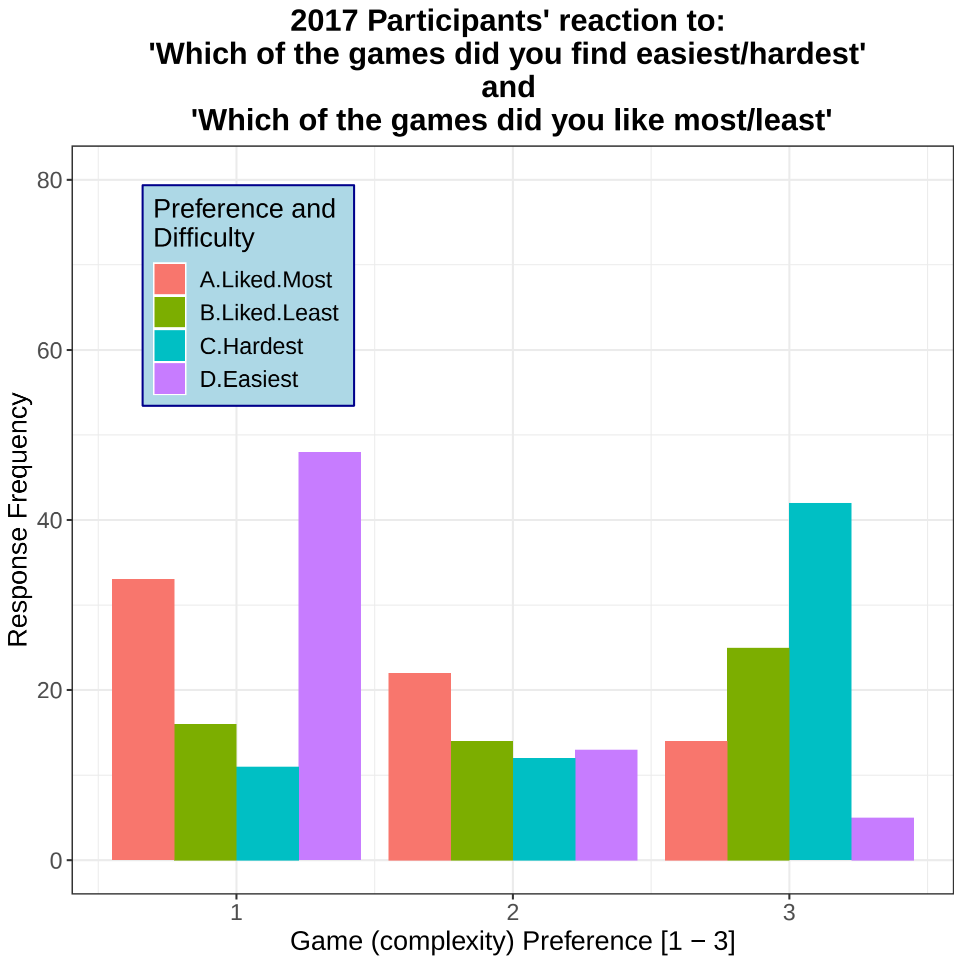

Figure 6.18: Participant ranking of the game classes, with variable goal-state, by preference for playing and by perceived difficulty. Results are generally symmetrical but indicate that enjoyment and perceived difficulty are not strongly correlated.

reflect insufficient time for participants to gain familiarity with spectrogram representations as tools for open-ended play, or it may indicate that open-ended play goal-state complexity is too abstract for the target age-group. As play-order was semi-structured for teaching purposes, it is not feasible to fully distinguish whether these results are a function of player fatigue or game quality.

The 2017 participants were asked to identify which of the three games they liked most and least[22]. While results were roughly symmetrical, the first game was preferred, followed by the second, with the third liked least, bias was less extreme than for questions of difficulty, indicating that enjoyment and perceived difficulty are not strongly correlated. This trend was evident when the question was stated in both the affirmative and negative – fewest participants liked SpectroPuzzle least. As play order was semi-structured, result biases are influenced by those whose poor BirdMatch performance preempted their ability

[22] 2017 post-survey Q.3/Q.4.

to explore the more complex games. These results indicate that my interaction designs support user development of spectral data-representation comprehension, although the relaxed goal-state of open-ended play may be better suited to an older audience.

6.3.3.2 Evaluating BirdMatch mode preference & difficulty ranking

The 2017 participants were asked to identify which modes of the memory game they liked most and least[23]. I find that simplicity and data-representation familiarity are not strong intrinsic motivators. For the first mode, a single species utterance was matched with different images of male and female adults birds, for the second, the matched images were of the spectrograms of identical utterances, while the third mode required that participants correctly match identical utterances without visual feedback. BirdMatch mode preference results were evenly distributed. Slightly more participants ranked the third mode most liked, while the first mode was least liked (fig. 6.19, pg. 6.19). This is

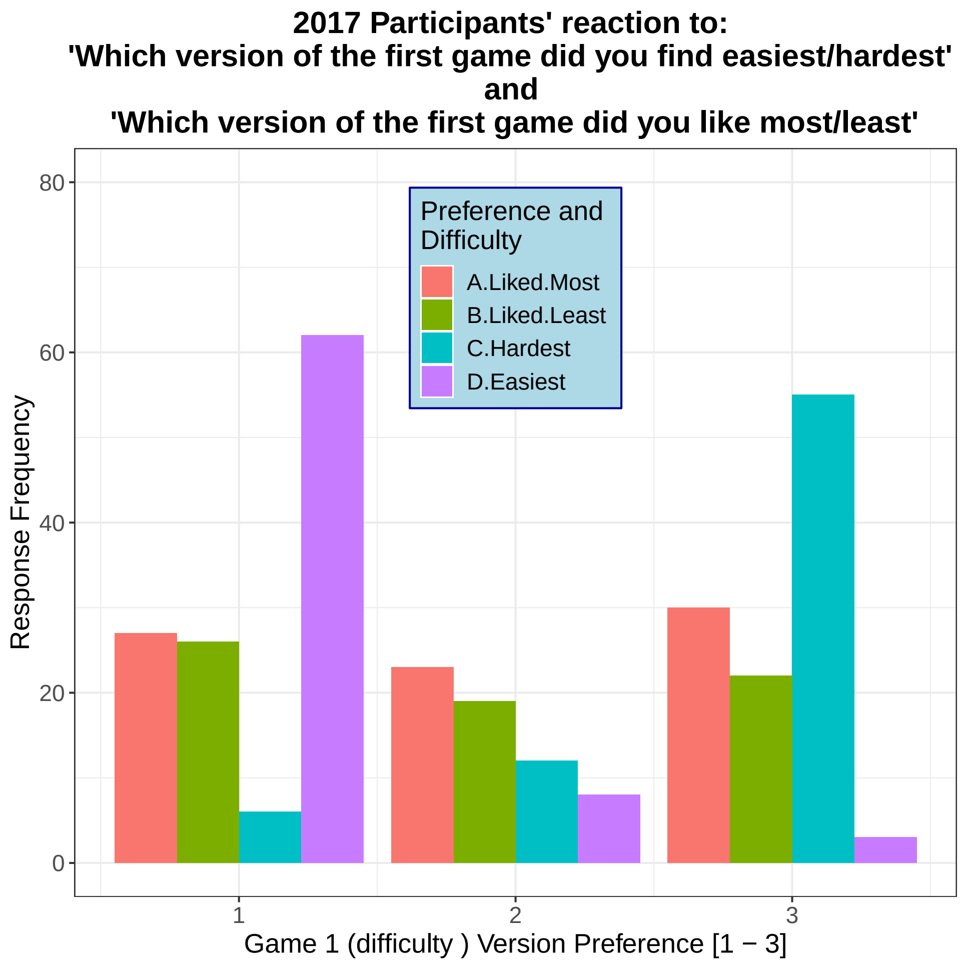

Figure 6.19: Participant ranking of modes of the memory game, with variable data representation, by preference for play and by perceived difficulty. Results support the premise that overly simple and familiar goals and data-representations are insufficient intrinsic motivators.

[23]2017 post-survey Q.7/Q.8.

surprising as the third mode, providing ear-training without visual support, was presumed to be hardest. Nevertheless, on aggregate more participants preferred some form of visual representation. The 2017 participants were asked to identify which BirdMatch modes they found hardest and easiest[24]. Results were symmetrical and strongly biased, with the first data-representation mode overwhelmingly considered easiest and the third hardest. These results identify limitations with Curtis' motivational model (see §2.5.1.1, fig. 2.5) and support my contention that overly simple and familiar goals and data-representations provide insufficient intrinsic motivation.

6.3.3.3 Evaluating: 'I was motivated to find the solutions to the games' & 'I am confident that my turns led to winning'

First cohort participants generally agreed with the statements 'I was motivated to find the solutions to the games' and 'I am confident that my turns in games 1/2 led to winning'[25]. Results provide insight into whether defined goal-states motivate user engagement and whether these game mechanics are sufficiently comprehensible to preempt need for additional training (fig. 6.20, pg. 188). Median responses to all questions were $\tilde{x} = 4$; participants agreed the games motivate goal-state achievement ($\mu = 4.13, \sigma^2 = 0.60, N = 92$) and were more confident that BirdMatch moves led to the goal-state ($\mu = 4.14, \sigma^2 = 0.66, N = 88$) than SpectroPuzzle moves ($\mu = 3.99, \sigma^2 = 0.64, N = 75$), where the goal-state path is less rigid. Participants agreed that the games provided motivation to search for, and the likelihood of reaching, their respective goal-states. Game mechanic adaptations were not deemed necessary for the second round of fieldwork, although feedback has included multiple SpectroPuzzle tutorial requests.

6.3.3.4 Discussing how data-representation familiarity & goal-state complexity affect motivation

Game preference results support my design approach of enabling users to develop spectral data-representation comprehension prior to pursuing complex interactions through games. The relaxed goal-state of open-ended play may be better suited to an older audience. BirdMatch mode difficulty ranking results showed the first overwhelmingly considered easiest and the third hardest, supporting my contention that visual data-representations enhance and ease learning. As the SpectroPuzzle goal-state path is less rigid, §6.3.3.3 results support my contention that confidence correlates with goal-state simplicity. Participants agreed that the games provided motivation to search for, and the likelihood of reaching, the goal-states so game-mechanic adaptations were not necessary.

[24] 2017 post-survey Q.9/Q.10.
[25] 2017 post-survey Likert Q.L8/Q.L9/Q.L10.

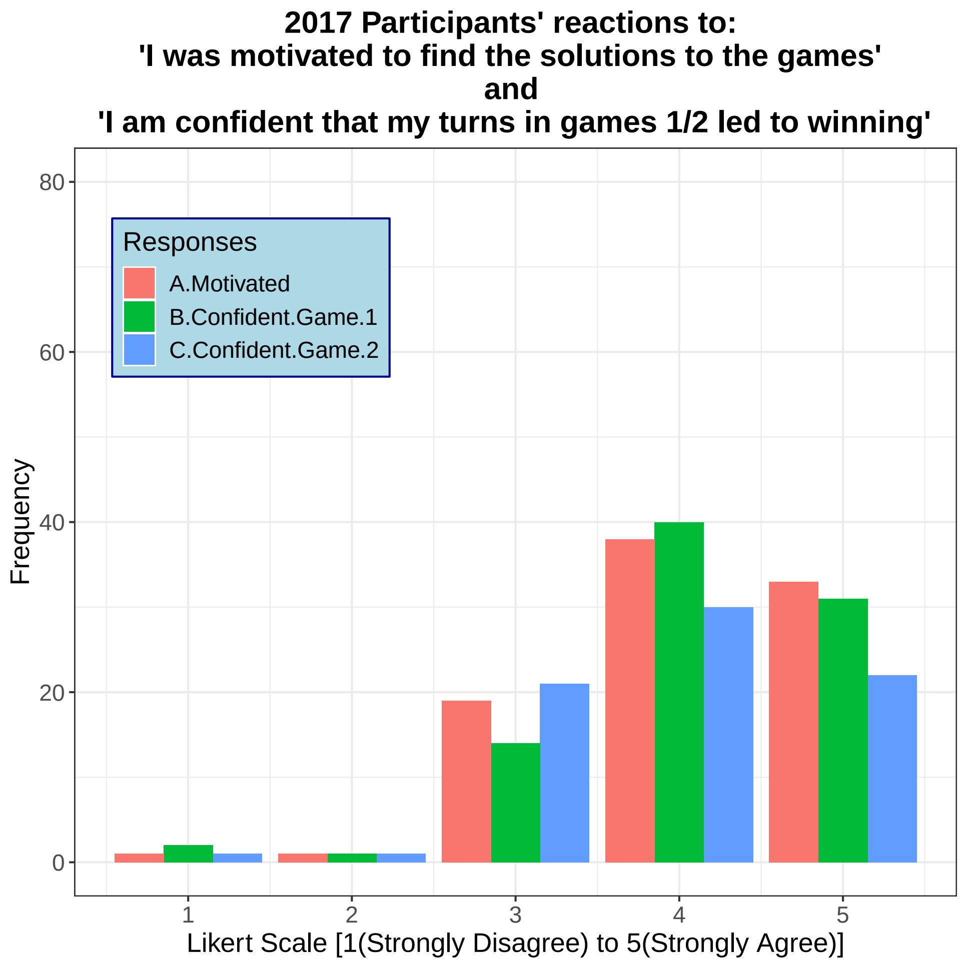

Figure 6.20: Histogram of Likert-scaled responses as to whether participants were motivated to find the game solutions, and whether they thought that their actions were guiding them towards a known goal-state. Agreement with the claims supports my contention that motivation and confidence are positively correlated.

6.3.4 Research exploration: motivating learning through play

Is motivation to play through the games' various data-representation and difficulty variants age-dependent? I examined game-play data exploring how participants across cohorts engaged with and traversed BirdMatch with each dataset and, for the first cohort, the option to engage with the data through other ludic interactions. Does learning success vary as different trajectories are pursued through the memory game variants? I investigate subsets of each cohort who present extremes of playful behaviours in BirdMatch, classified by proposed underlying motivations, and explore subsets aggregate learning success (L_s). I explore correlations between learning and engagement with SpectroPuzzle by identifying learning effects for the subset of participants who played most and compare their results with first cohort W_ϵ). I subsequently explore whether open-ended play engages target end-users by observing engagement metric evolution across iterative compositions produced with ChorusComposer. I conclude by examining learning for the subset of individuals whose trajectories through all games presented as extremes of engagement behaviours.

6.3.4.1 BirdMatch engagement & exploration by age

Ludic designs allowing multiple trajectories for play while supporting participant interest in the target species encouraged game engagement. Prior to game development, whether unfamiliar spectral visual representations would support engagement or be too esoteric for primary students was unknown. From the first cohort, 100 participants played 734 iterations of all 9 data-representation/difficulty BirdMatch variants. Individual participants played from 1 to 21 variant iterations ($\mu = 7.27$, $\tilde{x} = 6$, $\sigma^2 = 21.02$). The first cohort was presented with all three game classes. Despite being advised to leverage the memory game for initial learning, ~40 participants minimised BirdMatch interactions, preferring other types of play. In the second cohort, 102 participants were recorded playing 759 iterations across the 9 BirdMatch variants. Individual participants played from 2 to 21 variant iterations ($\mu = 7.44$, $\tilde{x} = 7$, $\sigma^2 = 13.12$). As a large proportion of the first cohort avoided BirdMatch, initial game choice was constrained for the second cohort. Engagement was similar across cohorts; diverse engagement behaviours were evident among those who played most within each cohort.

Characteristic extremes of engagement behaviour associated with flow were common to both cohorts: some chose extreme repetition of particular data-representation/difficulty variants, focussing on maximising gamified rewards, with presumed ancillary learning; others pursued the extreme of completing as many variants as possible. Investigating the engagement-through-play continuum, I found participants in the first cohort engaged with $\mu = 4.10, \tilde{x} = 4$, $\sigma^2 = 3.17$ variants, those in the second explored $\mu = 3.93, \tilde{x} = 3$, $\sigma^2 = 4.34$ variants. While second cohort participants played more iterations on average than those in the first, they explored fewer possible BirdMatch variants. This was likely because owl images, associated with multiple utterances, provided less information, so learning required more repetition. This may have been further confounded by the higher score requirements necessary for advancement from §5.4.2.1, pg. 144. Both cohorts included individuals who presented either extreme of engagement behaviour and many who combined both; extreme subsets are examined in §6.3.4.2.

Applying the E_q metric introduced in (5.6), I present log_{10}-scaled results for participant engagement by cohort (fig. 6.21, pg. 190). In the first cohort, summary statistics for participant E_q were $\mu = 0.68, \tilde{x} = 0.67$, $\sigma^2 = 0.06$, N=136, for the second cohort these were $\mu = 0.56, \tilde{x} = 0.50$, $\sigma^2 = 0.05$, N=105. For a breakdown of exploration quotients by age across cohorts, see fig. 6.22, pg. 190. The first cohort was less motivated than the second to learn through variant repetition, instead exploring and gaining increased data-representation exposure, perhaps due to less data-set complexity. These results demonstrate the need to motivate a variety of participant types and validate my contention

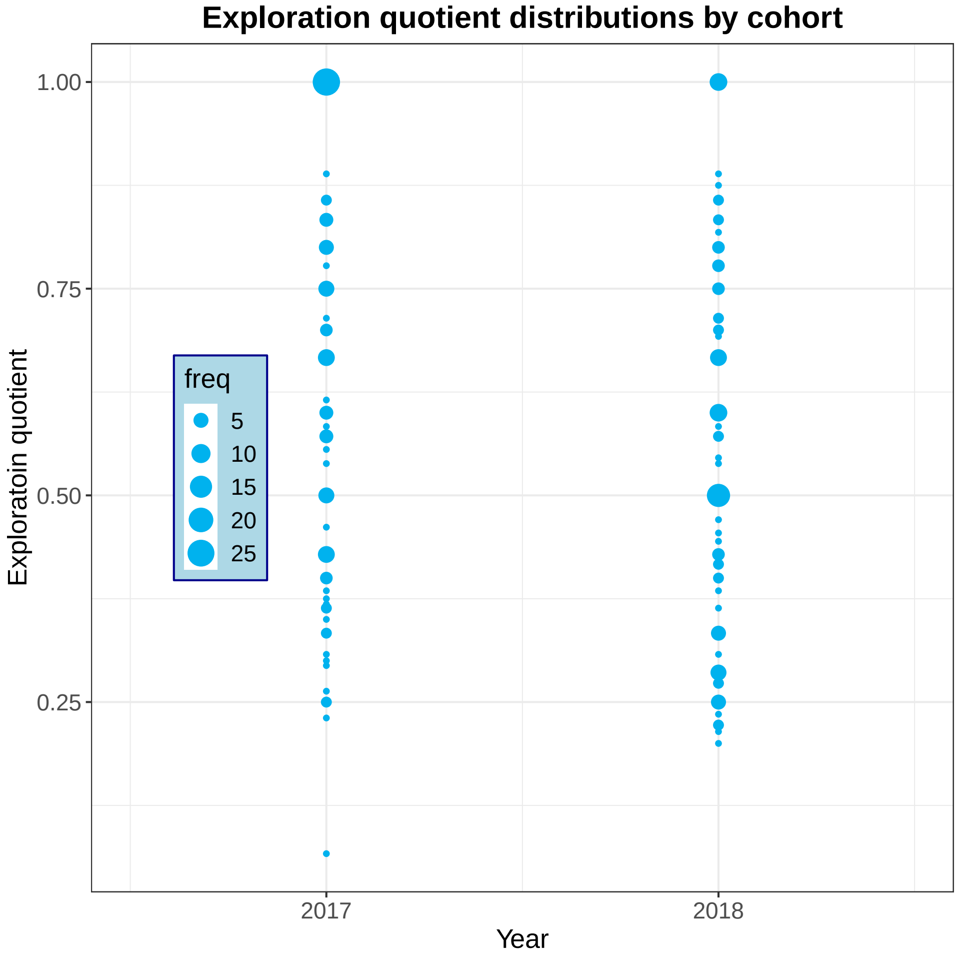

Figure 6.21: Comparison of exploration quotients across cohorts. As the 2018 cohort did not have further games to play, those who played more than 9 times inevitably played duplicate variants, reducing the preponderance of players with $E_q = 1$.

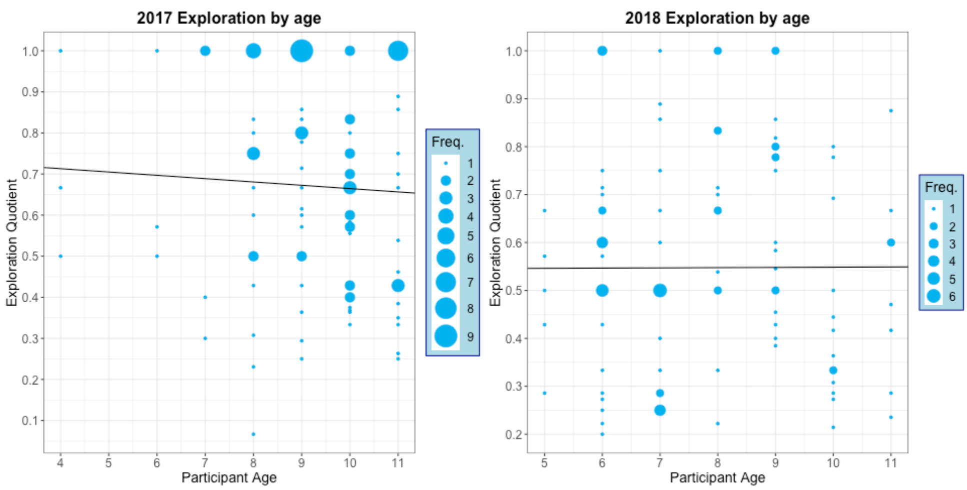

Figure 6.22: Comparison of exploration quotients by age for each cohort with weighted linear regression lines imposed. Multiple approaches to engagement exist for participants of all ages.

that multiple approaches to engagement exist for participants of all ages. As age dependency is not found, my designs are suitable for motivating students across key stages.

6.3.4.2 Exploration vs. learning: divergent ludic engagement behaviours in BirdMatch

I examine three subsets of each cohort who displayed particular play behaviours which represented diverse motivational extremes resulting in interaction flow states: (1) those who played the most total BirdMatch iterations regardless of variant, subset A; (2) those who played the most iterations of each single BirdMatch variant, for which at least 4 iterations were played, subset B; (3) those who explored 7 or more of the BirdMatch variants, ensuring that they played at minimum each difficulty and data-representation, subset C. For each subset, I calculate summary E_q statistics and L_s scores for comparison with their respective cohort totals. Presupposing engagement is positively correlated with learning, with the simple data-set exploring most variants yielded highest comparative L_s, whereas with the complex data-set, single variant repetition yielded most increased L_s.

Relationships between game engagement, exploration, and learning, measured by total game iterations played, E_q, and L_s metrics respectively, warrant discussion. Within the 2017 cohort, the most involved participants[26] played 14 – 21 iterations of 1 – 9 BirdMatch variants. Summary statistics for these players' E_q ($\mu = 0.36, \tilde{x} = 0.42, \sigma^2 = 0.01$, N=10) are significantly lower ($p = 1.62 * 10^{-14}$) than those for the cohort ($\mu = 0.68, \tilde{x} = 0.67$ and $\sigma^2 = 0.06$, N=136). Those who played most overall were more likely to pursue repetition than variant exploration. I expect that repetition correlates with stronger learning success. Within the 2018 cohort the most involved participants [27] played 13 – 20 iterations of 3 – 9 variants. E_q statistics for these players ($\mu = 0.38, \tilde{x} = 0.35, \sigma^2 = 0.02$, N=10) were significantly lower ($p = 1.38 * 10^{-08}$) than for the 2018 cohort ($\mu = 0.56, \tilde{x} = 0.50$, $\sigma^2 = 0.05$, N=105). Subset A engagement differs significantly from that of their cohort. Given the selection criteria for subset A, no directed drift in E_q was expected. Flow associated with focussed engagement significantly reduced exploration. I therefore consider whether focussed engagement correlates with increased learning success. To this end, table 6.6, pg. 192, compares mean learning success metrics by year between cohorts and their respective subsets who played the most total iterations. Both cohorts' most engaged subsets achieved higher mean L_s scores across BirdMatch on aggregate and at each difficulty level. Participant motivation correlates with increased learning, regardless of difficulty.

Participants who maximised iterations played of each single variant, subset B, are examined. As these individuals by definition explore less than their respective cohorts

[26] 2017 UID's: 77, 86, 96, 87, 89, 78, 97, 94, 85, 79.
[27] 2018 UID's: 30, 33, 50, 69, 78, 80, 84, 101, 102, 106.

Mean L_s scores comparing cohorts to respective subsets [A]						
Difficulty	2017			2018		
	Cohort 1	Subset 1.A	% +/-	Cohort 2	Subset 2.A	%+/-
$Easy_{L_s}$	0.135	0.187	+34%	0.151	0.172	+14%
$Medium_{L_s}$	0.071	0.097	+37%	0.048	0.080	+67%
$Hard_{L_s}$	0.032	0.055	+72%	0.019	0.032	+68%
$Total_{L_s}$	0.095	0.125	+32%	0.108	0.129	+19%

Table 6.6: Comparison of mean L_s scores for each subset A participants with cohort mean cumulatively and for specified difficulties. The most engaged subsets perform better both in total and at each difficulty level compared to their respective cohorts.

I compare their E_q for significance validation. I then examine whether maximising play of a single variant results in higher L_s scores or if a learning effect is present from the first few iterations. From the 2017 cohort, 7 variants saw individual participants play 5 – 15 iterations. E_q statistics for these participants[28] ($\mu = 0.31, \tilde{x} = 0.31, \sigma^2 = 0.01$, N=8) are significantly lower ($p < 2.20 * 10^{-16}$) than for the entire cohort ($\mu = 0.68, \tilde{x} = 0.67$, $\sigma^2 = 0.06$, N=136), as expected given that E_q is inversely affected by repeated single variant play. Comparing subsets B to respective subsets A participants' E_q, all of whom expressed strong engagement behaviour, albeit those in B with only a single variant, are not significant in either year ($p_{2017} = 0.07, p_{2018} = 0.22$). I explore whether subsets B participants achieved higher mean L_s scores for each difficulty and data-representation than their respective cohorts. Insufficient data existed for the two harder difficulties of the third mode (see table 6.7, pg. 192). The first cohort's simpler data-set did not support increased learning by

Mean L_s scores comparing cohorts to respective subsets [B]							
Variant		2017			2018		
Difficulty	Representation	Cohort	Subset B	% +/-	Cohort	Subset B	% +/-
Easy	Avian Images	0.132	0.212	+61%	0.145	0.188	+30%
	Spectrogram	0.173	0.287	+66%	0.185	0.156	-16%
	Blank	0.092	0.057	-38%	0.127	0.187	+47%
Medium	Avian Images	0.072	0.020	-72%	0.051	0.118	+131%
	Spectrogram	0.089	0.051	-43%	0.044	0.049	+11%
	Blank	0.039	N/A	N/A	0.037	N/A	N/A
Hard	Avian Images	0.034	0.024	-29%	0.020	0.032	+60%
	Spectrogram	0.038	0.017	-55%	0.017	N/A	N/A
	Blank	0.029	N/A	N/A	0.017	0.019	+12%

Table 6.7: Comparison of mean L_s scores for each subset B participants with cohort mean for specified variants. Blue denotes subset performs better than cohort mean, red denotes worse performance. Where no participant played more than 3 iterations, N/A is reported, marked yellow. Where variant iteration count ties occurred, score mean are reported.

[28] 2017 UID's: 78, 75, 84, 42, 89, 92, 87, 23.

those who played a single variant repetitively with subset B performing worse than their cohort in 5 of 7 variants. The blank visual data-representation failed to support learning for the subset, regardless of difficulty. While increased difficulty motivated repetition for the other representations, this did not support learning. Frustration associated with increased difficulty is not a desirable means of motivating repetitious play.

For seven 2018 variants individual participants played 4 – 10 iterations. These participants'[29] E_q statistics ($\mu = 0.35, \tilde{x} = 0.35, \sigma^2 = 0.02$, N=7) were significantly lower ($p = 8.09 * 10^{-12}$), while deviating less than in 2017, when compared with results for the 2018 cohort ($\mu = 0.56, \tilde{x} = 0.50$, $\sigma^2 = 0.05$, N=105). Examining whether 2018 subset B showed increased L_s relative to cohort average is of interest (see table 6.7, pg. 192). In all but one sufficiently played variant subset participants' mean L_s scores increased. With increasingly complex second data-set utterances, repetitive play better supports learning across difficulties and data-representations.

Finally, I discuss characteristics of play for participants who explored the most Bird-Match variants, subset C. I calculate mean L_s scores for this subset for each variant for which multiple iterations were played and compare results to those of their respective cohorts to examine whether such engagement is correlated with learning. Participants who played at least 7 of the 9 variants are guaranteed to have played at least one iteration with each data-representation and difficulty; in 2017, 13 participants[30] and in 2018, 15[31] met this criterion. As expected E_q for subsets C were higher than cohort average, ($\mu = 0.64, \tilde{x} = 0.67, \sigma^2 = 0.04$, N=13) for the first cohort and ($\mu = 0.73, \tilde{x} = 0.78, \sigma^2 = 0.02$, N=15) for the second. Comparing mean cohort L_s scores by variant with mean L_s scores for subset C players indicates whether participants who are motived to play most variants with minimal repetition still see a learning effect (see table 6.8, pg. 194). In the easy mode both cohorts' subset C learning effects were stronger than for the cohort in all but one instance. However at medium difficulty mean L_s scores were greater only half the time — higher predominantly for the second cohort, and at hard difficulty only one third of the time — higher predominantly for the first cohort. 2017 subset C, trained on a less complex dataset, performed better than their cohort at hard difficulty on all but the blank visual representation. None performed better than their cohort at this difficulty in 2018 when more complex utterances were introduced. For games which endeavour to teach a single dominant call per species, motivating engagement with as many variants as possible produces valuable learning trajectories. With more complex training data, gamified mechanics ought to predispose participants to focus longer on each variant before continuing. For the second cohort, the higher scoring requirement for progression, noted in §5.4.2.1, was insufficient to ensure learning, given the more complex dataset at harder

[29] 2018 UID's: 80, 30, 101, 106, 50, 104, 37.

[30] 2017 UID's: 76, 77, 86, 90, 93, 96, 98, 99, 100, 101, 102, 125, 132.

[31] 2018 UID's: 29, 31, 33, 34, 35, 36, 38, 46, 47, 54, 75, 78, 84, 87, 89.

Mean L_s scores comparing cohorts to respective subsets [C]							
Variant		2017			2018		
Difficulty	Representation	Cohort	Subset C	% +/-	Cohort	Subset C	% +/-
Easy	Avian Images	0.132	0.174	+32%	0.145	0.173	+19%
	Spectrogram	0.173	0.209	+21%	0.185	0.154	-17%
	Blank	0.092	0.154	+67%	0.127	0.173	+36%
Medium	Avian Images	0.072	0.062	-14%	0.051	0.064	+25%
	Spectrogram	0.089	0.137	+54%	0.044	0.064	+45%
	Blank	0.039	0.039	+/-0%	0.037	0.035	-5%
Hard	Avian Images	0.034	0.048	+41%	0.020	0.018	-10%
	Spectrogram	0.038	0.051	+34%	0.017	0.017	+/-0%
	Blank	0.029	0.009	-69%	0.017	0.017	+/-0%

Table 6.8: Comparison of mean L_s scores for each subset C participant with cohort mean for specified variants. Blue denotes subset performs better than cohort mean, red denotes worse performance. Green, no change in mean.

difficulties.

6.3.4.3 Correlating engagement with learning from SpectroPuzzle

Does continued engagement with the less-constrained goal-state game SpectroPuzzle lead to ongoing learning?. Is data-representation familiarity prerequisite for game success? Are goal-states sufficiently self-evident that comprehension rapidly plateaus? Learning results from the nearly 200 iterations of SpectroPuzzle played are presented in §6.2.3.1, I here investigate whether the subset of participants who were most motivated to play learned more. I select the subset who played <5 iterations of SpectroPuzzle[32], calculate W_ϵ (5.8), for this subset over the first two iterations of play, and see whether continued play increases W_ϵ . By comparing subset mean success to cohort mean success, I identify whether the subset differs significantly from the cohort (see table 6.9, pg. 195). I observe that those who played most were initially significantly worse, regardless of difficulty. For subsequent iterations at all but easy level, I can no longer conclude that these participants differ from their cohort. Motivated by desire for learning, those for whom the games support knowledge development repeated play.

6.3.4.4 Evaluating engagement through open-ended play in ChorusComposer

ChorusComposer was presented to the first cohort only. Observing participants who interacted with this toy allows me to examine whether play motivates participants who prefer open-ended exploration to goal-state achievement while interacting with spectral data. Unlike games for which solutions exist, the sound toy encourages creation of

[32] 2017 UIDs: 52, 65, 66, 76, 92, 93, 94, 98, 100, 124.

Cohort vs. subset mean W_ϵ (μ_{W_ϵ}) scores over first two iterations of play						
Difficulty	First Iteration of Play			Second Iteration of Play		
	Cohort μ_{W_ϵ}	Subset μ_{W_ϵ}	Wilcoxon's p-value	Cohort μ_{W_ϵ}	Subset μ_{W_ϵ}	Wilcoxon's p-value
Easy	0.427	0.287	$3.37 * 10^{-5}$	0.513	0.354	0.016
Medium	0.521	0.348	$1.00 * 10^{-5}$	0.550	0.411	0.395
Hard	0.476	0.340	$2.54 * 10^{-4}$	0.540	0.376	0.077

Table 6.9: A comparison of mean success scores for the high engagement subset [C] with the entire 2017 cohort. Blue reflects significant differences between cohort success and subset success. While those most engaged performed significantly worse ($p < 0.05$) than cohort mean when first exposed to the game, this ceased to be the case with repeated play at the medium and hard difficulties. Desire for enhanced knowledge, required more by those with low initial success, motivates engagement and increases learning success. Conversely, increased learning success motivates continued engagement.

abstract compositions built through overlaying avian utterances. Composition interaction presupposes that participants are capable of predicting audio output by viewing sample spectrograms.

Participants who engaged with ChorusComposer created numerous compositions of varying complexity, 22 are considered in this analysis. Engagement metrics, introduced in section §5.4.3.3, equation (5.9), as a function of turns, and (5.12), as a function of time, quantify the potential for increasing participation through play. Both mean and median results for time spent and engagement scores, a function of time and compositional complexity, increased between the first two composition iterations. While some participants

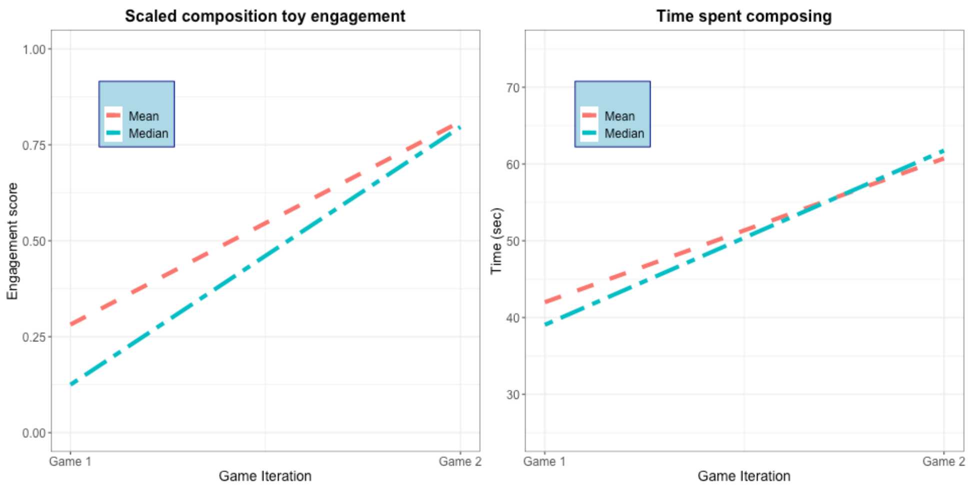

Figure 6.23: Time spent and scaled engagement quotient are presented between the first two iterations of composition by those participants who engaged with the sound toy. Engagement scaling to [0, 1] results from dividing by the maximal engagement score for each iteration.

failed to engage with ChorusComposer, most who did created increasingly complex compositions in subsequent iterations of play, once they understood the toy's interaction mechanics. Engagement increased as a function of composition duration from $\mu = 41.99$ seconds for the first compositions to $\mu = 60.72$ seconds for the second compositions (fig. 6.23, pg. 195). Likewise the mean number of samples used between the first two compositions increased from $\mu = 13.84$ to $\mu = 24.66$. The open-ended nature of the sound-toy, which lacks gamified rewards or limiting interaction mechanics, motivated only a subset of participants. However, those who engaged with the available interactions became more involved upon repetition and familiarity and found creation motivating. The relative success of this design validates the incorporation of creative interactions with data when new knowledge is taught.

6.3.4.5 Game progression evolution

While those who played the most total games consistently played BirdMatch most, due to a combination of the play protocol introduced in §5.4.2.1, the addictive motive for repetition, and overt gamified rewards, those who played the greatest game variety are also of interest (table 6.10, pg. 197). Selected participants[33] had: played at least one BirdMatch variant; played multiple iterations of at least two SpectroPuzzle difficulty levels; and engaged with ChorusComposer. This subset, selected for their motivation to play as many games as possible, achieved similar BirdMatch L_s to their cohort. While in the spectral data-representation/easy level, L_s was significantly higher ($p = 6.46 * 10^{-3}$), this apparent comfort and familiarity with spectral representations did not result in higher SpectroPuzzle W_ϵ. Across all three difficulties, SpectroPuzzle W_ϵ was significantly below cohort average. This suggests that multiple SpectroPuzzle iterations ought to be played (which this cohort failed to do) for learning effects to occur. Competence in BirdMatch spectrogram mode, while sufficient for teaching the data-representation, provides no basis for comprehending SpectroPuzzle game-mechanics or goal-state. These participants, while above average with BirdMatch spectrograms, albeit less competent when offered relaxed goal-state requirements, continued to engage with the spectral data-representation through the composition toy. As few participants played more than 1 iteration per difficulty of, and had been less successful at the outset with, SpectroPuzzle, there is insufficient data to determine whether they would have built upon their success from initial BirdMatch exposure. Observing performance similarities between this subset and the cohort when engaging with the sound toy, I contend that the third interaction mode provides motivation for participants less rewarded by goal-state mechanics, who nevertheless remain motivated to engage with spectral data. Participants motivated to pursue all possible game combinations with minimal success in prior data-representation modes or difficulty levels represent a behavioural trajectory that engages with game-mechanics primarily and game training intentions secondarily.

[33] 2017 UIDs: 3, 111, 118, 119, 120, 123, 129, 131.

2017: Most Engaged Participants Compared to Cohort by Game						
Game Design	Comparison Metric		Cohort Mean μ	Subset Mean μ	% +/-	Wilcoxon's p-value
Match	E_q		0.675	0.789	+17%	0.9989
	L_s Easy	Avian Image	0.132	0.203	+54%	0.9987
		Spectrogram	0.173	0.138	-20%	0.9998
		Blank	0.092	0.057	-38%	0.9912
	L_s Med.	Avian Image	0.072	0.127	+76%	$6.46{*}10^{-3}$
		Spectrogram	0.089	N/A	N/A	N/A
		Blank	0.039	N/A	N/A	N/A
	L_s Hard	Avian Image	0.034	0.129	+279%	0.9771
		Spectrogram	0.038	N/A	N/A	N/A
		Blank	0.029	N/A	N/A	N/A
Puzzle	Easy Difficulty W_ϵ		0.453	0.333	-26%	$3.97{*}10^{-6}$
	Medium Difficulty W_ϵ		0.510	0.371	-27%	$3.82{*}10^{-7}$
	Hard Difficulty W_ϵ		0.532	0.458	-14%	$7.19{*}10^{-3}$
Compose	Game 1 Time		41.99	39.88	-5%	0.517
	Game 1 Turns		13.84	15	+8%	0.526
	Game 1 Engagement		837.6	969.5	+16%	0.539

Table 6.10: Comparison of the subset of participants who are most engaged across all games with the 2017 cohort. For each game, relevant metrics are compared and tested for significance. Blue represents better subset performance, red worse, green highlights those differences which are significant. Yellow denotes insufficient data.

6.3.5 Summary: questions regarding motivation

In summary, I explored mechanisms for increasing avocational participants' engagement in a regional conservation project through games. Strong agreement that my games were enjoyable and 'satisfyingly addictive' supports my contention that game-mechanics motivate engagement (§5.5.2.1). Implementing new local datasets for these games is unlikely to diminish engagement, as participants agreed my games provided motivation to search for, and the likelihood of reaching, goal-states. Therefore game-mechanic adaptations are not necessary.

Play did not significantly change initially high mean interest in learning about wildlife or desire for engagement with nature (§5.5.2.2). The marginal decrease in second cohort participant belief that knowledge of nature increases enjoyment of nature may reflect the risk of providing a motivational interface for avian knowledge acquisition which does not require *in situ* engagement with nature. However, these results do not reinforce previously identified risks associated with technologically mediating nature engagement. Second cohort mean interest in learning about owls rose post-play. Participants' somewhat strongly agreed desire for knowledge validates introducing students to owl calls through games.

BirdMatch, with simple game-mechanics and clearly defined goal-state, was strongly favoured as easiest, while the composition game was considered most challenging showing how varying data-representation familiarity and goal-state complexity affect motivation (§5.5.2.3) . Game preference results support enabling users to develop spectral data-representation comprehension prior to pursuing more complex game-mechanic interactions.

For games which endeavour to teach a single dominant call per species, motivating engagement with as many variants as possible produces valuable learning trajectories. However, with more complex training data, gamified mechanics ought to predispose participants to focus longer on each variant before continuing (§5.5.2.4). Examining BirdMatch performance of subsets of highly motivated participants, I found both cohorts' most engaged subsets achieved higher mean L_s scores across BirdMatch on aggregate and at each difficulty level. Participant motivation correlates with increased learning, regardless of difficulty. As the SpectroPuzzle goal-state path is less rigid, §6.3.3.3 results support my contention that confidence correlates with goal-state comprehension. Motivated by desire for learning, those for whom SpectroPuzzle supported knowledge development repeated play. Sound toys, supporting engagement with data without a goal-state, intrinsically motivate participants through creation of aesthetically interesting artefacts. The open-ended nature of the sound-toy, which lacks gamified rewards or limiting interaction mechanics, motivated only a subset of participants.

6.4 Engaging stakeholders in collaborative design

Through focus groups with local and regional individual, non-governmental organisation (NGO), and governmental stakeholders, I endeavour to identify how my games support individual and institutional needs and how multiple stakeholders can contribute to collaborative design, creating effective interactions for citizen scientist training and engagement. In 2017, prior to project launch, I recorded conversations with two focus groups. The first elicited feedback from local farmers, gamekeepers, and landowners, and representatives from the Nidderdale AONB, the Game and Wildlife Conservation Trust (GWCT), and the Yorkshire regional government. The second targeted a community choir whose members, while not presumed to have prior motivation for conservation engagement, provided insight into the aesthetic value of birdsong within the community and identified contributions people not immediately involved in local ecology could make. In 2018, local participants who had engaged as citizen scientists with the Wild Watch project in its first year were invited to provide feedback. All groups were prompted to discuss the topics introduced in §5.4.4. I look at how these discussions inform the questions asked in §5.5.3. I examine stakeholders' roles as participant collaborators, the feedback they desire for contribution validation, and whether design feedback augments learning through play.

In response to the research question from §5.5.3.1 I will present results exploring familiarity with and concerns about citizen science across diverse focus groups, and findings of whether avian bioacoustic citizen science was considered novel. Prior projects' failure to successfully market their existence remains problematic. In response to the research question from §5.5.3.2 I will identify how collaborative design yields more engaged stakeholders and present expert participants' suggestions including that potential educational games should target visitors who lack prior local knowledge, not only the local populace.

6.4.1 Research examination: familiarity with & concerns about avian citizen science

Prior to contributing opinions to my evolving citizen science training game designs, focus group participants discussed whether they were cognisant of the roles citizen scientists play in avian conservation projects and whether they considered contributions to such projects valuable. What follows is a summary and critique of each group's conceptual familiarity with citizen science and whether participants found value from avian utterance recognition training prior to contribution to the Wild Watch. Participants were initially prompted to discuss the term 'citizen science'. This triggered discussion of whether they had prior knowledge of, or had previously contributed to, citizen science projects.

6.4.1.1 Stakeholder familiarity with avian citizen science

The first focus group comprised some participants with prior involvement in citizen science data collection for avian surveying and some for whom this was novel. One landowner had organised a pilot programme where gamekeepers collected avian presence data along trap lines in collaboration with the British Trust for Ornithology (BTO); this involved gamekeepers performing 90% of data collection, whilst BTO professionals provided data validation. Such methods offers scientific validity, as trap lines represent replicable, albeit nonlinear, transects. Such transects require new data-modelling approaches as most prior models rely upon linear or grid transects. Collaborating with gamekeepers, who are intrinsically protective of their lands, necessitates mechanisms for maintaining trust. Such collaborations offer data-collection potential while minimising additional environmental impact. The GWCT representative noted that their organisation had developed approaches to counting presence by call, contended erroneously that the BTO did not apply such methods, and noted that breeding bird surveys undercounted, as only adult pairs were counted, without considering presence or number of chicks. The GWCT were only interested in game birds and lacked coherent knowledge of the current state of surveying. The majority of remaining farmers and landowners were unfamiliar with citizen science as premised by the Wild Watch. Educating such stakeholders, both those with and without prior citizen science familiarity, remains a valuable long-term policy goal. A number of participants in the second 2017 group, musicians, were familiar with citizen science projects, citing BBC weather watchers[34] and the Royal Society for the Protection of Birds (RSPB)'s Big Garden Birdwatch (BGBW). They lacked prior avian bioacoustic knowledge or citizen science practice but expressed interest in my games' training potential. The 2018 focus group was comprised of people who had participated in some Wild Watch citizen science data collection events. While their familiarity was presumed, I was curious to identify the value they associated with citizen science data collection and concerns participation raised. Across these diverse focus groups, avian bioacoustic citizen science was generally considered novel, despite prior work in the domain (see §3.3.3.3); prior projects' failure to successfully market their existence beyond existent avocational ornithologists remains an issue.

6.4.1.2 Stakeholder concerns with my design implementations

A common concern for the first focus group, which had the most prior citizen science data collection project exposure, was that 'citizen science' was frequently used to crowd-source contributions, without either a commitment to scientific results validation or benefit to the crowd. The Wild Watch's representative identified their project goal as collecting sufficient survey data for evidence-based assessment, while I identified my games' purpose as developing participants' knowledge, supporting validation of collected data. Participants appreciated involvement in preliminary discussions regarding contributions they, as citizen scientists, might make and were interested in the role mobile interfaces could play. They were broadly supportive of my teaching interface designs, with one participant reporting that we had visited their child's school and the child reflected positively on the games' motivational affordances. Specific contributions and concerns are discussed in the following section. Underlying concerns about eventual uses made of contributions were again identified by the second focus group. They asked for clarity regarding ownership of contributed data and, correspondingly, whether contribution would be rewarded. Further concerns involved tracking and concurrent loss of privacy should collected data be geolocated. This group was primarily older and expressed unfamiliarity and potential discomfort with mobile devices as data collection tools. When playing my games they indicated satisfaction with the interactions enabled, and found the short time requirements for a single training game to support dabbling motivation. Chief amongst the concerns of the 2018 group who had participated in data collection exercises were ownership of collected data and how participants might be credited. While these participants were generally middle-aged and more familiar with mobile devices than the 2017 musical focus group, they similarly conveyed concerns regarding mobile software

[34]www.bbc.co.uk/weatherwatchers

tracking and corresponding privacy loss. For now, as my game designs primarily elicit training and engagement with avian utterances, there is no requirement to geolocate participants during interactions. Developments incorporating user recording and audio upload, implemented and discussed in §4.3, may require this. In keeping with conservative data acquisition protocols, participants were informed that my software would only track locations upon active application interaction, and users would be warned and asked to confirm comfort with recording geo-location, if enabled. Another concern for all groups was the ownership of collected data. The Wild Watch representative reassured participants that collected survey data would be publicly available, while I noted that, although game data collected in schools could not be public due to educational privacy rules, my resulting software designs are open-source and available for download[35].

6.4.2 Research examination: does collaborative design yield more engaged stakeholders

Do focus group participants desire to contribute? Do they think their game design contributions influence project outcomes? Does this desire motivate engagement? Design contributions will be assessed by summarising and critiquing focus groups' responses regarding their perceived roles as citizen scientists in conservation and their self-representation as engaged stakeholders. Cumulatively, ∼60 participants, introduced in §5.4.4, responded. Ultimately, both novice and expert focus group contributors reacted favourably to the process by which their contributions to iterative game design were introduced. They expressed satisfaction that they were playing an integral role in problem definition and could positively influence project outcomes through collaborative design.

6.4.2.1 Experts' design considerations

Several gamekeepers in the first focus group expressed concern that, while they are confident with their local avian utterance knowledge, they were uncertain such knowledge was sufficiently static to be conferred through games. An example given was the curlew, a project target species, which has over 30 distinct vocalisations. Several gamekeepers contended that these utterances are inordinately complex for citizen scientist identification as they can be context-dependent and may so vary between individuals in a population that game ground-truth is untenable. This highlights concerns that limiting game scope may require unfeasible standardisation of target knowledge. While I noted that my game designs constrain participant exposure to professionally identified utterances, several gamekeepers were concerned that ambiguous identifications might be erroneously standardised. I

[35] Compiled games at https://bioacoustic.games and underlying source code with first and second cohort datasets on GitHub at https://github.com/isakh/BridgeGames and https://github.com/isakh/BridgeOwls

clarified my games' initial purpose as training citizen scientists to provide ground-truth for presence models; as these models require only species identification, not utterance intention identification, each species' least ambiguous call suffices for data collection training.

Another proposed design goal was to encourage better land-user behaviours through games. The GWCT representative posited considerable value in teaching the purported intention of certain call types. Familiarising the local population with nesting and alarm calls may reduce impacts on vulnerable populations by informing land-users where they should not go. Participants in the first group discussed the potential for games to educate both the local populace and people who visit the region without prior knowledge of the countryside. One gamekeeper, identifying the desire to teach farmers to recognise call types to trigger changes in farming behaviours and practices such as not mowing silage when this interrupts nesting, stated: 'we need to be thinking about what can be changed in farm scheduling and behaviour as [many practices are] no longer necessary and [are] only done because of historic behaviour.' Several farmers countered that their schedules require practices which inevitably interfere with breeding seasons.

The GWCT representative proposed a class of game interactions be designed to correlate land-user actions with avian observations. An estate manager noted that since nearly all AONB land is controlled by estates or farmers, developing a stakeholder knowledge-transfer framework is imperative. The Wild Watch representative reiterated their project goal of collecting sufficient data for habitat suitability modelling (HSM), with the long-term outcome of identifying new viable habitats for at-risk species. This elicited objections from both gamekeepers and farmers that increasing avian populations without effective means of population redistribution was unwelcome, insofar as they associated this with increased predation; this claim is contentious as a type II or III functional response will reach a limit[36]. Participants considered land-use conflicts of interest, and asked how stakeholder needs can be integrated with model output to inform policy. Participants were enthusiastic that the project was eliciting their contributions from the outset and expressed hope that their concerns could influence long-term goals.

In conclusion we discussed how success might be measured in light of changing environmental pressures, without imposition on land-users and owners. The information mismatch can only be overcome if there is a mechanism for inter-stakeholder information transfer. While gameful interaction mechanisms can validate stakeholder knowledge, concerns remained as to whether results will be used solely for monitoring, or could instigate unintended land-use policy changes. Further game development requires local experts' knowledge, while my games' target audience, from a motivational learning perspective, is primarily new citizen scientists rather than professional and avocational bird watch-

[36]Functional responses [102] describe predation rates as a function of prey density. A type I response is linear, type II rectangular hyperbolic, while type III builds upon type II but proposes a prey density saturation level beyond which predation will not increase.

ers. The purpose of these focus groups was to build a network of engaged stakeholders capable of providing the information necessary to devise interactive applications for the aforementioned audience. Experts expressed interest in offering design contributions and have continued to provide influential feedback.

6.4.2.2 Novices' design consideration

A participant in the musicians' group noted that invasive species were now in his garden, part of a complex community full of many species which need preservation. He asked whether the project could offer mechanisms for incorporating non-target species information. A farmer in the first focus group said that he hears bird calls as background noise; as he, in contrast with gamekeepers, lacks knowledge, he does not pay significant attention to birds. A potential role for my games was identified as supporting novices through active listening and subsequent learning. Distinguishing between target and non-target species was brought up as a concern. Several farmers in the first group stated that despite lacking expert knowledge, they want to preserve environmental birdsong as it provides an emotional dimension to their attachment to the land. Learning a limited number of calls and songs for identification would be rewarding for them. The most common design adaptation proposed by novices was for interfaces to let users extend data-sets. While I discussed the potential for collaborative filtering to be applied to validation of extended data-sets, the scope of my current research relied on experts for call identification. While such extensions are of interest, particularly as collection engages and motivates avocational citizen scientist participation, incorporation of non-target species into data-sets reduces potential scientific output in the absence of suitably dynamic models.

Participants in the musician focus group asked whether the composition game could be played earlier in the learning process, as they were more familiar with this approach to musical learning. In iterative designs for broader release, play-order constraints may be relaxed. 2018 participants, who had some prior data-collection exposure, asked whether links to additional target-species information, such as seasonal behaviours and habitats, might be presented once auditory recognition was mastered. While this might distract from flow within single-iteration play, adding links upon mastery can be incorporated into subsequent releases.

A final design concern common to both musicians and citizen scientists, both of whom were on average older, was whether the visual representations and scale could be adapted for participants with poor vision. As my original game designs were targeted to primary students, visual constraints had not played a significant role in my design criteria. While designing for significantly variable screen sizes remains future work for subsequent broader release of the games, general android adaptations for those with poor vision, including speech to text output for species identification in BirdMatch and SpectroPuzzle, are now implemented.

6.4.3 Summary: questions regarding collaborative design

In summary, exploring familiarity with and concerns about citizen science (§5.5.3.1), I found, across diverse focus groups, that avian bioacoustic citizen science was considered novel. Prior projects' failure to successfully market their existence to those not already avocational ornithologists remains problematic. Participants appreciated opportunities beyond data-collection, including project design-feedback contributions.

Collaborative design yielded more engaged stakeholders (§5.5.3.2) with expert participants suggesting potential educational games targeting visitors who lack prior local knowledge, not only the local populace. Novices likewise engaged, proposing design adaptations, including varying visual representations and scale for participants with poor vision, and providing ancillary content regarding learned species. Focus group participants desired to support avian conservation citizen science, believed my designs teach effectively and motivate participation, and thought their contributions would influence project outcomes with games designed to benefit individual and project goals.

6.5 Summarising learning efficacy from gameplay

Results from §6.2.1.1, summarised in §6.2.4, show that sound training adds value. My game designs introduced acoustic visualisation analysis to participants who had not previously considered multi-sensory identification. The greatest learning success occurred with multi-sensory input; while participants trusted visual learning more, spectrograms reinforced auditory learning.

Participant confidence that the games developed visual and audial knowledge was supported by the second cohort's pre- and post-play identification task results. Decreases to BirdMatch mean L_e and corresponding increases to mean L_s, with repeated play for all data-representations, were summarised in §6.2.4. Data-representation difficulty rankings reinforced the expectation that participants were biased towards visual learning and that spectrograms bridged the gap between learning visual and audial identification. SpectroPuzzle learning was observed at the α-transition regardless of difficulty, showing that initial BirdMatch exposure was sufficient to familiarise players with spectral data-representations.

Self-perception of increased knowledge from repeated play increases motivation for participant engagement and correlates with learning-effects observed across data-representations. Age was not correlated with variable performance; my designs are suitable across the target age range. In BirdMatch I find that engagement and flow correlate with increased learning, regardless of difficulty; however, L_s maximisation approaches vary with training data-set complexity. Specifically, with the first data-set, exploring the most variants yielded highest L_s, whilst with the second data-set, where more nuanced utterances were introduced, repetition of a single variant led to L_s maximisation. Learning effects remained evident when goal-state constraints were relaxed in SpectroPuzzle.

CHAPTER 7

Conclusions, discussion, & future work

My research has developed and applied a collaborative design approach grounded in theories of motivation and play for developing applications that support conservation engagement and learning for citizen science. This chapter summarises my design explorations and discusses the validity of the frameworks through which they were analysed. The discussion of research contributions includes implications for human computer interaction (HCI) research for citizen science and also reflection on the biases embedded in my collaborative research process. Finally, I consider the potential for iterative design, deploying my software in public-facing citizen science projects and educational environments.

7.1 Summary & Conclusions

The previous chapters have presented results including: validation that primary students can interact with spectrogram data-representations; validation of games as engagement tools for connecting people with nature; validation of games as tools for training participants otherwise uninterested in avian bioacoustics; and validation of short-term learning from games, prerequisite for amateur contributions to science. These results have demonstrated that interaction design research methods can be incorporated into citizen science research. They also demonstrate the effectiveness of collaborative processes for interface design in this application domain. Chapter 2 presented two frameworks for design and analysis of citizen science games. I present conclusions regarding the validity of my design frameworks as contributions to citizen science design research, evaluate the efficacy of my software interface implementations, and clarify conclusions regarding research explorations introduced in §5.5.3. I identify what level of data-representation complexity best supports analysis and region of interest (ROI) selection, whether games can effectively teach novel data-representations, and clarify conclusions regarding research explorations introduced in §5.5.1. I summarise results from my investigation into whether my software interfaces enabled interactions enhancing learning while motivating engagement with citizen science and clarify conclusions regarding research explorations introduced in §5.5.2.

7.1.1 Design results & conclusions

I designed frameworks for classifying participant motivation with game-mechanic, data-representation, and goal-state complexity variables which describe player engagement in citizen science games and developed a software framework for rapid prototyping of games which teach localised bioacoustic datasets. I have created interaction designs which fulfil the United Kingdom (UK) academic prerequisites for a *structured experience unit* teaching children about *working scientifically*, including methods of data collection, data analysis, and extracting meaning from data. My collaborative design approach has taken into account multiple stakeholders, including land users and managers, and has been sensitive to risks that games may motivate interaction but not optimise project on-boarding. If initial encounters with nature are device-mediated, this will inform subsequent encounters and while visual feedback can aid auditory learning, screen-mediated interactions limit participants' observation of surrounding nature; I explored audio-focussed interaction potential.

7.1.1.1 Research framework design considerations

In chapter 2, (fig. 2.7, 2.8) I proposed novel frameworks for discussing game interaction design and motivating citizen scientist engagement through games. These provided a basis for analysing my interface artefacts. My first framework delineates game designs along axes of data-representation, goal-state complexity, and solution difficulty; for where implemented games fall, see figures 5.8, 5.11, and 5.14. My second framework provides dimensions for HCI practitioners to consider when designing for public-facing citizen science which encompass sources of motivation for engagement with biodiversity through games. My HCI research has involved observations of changes in learning and motivational success, when participants play multiple iterations of several games which fill diverse spaces in these frameworks, validating axes.

7.1.1.2 Interface design results & conclusions

My research involved identification of an interface suitable for unfamiliar tasks, birdsong spectrogram ROI selection, and subsequent validation that the selected familiar interface – touchscreens – support such tasks. My design research has compared the efficacy of various data-representations for bioacoustic signals and variable game goal-states for encouraging participant learning success and potential project contribution. I have produced novel choropleth mapping preference results for bioacoustic ROI selection. I have concluded that spectral representations provide sufficient dimensionality and are comprehensible for bioacoustic knowledge development as I validate primary students' capacity for and interest in play with such representations. My data-collection interface (§4.3.2) supports interactions for novices to build libraries of recorded birdsong and my games (§5.3) support science involvement in schools with interactive approaches to extracting meaning from data. My software framework supports extensible datasets, motivating users to collect field recordings, and allows rapid prototyping of new micro-targeted location-specific training games.

7.1.1.3 Collaborative design results & conclusions

I have taken into account different stakeholder cultures' definitions of conservation and engagement and have designed for the fundamental goal of creating new citizen scientists who, upon project on-boarding, progress further through the motivational arc than they would if the project were not augmented by my designs. I have considered the needs and wants of multiple stakeholders – including educators, conservationists, and various land-users – while evaluating impacts of conservation engagement through education. My work in schools has supported the ancillary need to meet curriculum requirements while engaging students with avian bioacoustics. Participants appreciated opportunities beyond data-collection, including project design-feedback contributions.

Collaborative design processes engaged stakeholders with expert participants, suggesting potential educational games targeting visitors who lack prior local knowledge, not only the local populace. Novices likewise engaged, proposing design adaptations, including varying visual representations and scale for participants with poor vision, and providing ancillary content regarding learned species. Land managers' contributions have identified the need to regulate potential environmental impacts of citizen science data-collection on land being conserved. Focus group participants desired to support avian conservation citizen science. They believed my designs teach effectively and motivate participation. They likewise thought their contributions would influence project outcomes with games designed to benefit individual and project goals.

7.1.2 Learning results & conclusions

Results from §6.2.1.1 support my contention that sound training adds value for citizen science data-collection. My games introduced acoustics to those participants who had not previously considered multi-sensory identification. The games were perceived to enhance visual and audial learning, yielding greatest learning success with multi-sensory input. This result supports incorporating visual audio depictions — spectrograms — into training. Results do not show a single play session eliciting significant shifts in perception of multi-sensory knowledge. Participants trusted visual learning more, although spectrograms reinforced auditory training.

7.1.2.1 Learning exploration results

Participants were confident that my games developed visual and audial knowledge across data-representations. Their confidence was supported by the second cohort's pre- and post-play owl identification task results, changes to which were positive for 3 of 4 species, indicating that my games have training value. In BirdMatch, both cohorts saw broadly decreasing mean L_e and correspondingly increasing mean L_s over the first three iterations of play across data-representations. Regardless of data-representation, learning effect strength declined after initial exposure but remained positive. BirdMatch representation difficulty ranking results showed the first mode overwhelmingly considered easiest and the third hardest, supporting my contention that visual data-representations enhance and ease learning. Examining BirdMatch performance of highly motivated participants, I found repetitive play better supported learning across difficulties and data-representations for the increasingly complex second data-set utterances, although effects remained positive for both cohorts. SpectroPuzzle learning was observed at the α-transition across all three difficulties. This may be attributable, at easy difficulty, to learning game-mechanics rather than increased data-representation comprehension, but implies learning spectrogram content at harder difficulties.

7.1.2.2 Learning exploration conclusions

Self-perception of increased knowledge resulting from game performance increases motivation for further project engagement, although survey data-quality validation is necessary prior to data inclusion in project databases. Learning effects were present but strength varied with data-representation. Exploring learning through play, I found no age dependency, demonstrating that my designs are suitable for students across key stages. As in-school experiments do not capture data necessary to quantify long-term learning, for the second cohort I built a website with a follow-up survey[1] asking participants to again complete the species identification tasks from the pre- and post-surveys; poor uptake produced insufficient responses for analysis. I find that engagement and flow correlate with increased learning, regardless of difficulty. With the first data-set, exploring most variants yielded highest comparative L_s, whereas with the second data-set, single variant

[1] http://bioacoustic.games/html/owl_survey.html

repetition yielded most increased L_s. Learning effects remained present when goal-state constraints were relaxed in SpectroPuzzle. As the 2-row version of SpectroPuzzle provides primarily game-mechanic training, and subsequent feedback has requested game-mechanic explication, this version may become an optional stand-alone training mode in future work.

7.1.3 Motivation results & conclusions

The desire for engagement with nature is widely promulgated as a core value of projects in the conservation domain. Conservation organisations seek to operationalise their understanding of engagement in order to define metrics for evaluating project success. Motivating engagement with citizen science projects presupposes participant desire for engagement with nature. Projects enabling participant feedback create a discourse between scientists and participants, giving participants agency in subsequent design iterations. My research explored increasing motivation for nature engagement through development of data-collection applications and data-interaction games. I formalised structures for citizen scientists to learn about acoustic biodiversity as a form of ecological engagement. I investigated how project participants are motivated to engage by direct experience of nature and by gamised rewards which provide a mixture of intrinsic and extrinsic motivation through internalisation of positive feedback. However, I cannot claim that associating avian recognition with enjoying nature significantly changed with play, perhaps because games enabled birdsong exposure without going outside.

7.1.3.1 Motivation exploration results

Strong agreement that my games were enjoyable and 'satisfyingly addictive' supports my contention that my game-mechanics motivate engagement. While play did not significantly change reported mean interest in learning about wildlife or desire for engagement with nature, these were high both pre- and post-engagement. Results do not reinforce previously identified risks that technologically-mediated interactions with nature misdirect engagement to the interface. Game preference results support enabling users to develop spectral data-representation comprehension prior to pursuing more complex game-mechanic interactions. For games which endeavour to teach a single dominant call per species, motivating engagement with as many variants as possible produces valuable learning trajectories. Participant motivation correlated with increased learning, regardless of difficulty. With more complex training data, gamified mechanics should predispose participants to focus longer on each variant before continuing. Confidence in winning correlated with goal-state comprehension across games. A positive learning feedback loop developed for those who found SpectroPuzzle that play enhanced knowledge development. Open-ended sound toys — supporting engagement through mechanics and iterative feedback while lacking a goal-state — motivated only the subset of participants interested in the creation of aesthetic artefacts, and may be better suited to an older audience.

7.1.3.2 Motivation exploration conclusions

My analyses examined whether gamified rewards correlated with increased motivation for engagement. Initial gamified interactions provided extrinsic motivation for participation. I investigated whether these are a viable way to encourage on-boarding. I found that my games yielded further declared interest above a baseline for prior engagement with educational games, although interest in nature and desire for engagement with wildlife were not significantly affected through play. I found no age-dependency for motivation to play across data-representation and difficulty variants in BirdMatch and while L_s varied as different trajectories were pursued, participants most motivated to engage, regardless of trajectory, developed knowledge. I found that constrained goal-state games continued to motivate engagement. While mechanics extrinsically motivate desire for play, they may demotivate non-gamified engagement with avian utterances, thus decreasing engagement with nature, when nuanced content, such as multiple calls per species, is taught.

7.2 Discussion

I discuss the scope of my research contributions, focussing on the relevance of my design framework and the effectiveness of my designs for promoting learning and engagement with bioacoustic citizen science. I summarise implications, identifying design efficacy and considering directions for future investigation. I conclude with a consideration of the limitations of my research, identifying gaps and biases inherent to my methods, particularly those associated with research in educational settings, as well as possible divergent explorations and questions that remain open-ended.

7.2.1 Research contributions

My research has contributed to clarifying and expanding the role of HCI in avian bioacoustic citizen science projects. I have generated novel design analysis frameworks encompassing game-design dimensions and the role of play in motivating engagement with citizen science. I have performed novel analysis of the efficacy of choropleth maps for representation of, navigation through, and ROI selection with bioacoustic spectrograms. I have validated the efficacy of several game-mechanics for learning birdsong and explored mapping game performance to motivation for non-game activities relevant to avian citizen science.

7.2.1.1 Efficacy of design framework & collaborative design

My bioacoustic game analysis framework codifies dimensions which combine familiar and novel features guiding subsequent design implementations. Varying difficulty is a game design convention which allows my analyses to confirm expected learning behaviours and test for boundaries on learning complexity in games for dabblers. While goal-state complexity has been varied in prior game design research, exploring this dimension is novel in citizen science games as most projects endeavour to expedite data-production from games, rather than to engage multiple player archetypes. Varying data-representation in citizen science training games is a novel approach which has allowed me to investigate spectral representations as visual support to ear-training. My engagement framework introduced novel dimensions for discussing sources of participant motivation and scope of participant engagement with biodiversity. These design frameworks guided my game prototyping software which enabled me to explore the roles of game-mechanics, data-representation, and goal-state complexity, in engaging participants who lack intrinsic motivation for citizen science. These dimensions and subsequent game designs bolster the argument that games enhance participant motivation to become citizen scientists as they increase baseline knowledge prerequisite for project contribution.

In response to questions of familiarity with and concerns about avian citizen science I found participants interested in engaging. Concerns included ownership of collected data, loss of privacy inherent in data-collection protocols, and whether project design focussed on crowd-sourcing contributions without concomitant commitment to scientific result validation and benefit to the crowd. In investigating whether collaborative design yielded more engaged stakeholders, several experts expressed concern that, while they were confident in their local avian utterance knowledge, they were uncertain such knowledge was sufficiently static to be conferred through games. This highlights concerns that limiting game scope may require unfeasible standardisation of target knowledge. Participants were enthusiastic that the project was eliciting their contributions from the outset and expressed hope that their concerns could influence long-term goals. In iterative designs for broader release, play-order constraints may be relaxed as these, while valuable for research analysis, had constrained certain user archetypes, such as the musicians who desired to engage with ChorusComposer from the outset.

7.2.1.2 Efficacy of learning games

Participants were confident that my games developed avian knowledge, irrespective of data-representation. While prior visual familiarity was consistently greater than audial familiarity, in neither cohort was this difference significant. Continued audio training though play aids learning. Although both cohorts marginally believed that they could recognise taught species by sound post-play, they significantly more strongly believed in future visual recognition. Regarding the effects of data-representation on learning, the first cohort showed strongest learning effects with the avian image data-representation, likely due to lack of ambiguity between images and calls, while the second cohort showed the strongest continued learning effect with the spectral data-representation. This supports my contention that spectrogram images benefit utterance comprehension in cases of teaching one-to-many relationships between species and calls. The most effective data-representation depends on ground-truth complexity. With short, static, ground-truth utterances, participants successfully increase their underlying knowledge with play. While my games sufficed to provide short-term knowledge acquisition for young students, testing for long-term retention remains: I constructed a web-based follow-up test for the second cohort to validate long-term learning. Issues acquiring data through schools after our permitted study hindered collection of data sufficient to validate long-term learning.

Generally, learning effects were strongest in the first iterations of play, but distinguishing learning utterances from learning game-mechanics was beyond the scope of my research. BirdMatch L_s scores on aggregate increased over multiple iterations of play; however, for some individuals, an initial spike was followed by a subsequent performance trough before knowledge development eventually plateaued. SpectroPuzzle success, which required comprehension not only of more complex game-mechanics but of more complex data interactions, appeared better suited to older players, although no strong age-dependency was found.

My game designs produced interactions which supported dabbling while enhancing learning of pre-selected ground-truth. While the software framework supports extensible data-sets, I identified issues with construction of ground-truth for data-sets encompassing multiple utterances per species and when I attempted to classify each utterance (*e.g.* alarm, flight, nesting and other calls). Ultimately, my training game designs ignore ambiguities which arise in interpretating birdsong when experts or amateurs are asked to classify utterances. My games teach a ground-truth, whether such a thing exists is a topic for ornithologists.

7.2.1.3 Efficacy of games as motivators

Prior to development of my games, whether unfamiliar spectral visual representations would support engagement or be too esoteric for primary students was unknown. Examining how varying data-representation familiarity and goal-state complexity affects motivation shows that game preference results may reflect insufficient time for participants to gain familiarity with spectrogram representations as data for open-ended play, or may indicate that open-ended play goal-state complexity is too abstract for the target age-group. That slightly more participants ranked the audio-only BirdMatch data-representation mode most liked is surprising, as this provides ear-training without visual support and was presumed to be hardest. On aggregate more participants preferred some form of visual representation.

While participant interest in nature has not been shown to change significantly with play, my games have served to engage those not previously interested in educational gaming. Interest in educational games hints that implementing regional and local datasets for these games is unlikely to diminish engagement, as mechanics remain unchanged. Exploring interest in nature and avian bioacoustics, I found participants in both cohorts comfortable associating avian utterance familiarity with enjoyment of nature. Reported interest in nature and bioacoustics started high, but no results associating changes to responses to survey questions with play were significant. This may reflect the risk of providing a motivational interface for avian knowledge acquisition which does not require *in situ* engagement with nature.

For games which teach a single dominant call per species, motivating engagement with as many variants as possible produces valuable learning trajectories. With more complex training data, gamified mechanics should predispose participants to focus longer on each variant before continuing. Those who engaged with ChorusComposer became more involved upon repetitive play, and with ensuing familiarity found motivation to create compositions. Participants motivated to pursue all possible game combinations despite minimal success in prior data-representation modes or difficulty levels represent a behavioural trajectory that engages with game-mechanics primarily and game-training intentions secondarily.

My research contribution to the Wild Watch project encouraged participant feedback and presupposed the open nature of collected data. Stakeholders engaged with the scientific process and were thus more likely to trust project output. There was no age-dependency for motivation to play across data-representation and difficulty variants in BirdMatch. My game designs are suitable for engaging even young students with bioacoustic citizen science. I identified several subsets of participants, each of whom achieved flow and repeated play, albeit through different trajectories. My game framework supports both control- and autonomy-oriented participants by supplying multiple game-mechanics which enhance both intrinsic and extrinsic motivation. I found that less-constrained goal-state games also motivate engagement. Adapting regional and local datasets for these games is unlikely to diminish engagement.

7.2.2 Research implications

Considering the results previously summarised and discussed, I clarify implications of my design research results. I found that collaborative design benefited both project scientific output and local communities. Localising designs for a target community enhanced motivation to engage. Games can serve dual purposes in citizen science, motivating initial engagement and subsequent training of new users. Game narratives should not distract participants from project goals and designing for dabbling can increase project output. Future citizen science projects should involve local stakeholders in the design of interaction artefacts that motivate community formation around project goals.

7.2.2.1 Collaborative design implications

Citizen science projects benefit from incorporation of stakeholders throughout the design process. Involving local stakeholders can give conservation projects access to regions that would otherwise be off-limits, including private land. Considering results regarding familiarity with and concerns about my development of avian citizen science artefacts, I note that engaging participants in the design phase increased artefact effectiveness while enhancing participant perception of knowledge exchange, a dynamic supporting intrinsic motivation. Involving novice and expert stakeholders in collaborative design provided valuable mechanisms for engaging participants with the project prior to its public release, thus yielding more engaged stakeholders. Collaborative design generates participant interest, providing additional motivation for on-boarding, and supports the development of localised datasets, although mechanics ought to exist for expert validation of collected data before incorporation into training games.

7.2.2.2 Implications for learning game development

Familiar game-mechanics are well suited to citizen science projects as they obviate training tutorials predicted to reduce participant motivation. From results regarding confidence in claimed knowledge, I contend that multi-sensory data should be incorporated into training materials; acoustic information provides increased learning head-room given less expected prior knowledge. Implications from my exploration of the effects of data-representation on learning include that my artefact designs are sound teaching tools given significant falling error-rates; lack of concurrent significance in rising L_s-transitions introduces questions regarding metric validity. Exploring learning with relaxed goal-state constraints, I contend that data-set curation must consider circumstances where utterances with similar spectral representations cause confusion; in such cases, avian images might be more useful than utterance spectrograms. My results imply that matching is a familiar and addictive game-mechanic suited to learning through dabbling. Puzzles, requiring prior data-representation comprehension, are less suited to initial learning through dabbling and less effective as introductory games for participant on-boarding, but support increased comprehension necessary for validating scientific output.

7.2.2.3 Implications for motivational game development

Results regarding changes to interest in nature and avian bioacoustics with play, show that motivating engagement with games is not strongly linked with motivating engagement with nature. While game-mechanics extrinsically motivate desire for play, they may demotivate non-gamified engagement with avian utterances. This can decrease engagement with nature, particularly when nuanced content, such as multiple calls per species, is taught. Identifying means of motivating continued interest in nature after play remains worthwhile.

Game preference results showing how varying data-representation familiarity and goal-state complexity affects motivation, indicate that the goal-state complexity of open-ended play is too abstract for the target age-group. Simplicity and familiar data-representations were not strong intrinsic motivators. For the second cohort, the somewhat higher scoring requirements than those demanded of the first cohort were insufficient to ensure learning, with the more complex dataset at harder difficulties. This implies that adaptations to the minimal scores required for difficulty and data-representation progression are necessary for each dataset to maximise motivation while ensuring learning.

Composition complexity can provide aesthetic value which supports intrinsic motivation. Open-ended play motivates participants less driven to addictive interactions, but such designs must consider target user demographics. Therefore, age-specific game-mechanic adaptations are warranted with such designs.

7.2.3 Limitations of current research

In addition to proposed design revisions and adaptations for subsequent target demographics (see §7.3.1.1, pg. 220), several issues are worth noting in terms of the limitations of my current research. In the following section I consider biases inherent to my designs and methods, particularly regarding experimental approaches which must adhere to school requirements. I will identify gaps in the research I have performed, noting both space for future exploration and likely errors that should be avoided in follow-up explorations. I will discuss, with hindsight, choices I made in my design research path and consider branches along which further work remains relevant. I conclude by introducing remaining open-ended questions, some of which will be developed in the following section on future work.

7.2.3.1 Sources of bias

Performing classroom research introduces multiple potential sources of bias — do my games motivate engagement or should this be attributed to the environment? Collected survey data presupposed students could comprehend each question; teachers' aids helped students with reading difficulty. Analysis of results necessitated trusting student self-reporting; there was no mechanism for triangulating responses, raising concern that students over-stated interest at the outset from a desire to provide the 'right' survey answer; after a session they might have been prone to being more honest, yielding falling reported interest in nature. In light of results regarding changes in reported interest in nature and avian bioacoustics it is relevant to note that as play-order was semi-structured for teaching purposes, it was not feasible to fully distinguish whether game preference results were a function of player fatigue or game quality. As teaching tools, my artefact designs led to significant L_e declines but these might be attributed to bias from learning substitution effects — learning to match but not to recognise target species. In a more rigorous experimental setting this might have been mitigated with A/B testing on specific data-points and single game-mechanics to validate species recognition; however, this would have eliminated much of the freedom associated with play. Given time constraints for game interactions within sessions satisfying school curriculum requirements, only data exploring short-term learning were collected in schools. Testing long-term learning required building a subsequent web-based test environment, but few teachers encouraged follow-up interactions and insufficient data were collected. While my collaborative design requires gaining stakeholder trust, this is bidirectional. Expanding datasets requires trusting local experts to provide ground-truth and this can introduce bias in the absence of expert validation of knowledge; conversely there are concerns that limiting data-set scope may require unfeasible target knowledge standardisation.

7.2.3.2 Research gaps

I set out to explore a broad research space encompassing interaction design, citizen science, and exploration of motivation through play. Numerous questions remain because I primarily followed a directed path when divergent options were possible. I proposed design frameworks (§2.7) to provide a basis from which to build tools for future interactive citizen science projects and built a software development framework as a basis for rapid generation of games for targeted learning interactions in the field; additional analysis dimensions and game interactions remain to be explored. Lack of correlation between significant falls in L_e and insignificant rises in L_s means there is space for future research into why my metrics are not strongly linked and whether different metrics might produce more internally consistent results. From my consideration of the effects of data-representation on learning, I note that while overall L_s increased across the first three iterations of play, regardless of data-representation, declining learning effects after initial exposure should be explored further. Additional research could be done into the design of analysis metrics. For BirdMatch, perhaps luck over-influences success, or possibly the winning component too stringently reduces the range of possible scores, particularly as a perfect win would be cancelled out by perfect luck. While I proposed from the outset that game output could be used to construct biodiversity metrics of use to policy makers, this was ultimately beyond the scope of my research. However, users trained on my games have subsequently participated in data-collection exercises with the Wild Watch, with the ultimate intention that their data form the basis of a regional habitat suitability modelling (HSM).

7.2.3.3 Divergent exploration and open questions

While I explored a research space encompassing interaction design, citizen science, and motivation through play, at multiple points in the research process I identified trade-offs and followed just one of possible paths. My detailed investigation and novel results regarding choropleth preference presupposed spectral representations of audio; interaction design research into the efficacy of other, less familiar, audio features for birdsong analysis remains to be done. Touchscreens provided most familiarity and corresponding reduction in need to explain interface interactions to users but did not provide the highest dimensionality; research into the representation dimensionality vs. familiarity trade-off remains. While my game designs were primarily focussed on providing familiar mechanics enabling immediate play, designs could be adapted to better support testing of both data-representation effectiveness for learning and game mechanics as sources of motivation, possibly without hindering flow. A/B testing of a single dynamic instead of multiple concurrent game-design variables would make it easier to tease out the significance of design choices; however, doing so would be at the expense of game experience. Interaction designs for developing user trust metrics through collaborative filtering competition provide viable mechanisms for increasing engagement while maintaining data-quality. Although my research has not focussed on such implementations, future work involves obtaining game ground-truth without expert annotation.

7.3 Future work

Development of my research output continues as I identify steps for implementing a transition from academic investigation to practical project development. In this final section I discuss software design revisions and augmentations, both satisfiers and motivators, for more diverse users. I discuss extensions to my work, including practical applications currently in development, and identify steps to be taken to overcome existing design issues and satisfy requirements for potential future projects. I present new collaborations and datasets in development for diverse future projects and ongoing modifications to my software implementation framework. I conclude with a discussion of future directions for expanding the scope of bioacoustic games for citizen science.

7.3.1 Ongoing development

While the collaboration with the Wild Watch has now reached its intended conclusion with the end of our three years of Heritage Lottery Fund (HLF) funding, several further collaborations are in development. My future work looks at the potential operationalisation of the outcomes of my game research for increasing engagement in public-facing citizen science projects. Here I will identify design updates relevant to iterative development of game artefacts for use outside the scope of UK primary education. I briefly introduce several collaborations in development, identifying the relevance of my prior work and the scope for further development across data-sets, target users, and game-mechanics. I conclude with a proposal for building an integrated platform encompassing my diverse interfaces suitable for assimilation into the broader bioacoustic-informatics ecosystem.

7.3.1.1 Design updates

Designing for significantly variable screen sizes for subsequent broader release of the games and adaptations for users with poor vision, including speech to text output for species identification in BirdMatch and SpectroPuzzle, are being implemented as satisfiers. As a motivator, play-order constraints will be relaxed to support additional user archetypes, such as the musical focus group who desired to engage with ChorusComposer from the outset. Species information supplied through Android toasts[2] was sometimes confounded with system messages which users were predisposed to ignore; revised screen overlays are being implemented as a motivator. Revised overlays can include links to additional target-species information, such as seasonal behaviours and habitats, requested by some users. While excess information might distract from flow within play of a single iteration, providing such links upon mastery is being incorporated for subsequent release. Future SpectroPuzzle design adaptations will offer 3 – 5 species rather than the current 2 – 4. Testing with 5 rows will be needed to identify whether solutions remain tractable for each target user group and visible on smaller screens. In subsequent public release, settings will allow personalised adjustment to data-representation characteristics, including choropleth mappings and spectrogram parameters, as well as tuning the scoring algorithm.

7.3.1.2 New collaborations & datasets

Several further collaborations for deployment of my software are in development, in the UK, and abroad. An upcoming HLF-funded project with the Cairngorms National Park[3] includes the requirement that I produce an adaptation of my games with utterances of

[2]A native android mechanism for overlaying data on-screen, also used by system messages, https://developer.android.com/guide/topics/ui/notifiers/toasts

[3]https://cairngorms.co.uk

the capercaillie, a rare and endangered endemic species. As there is need to preempt nest disruption, proposed games will primarily be used at the park's information centre to teach visitors the nesting and warning calls so they can recognise things to avoid. Further approaches to behaviour-modification games are being considered, including virtual reality interactions with the capercaillie as a means to reduce visitor attempts to find birds, while providing visitors with some semblance of contact and engagement.

Another proposed project is with the Norfolk Wildlife Trust at their Cley Marshes site[4], the oldest wildlife reserve in the region. A preliminary new dataset encompassing ducks has been produced for BirdMatch, with more detailed relevant spectrogram data pending for SpectroPuzzle and ChorusComposer. Collection of recordings for an additional wader dataset has also begun. For this collaboration I have proposed a combination of stationary tablets in the visitor centre and devices available for visitors to carry whilst exploring the site.

Another collaboration in progress is with the Rwanda Development Board (RDB), the government department which oversees development of their national parks. They have requested that suitable datasets be developed for games to train novice national park guides to recognise indigenous cryptic avian species by sound. I have proposed adding data-collection interfaces to the games for use by guides who will cover replicable transects as the RDB have requested that these be integrated with the eBird database.

7.3.1.3 Platform development

For now, my game designs primarily provide training on and elicit engagement with pre-recorded avian utterances; there is no requirement to geo-locate participants during interactions. Further development to the game platform includes incorporating user recording, library curation, and audio upload, as implemented and discussed in §4.3. For future distributed application release, output of ORM classes (see §A.I.1.1, §A.I.2, §A.I.3), which push data for each iteration played of each game to a database, can be toggled between on-device storage and transmission to networked centralised storage. As citizen science projects benefit from remotely updating target species lists to geo-located devices from a centralised database, this may also be implemented. I endeavour to combine data collected by professional land managers and avocational citizen scientists through my games into a meta-platform. This will include databases capable of integrating avian acoustic recordings and geographical information system (GIS) data with the underlying HSMs, allowing citizen monitoring of models which may influence subsequent policy. The Rwandan collaborators have proposed that their game output be incorporated into the eBird database, application programming interface (API) interoperability development ensues. Prior to upload and database inclusion, I am working on developing mechanics for

[4]https://www.norfolkwildlifetrust.org.uk/wildlife-in-norfolk/nature-reserves/reserves/cley-marshes

distributed verification of user-curated libraries.

7.3.2 Expanding research scope

This thesis has introduced a broad analysis of public engagement technology for bioacoustics and presented the particular path I have followed in the process of performing HCI research in developing citizen science games. Here I identify space for development beyond my current research, rather than within gaps in the process identified in §7.2.3.3. I propose useful directions for future contributions to ludified eco-informatics: the role of games and other technologies for building user trust, extensions to game-mechanics which maintain science-focussed game narratives, and the potential of new auditory analysis techniques for avian species recognition.

7.3.2.1 Citizen science community benefits to eco-informatics

Citizen science offers numerous benefits to eco-informatics datasets. Ongoing research investigates the potential for applying data-driven analysis techniques to collaboratively-filtering bioacoustic data for biodiversity analysis tasks, including species identification, population demographics, and behavioural shifts. As acoustic biodiversity indices describe broad geographic ranges, leveraging citizen science output can provide significantly more total and more precise acoustic data-points than do existent static sensor approaches. Building a community of citizen scientists can likewise provide the basis for incorporating collaborative filtering approaches to user trust metrics and subsequent data validation. Building distributed ledger technology for the databases which underpin proposed metrics provides the capacity for additional checks on user trust and mechanisms for weighting data-point validity prior to incorporation into biodiversity models. While data-driven approaches to biodiversity monitoring can provide necessary detail for models, engaging people with processes of biodiversity monitoring and assessment is necessary if human behaviour is to be modified.

7.3.2.2 Advancing bioacoustic game development

My future designs consider the potential for games to serve as tools for land-use behaviour modification. The comparative efficacy of population estimation from mobile recordings guides future designs, therefore building gamised transect game-mechanics remains future work. Drawing on prior work with image geo-caching, of flora, I am exploring the potential for motivating participants to explore nature through similar game mechanics where data-collection is open-ended play. Since avian fauna necessarily move, except in the case of nests, which we do not want game-motivators to encourage people to disrupt, reward mechanisms must be adapted. Augmented reality games provide a counter to virtual interfaces where participants may reap knowledge rewards without ancillary appreciation from exploring nature. In the case of the capercaillie, avoiding nest disruption is the target behaviour modification. Providing game interfaces which support intrinsic motivation for learning while extrinsically motivating engagement with nature will inspire a population of users to explore and contribute to bioacoustic citizen science.

Bibliography

[1] M. A. Acevedo, C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206–214, 2009.

[2] E. Aceves-Bueno, A. S. Adeleye, D. Bradley, W. T. Brandt, P. Callery, M. Feraud, K. L. Garner, R. Gentry, Y. Huang, Yuxiong, I. McCullough, I. Pearlman, S. A. Sutherland W. Wilkinson, Y. Yang, T. Zink, S. E. Anderson, and C. Tague. Citizen science as an approach for overcoming insufficient monitoring and inadequate stakeholder buy-in in adaptive management: criteria and evidence. *Ecosystems*, 18(3):493–506, 2015.

[3] I. Agranat. Method and apparatus for automatically identifying animal species from their vocalizations, November 18 2008. US Patent 7,454,334.

[4] L. Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

[5] L. Von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[6] T. M. Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013.

[7] A. Alabri and J. Hunter. Enhancing the quality and trust of citizen science data. In *e-Science (e-Science) and 2010 IEEE Sixth International Conference on*, pages 81–88. IEEE, 2010.

[8] T. P. Alloway, V. Bibile, and G. Lau. Computerized working memory training: Can it lead to gains in cognitive skills in students? *Computers in Human Behavior*, 29(3):632–638, 2013.

[9] T. Amano, J. D. L. Lamming, and W. J. Sutherland. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, 66(5):393–400, 2016.

[10] S. E. Anderson, A. S. Dave, and D. Margoliash. Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, 100(2):1209–1219, 1996.

[11] Y. Aono and K. Kazui. Phenological data series of cherry tree flowering in kyoto and japan and its application to reconstruction of springtime temperatures since the 9th century. *International Journal of Climatology*, 28(7):905–914, 2008.

[12] E. M. Avedon and B. Sutton-Smith. *The study of games*. John Wiley & Sons, 1971.

[13] D. E. Balmer, S. Gillings, B. Caffrey, R. L. Swann, I. S. Downie, and R. J. Fuller et al. *Bird Atlas 2007-11: the breeding and wintering birds of Britain and Ireland*. BTO Thetford, 2013.

[14] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K-H. Tauchert, and K-H. Frommolt. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31(12):1524–1534, 2010.

[15] L. Barrington, D. O'Malley, D. Turnbull, and G. Lanckriet. User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 7–10. ACM, 2009.

[16] R. F. Baumeister and M. R. Leary. The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin*, 117(3):497, 1995.

[17] I. Bogost. Gamification is bullshit. *The gameful world: Approaches and issues and applications*, pages 65–80, 2011.

[18] I. Bogost. Exploitationware. In *Rhetoric/composition/play through video games*, pages 139–147. Springer, 2013.

[19] R. Bonney, C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11):977–984, 2009.

[20] R. Bonney, T. B. Phillips, H. L. Ballard, and J. W. Enck. Can citizen science enhance public understanding of science? *Public Understanding of Science*, 25(1):2–16, 2016.

[21] R. Bonney, J. L. Shirk, T. B. Phillips, A. Wiggins, H. L. Ballard, A. J. Miller-Rushing, and J. K. Parrish. Next steps for citizen science. *Science*, 343(6178):1436–1437, 2014.

[22] D. N. Bonter and C. B. Cooper. Data validation in citizen science: a case study from project feederwatch. *Frontiers in Ecology and the Environment*, 10(6):305–307, 2012.

[23] F. Borges, M. Glemnitz, A. Schultz, Alfred, and U. Stachow. Assessing the habitat suitability of agricultural landscapes for characteristic breeding bird guilds using landscape metrics. *Environmental Monitoring and Assessment*, 189(4):166–187, 2017.

[24] A. Boulmaiz, D. Messadeg, N. Doghmane, and A. Taleb-Ahmed. Robust acoustic bird recognition for habitat monitoring with wireless sensor networks. *International Journal of Speech Technology*, 19(3):631–645, 2016.

[25] A. Bowser, D. Hansen, Y. He, C. Boston, M. Reid, L. Gunnell, and J. Preece. Using gamification to inspire new citizen science volunteers. In *Proceedings of the first international conference on gameful design and research and applications*, pages 18–25. ACM, 2013.

[26] A. Bowser, D. Hansen, J. Preece, Y. He, C. Boston, and J. Hammock. Gamifying citizen science: a study of two user groups. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 137–140. ACM, 2014.

[27] A. Bowser, A. Wiggins, L. Shanley, J. Preece, and S. Henderson. Sharing data while protecting privacy in citizen science. *interactions*, 21(1):70–73, 2014.

[28] R. Bragg, C. Wood, J. Barton, and J. Pretty. Measuring connection to nature in children aged 8-12: A robust methodology for the rspb. *Epublication https://www. rspb. org. uk/Images/methodology-report_tcm9-354606. pdf (accessed December 2015)*, 2013.

[29] T. S. Brandes. Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1):S163–S173, 2008.

[30] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012.

[31] H. K. Burgess, L. B. DeBey, H. E. Froehlich, N. Schmidt, E. J. Theobald, A. K. Ettinger, J. HilleRisLambers, J. Tewksbury, and J. K. Parrish. The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation*, 208:113–120, 2017.

[32] J. A. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. 2006.

[33] S. H. M. Butchart, M. Walpole, B. Collen, A. van Strien, J. P. W. Scharlemann, R. E. A. Almond, J. E. M. Baillie, B. Bomhard, C. Brown, J. Bruno, and John et al. Global biodiversity: indicators of recent declines. *Science*, 328(5982):1164–1168, 2010.

[34] J. L. Cappadonna, M. Brereton, D. M. Watson, and P. Roe. Calls from the wild: Engaging citizen scientist with animal sounds. In *Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems*, pages 157–160. ACM, 2016.

[35] R. Caruana, M. Elhawary, A. Munson, M. Riedewald, D. Sorokina, D. Fink, W. M. Hochachka, and S. Kelling. Mining citizen science data to predict prevalence of wild bird species. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 909–915. ACM, 2006.

[36] C. L. Catlin-Groves. The citizen science landscape: from volunteers to citizen sensors and beyond. *International Journal of Zoology*, 2012, 2012.

[37] H. R. Chapman and N. Kirby-Turner. Visual/verbal analogue scales: examples of brief assessment methods to aid management of child and adult patients in clinical practice. *British Dental Journal*, 193(8):447, 2002.

[38] Z. Chen and R. C. Maher. Semi-automatic classification of bird vocalizations using spectral peak tracks. *The Journal of the Acoustical Society of America*, 120(5):2974–2984, 2006.

[39] C. Cherry. The cocktail party problem. *Journal of the Acoustical Society of America*, 25:975–979, 1953.

[40] C-H. Chou and P-H. Liu. Bird species recognition by wavelet transformation of a section of birdsong. In *Ubiquitous and Autonomic and Trusted Computing and 2009. UIC-ATC09. Symposia and Workshops on*, pages 189–193. IEEE, 2009.

[41] C-H. Chou, P-H. Liu, and B. Cai. On the studies of syllable segmentation and improving mfccs for automatic birdsong recognition. In *Asia-Pacific Services Computing Conference and 2008. APSCC08. IEEE*, pages 745–750. IEEE, 2008.

[42] D. Christin, A. Reinhardt, S. S. Kanhere, and M. Hollick. A survey on privacy in mobile participatory sensing applications. *Journal of Systems and Software*, 84(11):1928–1946, 2011.

[43] J. P. Cohn. Citizen science: Can volunteers do real research? *AIBS Bulletin*, 58(3):192–197, 2008.

[44] D. Coleman, Y. Georgiadou, and J. Labonte. Volunteered geographic information: The nature and motivation of produsers. *IJSDIR*, 4(1):332–358, 2009.

[45] C. C. Conrad and K. G. Hilchey. A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, 176(1):273–291, 2011.

[46] C. B. Cooper, J. Shirk, and B. Zuckerberg. The invisible prevalence of citizen science in global research: migratory birds and climate change. *PloS one*, 9(9):e106508, 2014.

[47] M. Cottman-Fields, M. Brereton, and P. Roe. Virtual birding: extending an environmental pastime into the virtual world for citizen science. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 2029–2032. ACM, 2013.

[48] M. Cottman-Fields, M. Brereton, J. Wimmer, and P. Roe. Collaborative extension of biodiversity monitoring protocols in the bird watching community. In *Proceedings of the 13th Participatory Design Conference: Short Papers and Industry Cases and Workshop Descriptions and Doctoral Consortium papers and Keynote abstracts-Volume 2*, pages 111–114. ACM, 2014.

[49] C. Crawford. The art of computer game design. 1984.

[50] N. Cross. Design research: A disciplined conversation. *Design issues*, 15(2):5–10, 1999.

[51] K. Crowston and I. Fagnot. The motivational arc of massive virtual collaboration. In *Proceedings of the IFIP WG*, volume 9.5, 2008.

[52] K. Crowston and N. R. Prestopnik. Motivation and data quality in a citizen science game: A design science evaluation. In *System Sciences (HICSS) and 2013 46th Hawaii International Conference on*, pages 450–459. IEEE, 2013.

[53] M. Csikszentmihalyi. Flow: The psychology of optimal performance. *NY: Cambridge UniversityPress*, 1990.

[54] V. Curtis. Motivation to participate in an online citizen science game: A study of foldit. *Science Communication*, 37(6):723–746, 2015.

[55] T. Damoulas, S. Henry, A. Farnsworth, M. Lanzone, and C. Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *Machine Learning and Applications (ICMLA) and 2010 Ninth International Conference on*, pages 424–429. IEEE, 2010.

[56] F. Danielsen, N. D. Burgess, M. Funder, T. Blomley, J. Brashares, A. Akida, A. Jensen, M. Mendoza, G. Stuart-Hill, M. K. Poulsen, H. Ramadhani, M. K. Sam, and E. Topp-Jorgensen. Taking stock of nature in species-rich but economically poor areas: an emerging discipline of locally based monitoring. *Taking Stock of Nature. Participatory Biodiversity Assessment for Policy and Planning and Practice*, pages 88–112, 2010.

[57] F. Danielsen, N. D. Burgess, P. M. Jensen, and K. Pirhofer-Walzl. Environmental monitoring: the scale and speed of implementation varies according to the degree of peoples involvement. *Journal of Applied Ecology*, 47(6):1166–1186, 2010.

[58] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. User acceptance of computer technology: a comparison of two theoretical models. *Management science*, 35(8):982–1003, 1989.

[59] E. L. Deci and R. M. Ryan. Cognitive evaluation theory. In *Intrinsic motivation and self-determination in human behavior*, pages 43–85. Springer, 1985.

[60] E. L. Deci and R. M. Ryan. Self-determination theory. *Handbook of theories of social psychology*, 1:416–433, 2011.

[61] E. L. Deci and M. Vansteenkiste. Self-determination theory and basic need satisfaction: Understanding human development in positive psychology. *Ricerche di psicologia*, 2004.

[62] M. Depraetere, S. Pavoine, F. Jiguet, A. Gasc, S. Duvail, and J. Sueur. Monitoring animal diversity using acoustic indices: implementation in a temperate woodland. *Ecological Indicators*, 13(1):46–54, 2012.

[63] S. Deterding. Situated motivational affordances of game elements: A conceptual model. In *Gamification: Using game design elements in non-gaming contexts and a workshop at CHI*, 2011.

[64] S. Deterding. Coding conduct: Games and play and human conduct between technical artifacts and social framing. *Hans Bredow Institute for Media Research and November*, 5, 2012.

[65] S. Deterding. Gamification: designing for motivation. *interactions*, 19(4):14–17, 2012.

[66] S. Deterding, S. L. Björk, L. E. Nacke, D. Dixon, and E. Lawley. Designing gamification: creating gameful and playful experiences. In *CHI13 Extended Abstracts on Human Factors in Computing Systems*, pages 3263–3266. ACM, 2013.

[67] E. D'Hondt, M. Stevens, and A. Jacobs. Participatory noise mapping works! an evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. *Pervasive and Mobile Computing*, 9(5):681–694, 2013.

[68] J. L. Dickinson, B. Zuckerberg, and D. N. Bonter. Citizen science as an ecological research tool: challenges and benefits. *Annual review of ecology and evolution and systematics*, 41:149–172, 2010.

[69] T. G. Dietterich. Machine learning in ecosystem informatics and sustainability. In *IJCAI*, pages 8–13, 2009.

[70] B. DiSalvo and B. B. Morrison. A critique of g̈amificationïn khan academy. 2013.

[71] P. Divenyi. *Speech separation by humans and machines*. Springer Science & Business Media, 2004.

[72] A. Dolphin. Defining sound toys: play as composition. In K. Collins, B. Kapralos, and H. Tessler, editors, *The Oxford Handbook of Interactive Audio*. Oxford University Press, Oxford, 2014.

[73] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880. ACM, 2012.

[74] B. Elliott. An interactive system for identifying and extracting audio clips based on a spectrogram, 2014.

[75] A. Eveleigh, C. Jennett, A. Blandford, P. Brohan, and A. L. Cox. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2985–2994. ACM, 2014.

[76] F. Eyben. *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[77] S. Fagerlund and A. Härmä. Parametrization of inharmonic bird sounds for automatic recognition. In *13th European Signal Processing Conference (EUSIPCO 2005)*, pages 4–8, 2005.

[78] D. Fallman. The interaction design research triangle of design practice and design studies and design exploration. *Design Issues*, 24(3):4–18, 2008.

[79] E. H. Fegraus, S. Andelman, M. B. Jones, and M. Schildhauer. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (eml) and principles for metadata creation. *Bulletin of the Ecological Society of America*, 86(3):158–168, 2005.

[80] M. Fisher-Phelps, D. Schwilk, and T. Kingston. Mobile acoustic transects detect more bat activity than stationary acoustic point counts in a semi-arid and agricultural landscape. *Journal of Arid Environments*, 136:38–44, 2017.

[81] S. Follmer, D. Leithinger, A. Olwal, A. Hogge, and H. Ishii. inform: dynamic physical affordances and constraints through shape and object actuation. In *Uist*, volume 13, pages 417–426, 2013.

[82] R. A. Fuller, J. Tratalos, and K. J. Gaston. How many birds are there in a city of half a million people? *Diversity and Distributions*, 15(2):328–337, 2009.

[83] S. H. Gage, E. P. Kasten, and W. Joo. Development and application of environmental acoustics for monitoring bird pathways in wind resource areas. (MSU-REAL-09-1), 2009.

[84] S. Gaonkar, J. Li, R. R. Choudhury, L. Cox, and A. Schmidt. Micro-blog: sharing and querying content through mobile phones and social participation. In *Proceedings of the 6th international conference on Mobile systems and applications and services*, pages 174–186. ACM, 2008.

[85] S. D. Gardner. *Evaluation of the ColorBrewer color schemes for accommodation of map readers with impaired color vision.* PhD thesis, Pennsylvania State University, 2005.

[86] J. Geldmann, J. Heilmann-Clausen, T. E. Holm, I. Levinsky, B. Markussen, K. Olsen, C. Rahbek, and A. P. Tøttrup. What determines spatial bias in citizen science? exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11):1139–1149, 2016.

[87] T. Q. Gentner. Temporal scales of auditory objects underlying birdsong vocal recognition. *The Journal of the Acoustical Society of America*, 124(2):1350–1359, 2008.

[88] H. Goëau, H. Glotin, W-P. Vellinga, R. Planqué, A. Rauber, and A. Joly. Lifeclef bird identification task 2014. In *CLEF: Conference and Labs of the Evaluation forum*, 2014.

[89] C. P. Gomes. Computational sustainability: Computational methods for a sustainable environment and economy and society. *The Bridge*, 39(4):5–13, 2009.

[90] A. Gonsamo and P. D'Odorico. Citizen science: best practices to remove observer bias in trend analysis. *International journal of biometeorology*, 58(10):2159–2163, 2014.

[91] M. Gontier, B. Balfors, and U. Mörtberg. Biodiversity in environmental assessment - current practice and tools for prediction. *Environmental Impact Assessment Review*, 26(3):268–286, 2006.

[92] M. F. Goodchild. Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0. *International Journal of Spatial Data Infrastructures Research*, pages 24–32, 2007.

[93] M. Graciarena, M. Delplanche, E. Shriberg, A. Stolcke, and L. Ferrer. Acoustic front-end optimization for bird species recognition. In *Acoustics Speech and Signal Processing (ICASSP) and 2010 IEEE International Conference on*, pages 293–296. IEEE, 2010.

[94] A. Greenhill, K. Holmes, C. Lintott, B. Simmons, K. Masters, J. Cox, and G. Graham. Playing with science: Gamised aspects of gamification found on the online citizen science project-zooniverse. In *GAMEON2014*. EUROSIS, 2014.

[95] M. Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B: Planning and design*, 37(4):682–703, 2010.

[96] M. Haklay. Classification of citizen science activities. *Po Ve Sham*, 2011.

[97] M. Haklay. Citizen science and volunteered geographic information: Overview and typology of participation. In *Crowdsourcing geographic knowledge*, pages 105–122. Springer, 2013.

[98] M. Haklay, S. Basiouka, V. Antoniou, and A. Ather. How many volunteers does it take to map an area well? the validity of linus' law to volunteered geographic information. *The Cartographic Journal*, 47(4):315–322, 2010.

[99] K. Han, E. A. Graham, D. Vassallo, and D. Estrin. Enhancing motivation in a mobile participatory sensing project through gaming. In *Privacy and security and risk and trust (passat) and 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 1443–1448. IEEE, 2011.

[100] A. Harma and P. Somervuo. Classification of the harmonic structure in bird vocalization. In *Acoustics and Speech and Signal Processing and 2004. Proceedings.(ICASSP04). IEEE International Conference on*, volume 5, pages V–701. IEEE, 2004.

[101] A. Hirzel and A. Guisan. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological modelling*, 157(2):331–341, 2002.

[102] C. S. Holling. The functional response of predators to prey density and its role in mimicry and population regulation. *The Memoirs of the Entomological Society of Canada*, 97(S45):5–60, 1965.

[103] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[104] P-Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics, 2009.

[105] J. Huizinga and R. F. C. Hull. *Homo Ludens. A Study of the Play-element in Culture.[Translated by RFC Hull.]*. Routledge & Kegan Paul, 1949.

[106] I. Iacovides, C. Jennett, C. Cornish-Trestrail, and A. L. Cox. Do games attract or sustain engagement in citizen science?: a study of volunteer motivations. In *CHI13 Extended Abstracts on Human Factors in Computing Systems*, pages 1101–1106. ACM, 2013.

[107] D. Borland R. M. Taylor II. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, 27(2):14–17, 2007.

[108] A. Irwin. *Citizen science: A study of people and expertise and sustainable development.* Psychology Press, 1995.

[109] Y. Iwasaki and R. C. Mannell. Situational and personality influences on intrinsically motivated leisure behavior: Interaction effects and cognitive processes. *Leisure Sciences*, 21(4):287–306, 1999.

[110] C. Jennett, L. Kloetzer, D. Schneider, I. Iacovides, A. Cox, M. Gold, B. Fuchs, A. Eveleigh, K. Methieu, Z. Ajani, and Y. Talsi. Motivations and learning and creativity in online citizen science. *Journal of Science Communication*, 15(3), 2016.

[111] B. E. John and D. E. Kieras. The goms family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(4):320–351, 1996.

[112] A. Joly, H. Müller, H. Goëau, H. Glotin, C. Spampinato, A. Rauber, P. Bonnet, W-P. Vellinga, and B. Fisher. Lifeclef: Multimedia life species identification.

[113] C-F. Juang and T-M. Chen. Birdsong recognition using prediction-based recurrent neural fuzzy networks. *Neurocomputing*, 71(1):121–130, 2007.

[114] J. Juul. The game and the player and the world: Looking for a heart of gameness. *PLURAIS-Revista Multidisciplinar*, 1(2), 2010.

[115] E. Kanjo, S. Benford, M. Paxton, A. Chamberlain, D. S. Fraser, D. Woodgate, D. Crellin, and A. Woolard. Mobgeosen: facilitating personal geosensor data collection and visualization using mobile phones. *Personal and Ubiquitous Computing*, 12(8):599–607, 2008.

[116] A. B. Kaufman. Let's talk about citizen science: What doesn't work. *Animal Behavior and Cognition*, 1(4):446–451, 2014.

[117] S. Kelling, D. Fink, F. A. La Sorte, A. Johnston, N. E. Bruns, and W. M. Hochachka. Taking a 'big data' approach to data quality in a citizen science project. *Ambio*, 44(4):601–611, 2015.

[118] S. Kelling, W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, Grant, and G. Hooker. Data-intensive science: a new paradigm for biodiversity studies. *BioScience*, 59(7):613–620, 2009.

[119] C. J. Keylock. Simpson diversity and the shannon–wiener index as special cases of a generalized entropy. *Oikos*, 109(1):203–207, 2005.

[120] S. Kim, J. Mankoff, and E. Paulos. Sensr: evaluating a flexible framework for authoring mobile data-collection tools for citizen science. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1453–1462. ACM, 2013.

[121] B. T. Klingbeil and M. R. Willig. Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols. *PeerJ*, 3:e973, 2015.

[122] H. Kobori, J. L. Dickinson, I. Washitani, R. Sakurai, T. Amano, N. Komatsu, W. Kitamura, K. Koyama S. Takagawa, T. Ogawara, and A. J. Miller-Rushing. Citizen science: a new approach to advance ecology and education and conservation. *Ecological research*, 31(1):1–19, 2016.

[123] J. A. Kogan and D. Margoliash. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study. *The Journal of the Acoustical Society of America*, 103(4):2185–2196, 1998.

[124] M. E. Krasny and R. Bonney. Environmental education through citizen science and participatory action research. *Environmental education and advocacy: Changing perspectives of ecology and education*, pages 292–319, 2005.

[125] A. Krause, E. Horvitz, A. Kansal, F. Zhao, and Feng. Toward community sensing. In *Proceedings of the 7th international conference on Information processing in sensor networks*, pages 481–492. IEEE Computer Society, 2008.

[126] R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl. *Building successful online communities: Evidence-based social design.* MIT Press, 2012.

[127] G. M. Kuhn. Description of a color spectrogram. *The Journal of the Acoustical Society of America*, 76(3):682–685, 1984.

[128] C. Kullenberg and D. Kasperowski. What is citizen science?–a scientometric meta-analysis. *PloS one*, 11(1):e0147152, 2016.

[129] C. Kwan, G. Mei, X. Zhao, Z. Ren, R. Xu, V. Stanford, C. Rochet, J. Aube, and K. C. Ho. Bird classification algorithms: Theory and experimental results. In *Acoustics and Speech and Signal Processing and 2004. Proceedings.(ICASSP04). IEEE International Conference on*, volume 5, pages V–289. IEEE, 2004.

[130] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine and IEEE*, 48(9):140–150, 2010.

[131] N. D. Lane, Y. Xu, H. Lu, S. Hu, T. Choudhury, A. T. Campbell, and F. Zhao. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 355–364. ACM, 2011.

[132] E. Law and L. Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206. ACM, 2009.

[133] E. L. M. Law, L. Von Ahn, R. B. Dannenberg, and M. Crawford. Tagatune: A game for music and sound annotation. In *ISMIR*, volume 3, page 2, 2007.

[134] E. C. Leach, C. J. Burwell, L. A. Ashton, D. N. Jones, and R. L. Kitching. Comparison of point counts and automated acoustic monitoring: detecting birds in a rainforest biodiversity survey. *Emu*, 116(3):305–309, 2016.

[135] C. A. Lepczyk. Integrating published data and citizen science to describe bird diversity across a landscape. *Journal of Applied Ecology*, 42(4):672–677, 2005.

[136] J. C. R. Licklider. A duplex theory of pitch perception. *The Journal of the Acoustical Society of America*, 23(1):147–147, 1951.

[137] A. Light and P. J. Bartlein. The end of the rainbow? color schemes for improved data graphics. *Eos and Transactions American Geophysical Union*, 85(40):385–391, 2004.

[138] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[139] M. T. Lopes, L. L. Gioppo, T. T. Higushi, C. A. A. Kaestner, C. N. Silla, and A. L. Koerich. Automatic bird species identification for large number of species. In *Multimedia (ISM) and 2011 IEEE International Symposium on*, pages 117–122. IEEE, 2011.

[140] M. Luczak-Roesch, R. Tinati, E. Simperl, M. Van Kleek, N. Shadbolt, and R. J. Simpson. Why won't aliens talk to us? content and community dynamics in online citizen science. In *ICWSM*, 2014.

[141] R. Lukyanenko, J. Parsons, and Y. F. Wiersma. Emerging problems of data quality in citizen science. *Conservation Biology*, 30(3):447–449, 2016.

[142] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological informatics*, 2(3):279–296, 2007.

[143] B. Magerko, J. Freeman, T. McKlin, S. McCoid, T. Jenkins, and E. Livingston. Tackling engagement in computing with computational music remixing. In *Proceeding of the 44th ACM technical symposium on Computer science education*, pages 657–662. ACM, 2013.

[144] T. W. Malone. What makes things fun to learn? heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems*, pages 162–169. ACM, 1980.

[145] M. I. Mandel and D. P. W. Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008.

[146] I. G. Martí, L. E. Rodríguez, M. Benedito, S. Trilles, A. Beltrán, L. Díaz, and J. Huerta. Mobile application for noise pollution monitoring through gamification techniques. In *Entertainment Computing-ICEC 2012*, pages 562–571. Springer, 2012.

[147] I. C. Martínez, S. Malbrán, and F. Shifres. The role of repetition in aural identification of harmonic sequences. *Bulletin of the Council for Research in Music Education*, pages 93–97, 1999.

[148] A. H. Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

[149] A. H. Maslow and K. J. Lewis. Maslow's hierarchy of needs. *Salenger Incorporated*, 14, 1987.

[150] R. Mason, P. Roe, M. Towsey, J. Zhang, J. Gibson, and S. Gage. Towards an acoustic environmental observatory. In *eScience and 2008. eScience08. IEEE Fourth International Conference on*, pages 135–142. IEEE, 2008.

[151] P. Massa, K. Souren, M. Salvetti, and D. Tomasoni. Trustlet and open research on trust metrics. *Scalable Computing: Practice and Experience*, 9(4), 2008.

[152] A. L. McIlraith and H. C. Card. Birdsong recognition using backpropagation and multivariate statistics. *Signal Processing and IEEE Transactions on*, 45(11):2740–2748, 1997.

[153] E. D. Mekler, F. Brühlmann, A. N. Tuch, and K. Opwis. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, 2015.

[154] W. K. Michener. Meta-information concepts for ecological data management. *Ecological informatics*, 1(1):3–7, 2006.

[155] W. K. Michener and M. B. Jones. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolution*, 27(2):85–93, 2012.

[156] S. Moran, N. Pantidi, T. Rodden, A. Chamberlain, C. Griffiths, D. Zilli, G. Merrett, and A. Rogers. Listening to the forest and its curators: lessons learnt from a bioacoustic smartphone application deployment. In *Proceedings of the 32nd annual*

ACM conference on Human factors in computing systems, pages 2387–2396. ACM, 2014.

[157] K. Moreland. Diverging color maps for scientific visualization. In *Advances in Visual Computing*, pages 92–103. Springer, 2009.

[158] R. J. Murphy. *Facilitating Citizen Science Through Gamification*. PhD thesis, PhD thesis and Memorial University of Newfoundland, 2015.

[159] J. Nakamura and M. Csikszentmihalyi. The concept of flow. In *Flow and the foundations of positive psychology*, pages 239–263. Springer, 2014.

[160] D. A. Nelson. The importance of invariant and distinctive features in species recognition of bird song. *Condor*, pages 120–130, 1989.

[161] S. Ness and G. Tzanetakis. Human and machine annotation in the orchive and a large scale bioacoustic archive. In *Signal and Information Processing (GlobalSIP) and 2014 IEEE Global Conference on*, pages 1136–1140. IEEE, 2014.

[162] C. Newman, C. D. Buesching, and D. W. Macdonald. Validating mammal monitoring methods and assessing the performance of volunteers in wildlife conservation - 'sed quis custodiet ipsos custodies?'. *Biological Conservation*, 113(2):189–197, 2003.

[163] G. Newman, A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston. The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6):298–304, 2012.

[164] S. Nicholson. A user-centered theoretical framework for meaningful gamification. *Games+ Learning+ Society*, 8(1):223–230, 2012.

[165] D. A. Norman. Affordance and conventions and design. *interactions*, 6(3):38–43, 1999.

[166] O. Nov, O. Arazy, and D. Anderson. Dusting for science: motivation and participation of digital citizen science volunteers. In *Proceedings of the 2011 iConference*, pages 68–74. ACM, 2011.

[167] O. Nov, O. Arazy, and D. Anderson. Scientists@ home: what drives the quantity and quality of online citizen science participation? *PloS one*, 9(4):e90375, 2014.

[168] M. Osterloh and B. S. Frey. Motivation and knowledge transfer and organizational forms. *Organization science*, 11(5):538–550, 2000.

[169] N. Pantidi, S. Moran, K. Bachour, T. Rodden, D. Zilli, G. V. Merrett, and A. Rogers. Field testing a rare species bioacoustic smartphone application: challenges and future considerations. 2014.

[170] B. Paras. Game and motivation and effective learning: An integrated model for educational game design. In *Proceedings of DiGRA 2005: Changing Views - Worlds in Play*, 2005.

[171] C. Penone, I. Le Viol, V. Pellissier, J. Julien, Y. Bas, Yves, and C. Kerbiriou. Use of large-scale acoustic monitoring to assess anthropogenic pressures on orthoptera communities. *Conservation Biology*, 27(5):979–987, 2013.

[172] J. Porter, P. Arzberger, H-W. Braun, P. Bryant, S. Gage, T. Hansen, P. Hanson, C-C. Lin, F-P. Lin, T. Kratz, W. Michener, S. Shapiro, and T. Williams. Wireless sensor networks for ecology. *BioScience*, 55(7):561–572, 2005.

[173] J. Preece. How two billion smartphone users can save species! *interactions*, 24(2):26–33, 2017.

[174] J. Preece and A. Bowser. What hci can do for citizen science. In *CHI14 Extended Abstracts on Human Factors in Computing Systems*, pages 1059–1060. ACM, 2014.

[175] N. Prestopnik, K. Crowston, and J. Wang. Gamers and citizen scientists and data: Exploring participant contributions in two games with a purpose. *Computers in Human Behavior*, 68:254–268, 2017.

[176] N. R. Prestopnik. *Design science in human-computer interaction A model and three examples*. PhD thesis, Syracuse University, 2013.

[177] N. R. Prestopnik and K. Crowston. Gaming for (citizen) science: exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system. In *e-Science Workshops (eScienceW) and 2011 IEEE Seventh International Conference on*, pages 28–33. IEEE, 2011.

[178] N. R. Prestopnik and K. Crowston. Citizen science system assemblages: understanding the technologies that support crowdsourced science. In *Proceedings of the 2012 iConference*, pages 168–176. ACM, 2012.

[179] N. R. Prestopnik and J. Tang. Points and stories and worlds and diegesis: Comparing player experiences in two citizen science games. *Computers in Human Behavior*, 52:492–506, 2015.

[180] A. K. Przybylski, C. S. Rigby, and R. M. Ryan. A motivational model of video game engagement. *Review of general psychology*, 14(2):154, 2010.

[181] H. C. Purchase, N. Andrienko, T. J. Jankun-Kelly, and M. Ward. Theoretical foundations of information visualization. In *Information Visualization*, pages 46–64. Springer, 2008.

[182] R. Raisamo, S. Patomäki, M. Hasu, and V. Pasto. Design and evaluation of a tactile memory game for visually impaired children. *Interacting with Computers*, 19(2):196–205, 2006.

[183] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu. Ear-phone: an end-to-end participatory urban noise mapping system. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 105–116. ACM, 2010.

[184] J. Reed, W. Rodriguez, and A. Rickhoff. A framework for defining and describing key design features of virtual citizen science projects. In *Proceedings of the 2012 iConference*, pages 623–625. ACM, 2012.

[185] D. B. Resnik, K. C. Elliott, and A. K. Miller. A framework for addressing ethical issues in citizen science. *Environmental Science & Policy*, 54:475–481, 2015.

[186] A. D. Rice and J. W. Lartigue. Touch-level model (tlm): evolving klm-goms for touchscreen and mobile devices. In *Proceedings of the 2014 ACM Southeast regional conference*, page 53. ACM, 2014.

[187] K. Riede. Acoustic monitoring of orthoptera and its potential for conservation. *Journal of Insect Conservation*, 2(3-4):217–223, 1998.

[188] M. A. Roch, H. Batchelor, S. Baumann-Pickering, C. L. Berchok, D. Cholewiak, E. Fujioka, E. C. Garland S. Herbert, J. A. Hildebrand E. M. Oleson, and et al. Management of acoustic metadata for bioacoustics. *Ecological Informatics*, 31:122–136, 2016.

[189] Y. Rogers. New theoretical approaches for human-computer interaction. *Annual review of information science and technology*, 38(1):87–143, 2004.

[190] D. Rotman, J. Hammock, J. Preece, D. Hansen, C. Boston, A. Bowser, and Y. He. Motivations affecting initial and long-term participation in citizen science projects in three countries. *iConference 2014 Proceedings*, 2014.

[191] D. Rotman, J. Preece, J. Hammock, K. Procita, D. Hansen, C. Parr, D. Lewis, and D. Jacobs. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 217–226. ACM, 2012.

[192] Z. Ruan, E. C. Ngai, and J. Liu. Wireless sensor network deployment in mobile phones assisted environment. In *Quality of Service (IWQoS) and 2010 18th International Workshop on*, pages 1–9. IEEE, 2010.

[193] R. M. Ryan and E. L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.

[194] R. M. Ryan, J. Kuhl, and E. L. Deci. Nature and autonomy: An organizational view of social and neurobiological aspects of self-regulation in behavior and development. *Development and psychopathology*, 9(4):701–728, 1997.

[195] R. M. Ryan, V. Mims, and R. Koestner. Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of personality and Social Psychology*, 45(4):736, 1983.

[196] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60, 2005.

[197] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling. Towards the automatic classification of avian flight calls for bioacoustic monitoring. *PloS one*, 11(11):e0166866, 2016.

[198] K. Salen and E. Zimmerman. *Rules of play: Game design fundamentals*. MIT press, 2004.

[199] K. Sato. Perspectives of design research: In design: Collective views for forming the foundation of design research. *Visible language*, 38(2):218, 2004.

[200] K. Seaborn and D. I. Fels. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74:14–31, 2015.

[201] A. Selin, J. Turunen, and J. T. Tanttu. Wavelets in recognition of bird sounds. *EURASIP Journal on Applied Signal Processing*, 2007(1):141–141, 2007.

[202] J. L. Sherry. Flow and media enjoyment. *Communication theory*, 14(4):328–347, 2004.

[203] J. L. Shirk, H. L. Ballard, C. C. Wilderman, T. Phillips, A. Wiggins, R. Jordan, E. McCallie, M. Minarchek, B. V. Lewenstein, M. E. Krasny, and R. Bonney. Public participation in scientific research: a framework for deliberate design. *Ecology and Society*, 17(2), 2012.

[204] K. Siedenburg. An exploration of real-time visualizations of musical timbre. In *Welcome to the 3 rd International Workshop on Learning Semantics of Audio Signals*, page 17, 2009.

[205] J. Silvertown, C. D. Buesching, S. K. Jacobson, and T. Rebelo. Citizen science and nature conservation. *Key topics in conservation biology*, 2:127–142, 2013.

[206] B. F. Skinner. Operant behavior. *American Psychologist*, 18(8):503, 1963.

[207] M. Slaney. Web-scale multimedia analysis: Does content matter? *IEEE MultiMedia*, 18(2):12–15, 2011.

[208] J. O. Smith. *Mathematics of the discrete Fourier transform (DFT): with audio applications*. Julius Smith, 2007.

[209] P. Somervuo, A. Härmä, and S. Fagerlund. Parametric representations of bird sounds for automatic species recognition. *Audio and Speech and Language Processing and IEEE Transactions on*, 14(6):2252–2263, 2006.

[210] J. Springer, Z. Duan, and B. Pardo. Approaches to multiple concurrent species bird song recognition. In *The 2nd International Workshop on Machine Listening in Multisource Environments and ICASSP*, 2013.

[211] J. Sprinks, J. Wardlaw, R. Houghton, S. Bamford, and J. Morley. Task workflow design and its impact on performance and volunteers' subjective preference in virtual citizen science. *International Journal of Human-Computer Studies*, 104:50–63, 2017.

[212] M. Stange, A. Barry, J. Smyth, and K. Olson. Effects of smiley face scales on visual processing of satisfaction questions in web surveys. *Social Science Computer Review*, page 0894439316674166, 2016.

[213] M. Stevens, M. Vitos, J. Altenbuchner, G. Conquest, J. Lewis, and M. Haklay. Introducing sapelli: A mobile data collection platform for non-literate users. In *Proceedings of the 4th Annual Symposium on Computing for Development*, page 17. ACM, 2013.

[214] E. Stolterman. The nature of design practice and implications for interaction design research. *International Journal of Design*, 2(1), 2008.

[215] D. Stowell, S. Musevic, J. Bonada, and M. D. Plumbley. Improved multiple birdsong tracking with distribution derivative method and markov renewal process clustering. In *Acoustics and Speech and Signal Processing (ICASSP) and 2013 IEEE International Conference on*, pages 468–472. IEEE, 2013.

[216] D. Stowell and M. D. Plumbley. Acoustic detection of multiple birds in environmental audio by matching pursuit.

[217] D. Stowell and M. D. Plumbley. Birdsong and c4dm: A survey of uk birdsong and machine recognition for music researchers. Technical Report C4DM-TR-09-12, Centre for Digital Music and Queen Mary University of London, 2010.

[218] D. Stowell and M. D. Plumbley. Framewise heterodyne chirp analysis of birdsong. In *Signal Processing Conference (EUSIPCO) and 2012 Proceedings of the 20th European*, pages 2694–2698. IEEE, 2012.

[219] D. Stowell and M. D. Plumbley. Large-scale analysis of frequency modulation in birdsong databases. *arXiv preprint arXiv:1311.4764*, 2013.

[220] B. L. Sullivan, J. L. Aycrigg, J. H. Barry, R. E. Bonney, N. Bruns, C. B. Cooper, T. Damoulas, A. D'hondt, T. Dietterich, A. Farnsworth, D. Fink, J. W. Fitzpatrick, T. Fredericks, J. Gerbracht, C. Gomes, W. M. Hochachka, M. J. Iliff, C. Lagoze, F. A. La Sorte, M. Merrifield, W. Morris, T. B. Phillips, M. Reynolds, A. D. Rodewald, K. V. Rosenberg, N. M. Trautmann, A. Wiggins, D. W. Winkler, W-K. Wong, C. L. Wooda, J. Yu, and S. Kelling. The ebird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014.

[221] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

[222] P. Sweetser and P. Wyeth. Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3), 2005.

[223] A. Taylor, Andrew, and G. Grigg. Development of an autonomous (acoustic) biodiversity monitoring system. *Development*, 2011.

[224] A. M. R. Terry and P. K. McGregor. Census and monitoring based on individually identifiable vocalizations: the role of neural networks. In *Animal Conservation forum*, volume 5.2, pages 103–111. Cambridge University Press, 2002.

[225] E. J. Theobald, A. K. Ettinger, H. K. Burgess, L. B. DeBey, N. R. Schmidt, H. E. Froehlich, C. Wagner, J. HilleRisLambers, J. Tewksbury, M. A. Harsch, and J.K. Parrish. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, 181:236–244, 2015.

[226] W. Thuiller and T. Münkemüller. Habitat suitability modeling. *Effects of Climate Change on Birds. Oxford University Press and Oxford*, pages 77–85, 2010.

[227] A. T. Tierney, F. A. Russo, and A. D. Patel. The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences*, 108(37):15510–15515, 2011.

[228] R. Tinati, M. Van Kleek, E. Simperl, M. Luczak-Rösch, R. Simpson, and N. Shadbolt. Designing for citizen data analysis: a cross-sectional case study of a multi-domain citizen science platform. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4069–4078. ACM, 2015.

[229] R. Tinati, M. Luczak-Roesch, E. Simperl, and W. Hall. An investigation of player motivations in eyewire and a gamified citizen science project. *Computers in Human Behavior*, 73:527–540, 2017.

[230] A. Truskinger, M. Towsey, and P. Roe. Decision support for the efficient annotation of bioacoustic events. *Ecological Informatics*, 25:14–21, 2015.

[231] E. R. Tufte. The visual display of quantitative information, 2001.

[232] A. I. T. Tulloch, H. P. Possingham, L. N. Joseph, J. Szabo, and T. G. Martin. Realising the full potential of citizen science monitoring programs. *Biological Conservation*, 165:128–138, 2013.

[233] G. Tzanetakis and P. Cook. Marsyas: A framework for audio analysis. *Organised sound*, 4(3):169–175, 2000.

[234] M. Vansteenkiste, C. P. Niemiec, and B. Soenens. The development of the five mini-theories of self-determination theory: An historical overview and emerging trends and future directions. In *The decade ahead: Theoretical perspectives on motivation and achievement*, pages 105–165. Emerald Group Publishing Limited, 2010.

[235] V. Venkatesh. Determinants of perceived ease of use: Integrating control and intrinsic motivation and emotion into the technology acceptance model. *Information systems research*, 11(4):342–365, 2000.

[236] E. Vilches, I. A. Escobar, E. E. Vallejo, and C. E. Taylor. Data mining applied to acoustic bird species recognition. In *Pattern Recognition and 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 400–403. IEEE, 2006.

[237] T. Vitone, K. A. Stofer, M. S. Steininger, J. Hulcr, R. Dunn, and A. Lucky. School of ants goes to college: integrating citizen science into the general education classroom increases engagement with science. *J. Sci. Commun.*, 15:1–24, 2016.

[238] N. Vos, H. Van Der Meijden, and E. Denessen. Effects of constructing versus playing an educational game on student motivation and deep learning strategy use. *Computers & Education*, 56(1):127–137, 2011.

[239] H. Wang, J. Elson, L. Girod, D. Estrin, and K. Yao. Target classification and localization in habitat monitoring. In *Acoustics and Speech and Signal Processing and 2003. Proceedings.(ICASSP03). 2003 IEEE International Conference on*, volume 4, pages 840–844. IEEE, 2003.

[240] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[241] E. Wenger. *Communities of practice: Learning and meaning and identity.* Cambridge university press, 1998.

[242] R. W. White. Motivation reconsidered: The concept of competence. *Psychological review*, 66(5):297, 1959.

[243] Y. F. Wiersma, J. Parsons, and R. Lukyanenko. Data quality in citizen science–a research study. *Environ. Sci*, 25(2):75–78, 2016.

[244] A. Wiggins and K. Crowston. From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS) and 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

[245] J. Wimmer, M. Towsey, P. Planitz, P. Roe, and I. Williamson. Scaling acoustic data analysis through collaboration and automation. In *e-Science (e-Science) and 2010 IEEE Sixth International Conference on*, pages 308–315. IEEE, 2010.

[246] N. Xu. A survey of sensor network applications. *IEEE Communications Magazine*, 40(8):102–114, 2002.

[247] Y. Xue, I. Davies, D. Fink, C. Wood, and C. P. Gomes. Avicaching: A two stage game for bias reduction in citizen science. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 776–785. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[248] H-F. Yang, J. Zhang, and P. Roe. Using reputation management in participatory sensing for data classification. *Procedia Computer Science*, 5:190–197, 2011.

[249] H-F. Yang, J. Zhang, and P. Roe. Reputation modelling in citizen science for environmental acoustic data analysis. *Social Network Analysis and Mining*, 3(3):419–435, 2013.

[250] J. Yu, W-K. Wong, and R. A. Hutchinson. Modeling experts and novices in citizen science data for species distribution modeling. In *Data Mining (ICDM) and 2010 IEEE 10th International Conference on*, pages 1157–1162. IEEE, 2010.

[251] J. Zhang, K. Huang, M. Cottman-Fields, A. Truskinger, P. Roe, S. Duan, X. Dong, M. Towsey, and J. Wimmer. Managing and analysing big audio data for environmental monitoring. In *Computational Science and Engineering (CSE) and 2013 IEEE 16th International Conference on*, pages 997–1004. IEEE, 2013.

[252] P. Zhang. Technical opinion motivational affordances: reasons for ict design and use. *Communications of the ACM*, 51(11):145–147, 2008.

[253] D. Zilli, O. Parson, G. V. Merrett, and A. Rogers. A hidden markov model-based acoustic cicada detector for crowdsourced smartphone biodiversity monitoring. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2945–2951. AAAI Press, 2013.

[254] E. Zimmerman. Play as research: The iterative design process. *Design research: Methods and perspectives*, 2003:176–184, 2003.

[255] J. Zimmerman, J. Forlizzi, and S. Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502. ACM, 2007.

APPENDIX A

Quantitative and qualitative data collected

All information stored to each database during play of each game for Touch-Level Model Goals, Operators, Methods, and Selection rules (TLM-GOMS) analysis is presented, followed by the surveys completed by each cohort.

A.I Game-play data collected

The following UML diagrams outline the structure of my game application development framework and expand specific data classes, showing detailed structure of the relationship between data collection on device and data exported for analysis.

A.I.1 Game data classes

The overall software framework structure (fig. A.I.1, pg. 250).

Figure A.I.1: Android classes for the application encompassing a framework for the three game interaction models.

A.I.1.1 Match game data class and ORM

Data collected and stored during BirdMatch play (fig. A.I.2, pg. 251).

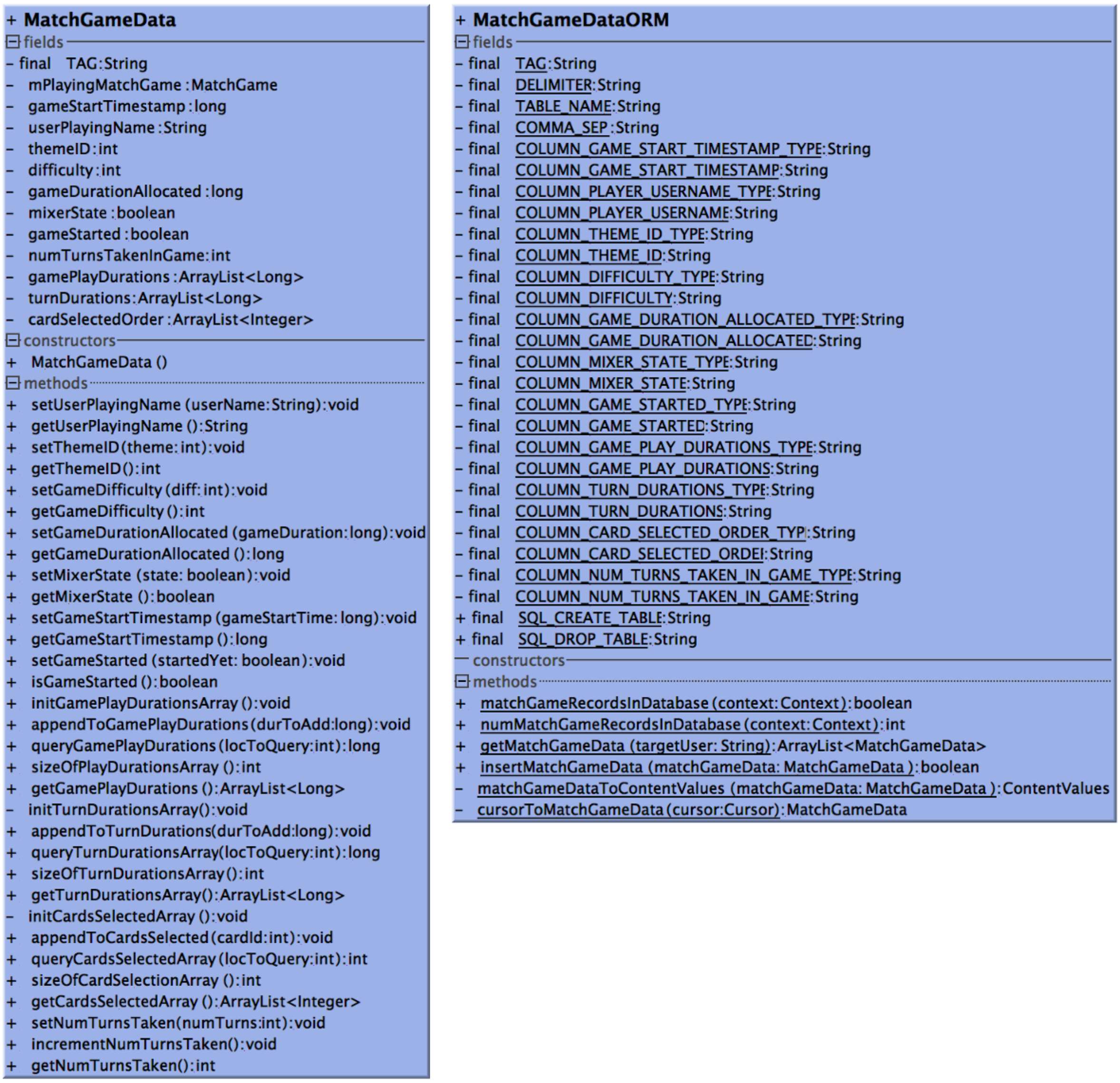

Figure A.I.2: Fields populated in the matchGameData class are saved to database via the matchGameDataORM class.

A.I.2 Swap game data class and ORM

Data collected and stored during SpectroPuzzle play (fig. A.I.3, pg. 252).

Figure A.I.3: Fields populated in the swapGameData class are saved to database via the swapGameDataORM class.

A.I.3 Composition toy data class and ORM

Data collected and stored during ChorusComposer play (fig. A.I.4, pg. 253).

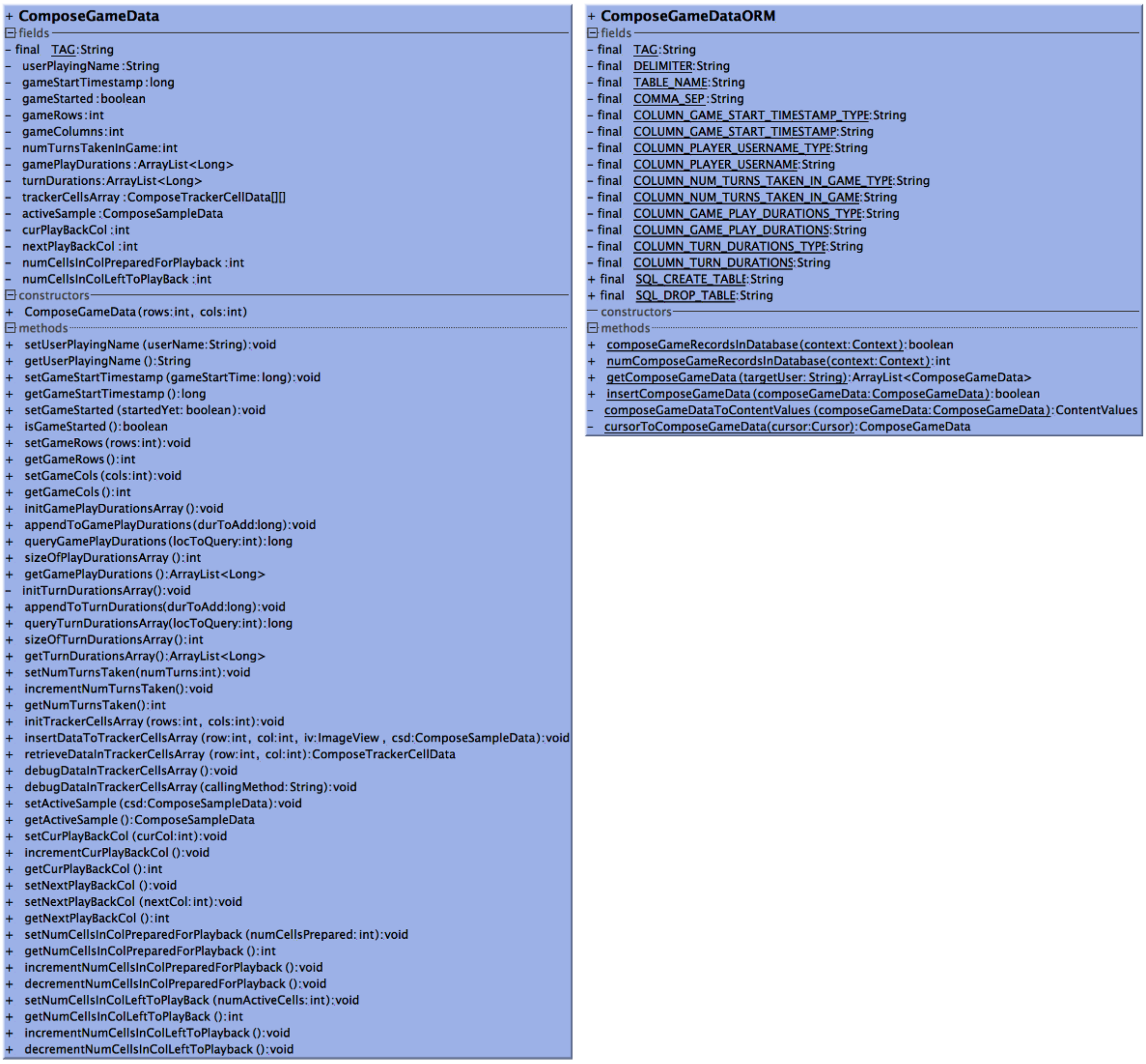

Figure A.I.4: Fields populated in the composeGameData class are saved to database via the composeGameDataORM class.

A.II Surveys

The questions asked of each cohort in the pre- and post-surveys are presented here, along with the expected format of response data.

A.II.1 2017 Pre-survey

Pre-Game Survey - First Game Deployment	
Question	**Response**
[Q.1] What is your name	Text
[Q.2] What is your age	Numeric
[Q.3] What is your gender	Text
[Q.4] How many kinds of wild birds can you identify by sight	Numeric
[Q.5] How many kinds of wild birds can you identify by sound	Numeric
[Q.6] Do you keep track of the birds that you see or hear	Boolean
[Q.7] When identifying birds, do you use both sight and sound	Boolean
[Q.8] If you hear a bird but do not see it, do you count it	Boolean
[Q.L1] I enjoy spending time in nature	Likert (1-5)
[Q.L2] When I am outdoors, I notice the wildlife around me	Likert (1-5)
[Q.L3] I am interested in learning more about wildlife	Likert (1-5)
[Q.L4] If I knew more about wildlife I would enjoy nature more	Likert (1-5)
[Q.L5] Being able to recognise bird songs makes being outdoors more enjoyable	Likert (1-5)

Table A.1: The pre-survey questions, and the type of data expected for each response. As results for questions 4 & 5 ranged widely, they were logarithmically scaled (base e) for analysis.

A.II.2 2017 Post-survey

Post-Game Survey - First Game Deployment	
Question	**Response**
[Q.1] What was your game name	Text
[Q.2] Have you previously played educational video games	Boolean
[Q.3] Which of the games did you like most	Numeric Rank
[Q.4] Which of the games did you like least	Numeric Rank
[Q.5] Which of the games did you find most challenging	Numeric Rank
[Q.6] Which of the games did you find easiest	Numeric Rank
[Q.7] Which version of the first game did you like most	Numeric Rank
[Q.8] Which version of the first game did you like least	Numeric Rank
[Q.9] Which version of the first game did you find most challenging	Numeric Rank
[Q.10] Which version of the first game did you find easiest	Numeric Rank
[Q.11] What part of the games caused the most confusion	Free-form
[Q.12] Explain what you liked and disliked about the games	Free-form
[Q.L1] I enjoyed playing the games	Likert (1-5)
[Q.L2] I would like to continue playing such games	Likert (1-5)
[Q.L3] The games helped me to learn birds by sight	Likert (1-5)
[Q.L4] The games helped me to learn birds by sound	Likert (1-5)
[Q.L5] I want to continue to learn more bird songs	Likert (1-5)
[Q.L6] Being able to recognise bird songs makes being outdoors more enjoyable	Likert (1-5)
[Q.L7] I would like to spend more time outdoors listening to birds	Likert (1-5)
[Q.L8] I was motivated to find solutions to the games	Likert (1-5)
[Q.L9] I was confident that my turns in game one led to winning	Likert (1-5)
[Q.L10] I was confident that my turns in game two led to winning	Likert (1-5)
[Q.L11] I am confident that I could recognise some of the birds from the games by sight	Likert (1-5)
[Q.L12] I am confident that I could recognise some of the birds from the games by sound	Likert (1-5)

Table A.2: The post-game survey questions, and the data type expected for each response.

A.II.3 2018 Pre-survey

Pre-Game Survey - Second Game Deployment	
Question	**Response**
[Q.1] What is your name	Text
[Q.2] What is your age	Numeric
[Q.3] What is your gender	Text
[Q.4 - Q.7] What do you hear	Text
[Q.8] How many kinds of owls can you identify by sight	Circle (0 - 4+)
[Q.9] How many kinds of owls can you identify by sound	Circle (0 - 4+)
[Q.10] When identifying birds, do you use both sight and sound	Boolean
[Q.L1] I enjoy spending time in nature	Likert (1-5)
[Q.L2] When I am outdoors, I notice the wildlife around me	Likert (1-5)
[Q.L3] I am interested in learning more about wildlife	Likert (1-5)
[Q.L4] If I knew more about wildlife I would enjoy nature more	Likert (1-5)
[Q.L5] I want to learn more about owls	Likert (1-5)
[Q.L6] Being able to recognise owl calls makes being outdoors more enjoyable	Likert (1-5)
[Q.L7] I would like to spend more time outdoors listening for owls	Likert (1-5)

Table A.3: The pre-survey questions for the second deployment of the games and the types of data expected for each response. Questions 4-7 ask participants to describe what they hear from a recording of an owl.

A.II.4 2018 Post-survey

Post-Game Survey - Second Game Deployment	
Question	**Response**
[Q.1] What is your name	Text
[Q.2] What was your *Login Name*	Text
[Q.3 - Q.6] What do you hear	Text
[Q.7] How many kinds of owls can you identify by sight	Circle (0 - 4+)
[Q.8] How many kinds of owls can you identify by sound	Circle (0 - 4+)
[Q.L1] I enjoyed playing the games	Likert (1-5)
[Q.L2] The games helped me to learn to recognise owls by sight	Likert (1-5)
[Q.L3] The games helped me to learn to recognise owls by sound	Likert (1-5)
[Q.L4] I want to learn more about owls	Likert (1-5)
[Q.L5] Being able to recognise owl calls makes being outdoors more enjoyable	Likert (1-5)
[Q.L6] I would like to spend more time outdoors listening for owls	Likert (1-5)
[Q.L7] I enjoy spending time in nature	Likert (1-5)
[Q.L8] When I am outdoors, I notice the wildlife around me	Likert (1-5)
[Q.L9] I am interested in learning more about wildlife	Likert (1-5)
[Q.L10] If I knew more about wildlife I would enjoy nature more	Likert (1-5)
[Q.L11] I am confident that I could recognise some of the owls from the games by sight	Likert (1-5)
[Q.L12] I am confident that I could recognise some of the owls from the games by sound	Likert (1-5)

Table A.4: The post-survey questions for the second deployment of the games and the type of data expected for each response. Questions 3 -6 ask participants to describe what they hear from playback of a recording of an owl.

APPENDIX B

Further survey results

The following survey results show inconclusive shifts in participant belief with play. For reference, summary belief statistics as well as significance test results are provided.

B.I Results for 6.1.2.1

While all response medians were $\tilde{x} = 5$, other summary statistics varied. The 2017 pre-survey results ($\mu = 4.48$, $\sigma^2 = 0.45$, N=136) are higher and less varied, respectively, than the 2018 pre-survey results ($\mu = 4.30$, $\sigma^2 = 1.01$, N=104). Play has a minor inhibitory effect on positive perceptions of nature, as 2018 post-survey results fell to ($\mu = 4.23\sigma^2 = 1.56$, N=69). I performed a one-tailed Wilcoxon's signed rank test, a non-parametric test — as results are not normally distributed — suitable for determining whether compared samples came from the same population with the same distribution. The test determines, in the first case, whether both cohorts can be assumed to have been drawn from the same population regarding affinity for nature and, in the second case, whether play has a discernible effect on participant engagement with nature. First comparison results are $p = 0.10$ and second $p = 0.21$; generally both cohorts agree they enjoy nature but in both cases p is insufficient to claim there is a significant difference between cohorts or with play at $p < 0.05$.

B.II Results for 6.1.2.2

While median responses for both cohorts prior to play were $\tilde{x} = 4$, for the second cohort, $\tilde{x}$ increased to 5 post play. The 2017 pre-survey results ($\mu = 4.09$, $\sigma^2 = 0.61$, N=135) are negligibly higher and narrowly less varied than the 2018 pre-survey results ($\mu = 4.08$, $\sigma^2 = 1.17$, N=101). The 2018 participants' belief that they would notice more wildlife when outdoors after game exposure fell insignificantly ($\mu = 4.07$, $\sigma^2 = 1.49$, N=67).

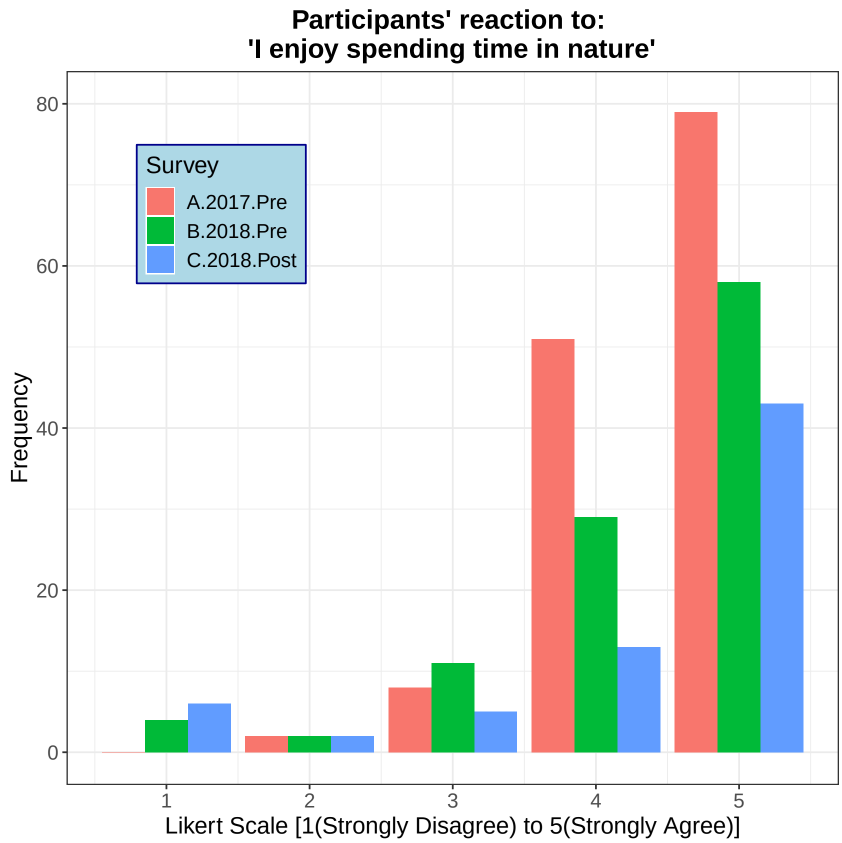

Figure B.I.1: Histogram of Likert-scaled responses to enjoyment of nature across both cohorts, the latter both pre- and post-play. Median support across cohorts is similarly strong, albeit dropping insignificantly ($p = 0.21$), with play.

One-tailed Wilcoxon's signed rank tests compared cohorts prior to play and the second cohort pre- and post-play. The former comparison was insignificant ($p = 0.47$) as was the latter for whether play changed second cohort beliefs ($p = 0.55$). Neither of these p-values is significant at $p < 0.05$, thus the null hypotheses, that there are no significant differences between Likert responses to this statement between cohorts, or pre- and post-play, cannot be rejected.

B.III Results for 6.3.2.1

The median response for the 2017 cohort before play was $\tilde{x} = 4$; for the second cohort, $\tilde{x} = 5$ pre- and post-play. The 2017 pre-survey results ($\mu = 4.16$, $\sigma^2 = 0.81$, N=133) are slightly higher and less varied, respectively, than the 2018 pre-survey results ($\mu = 4.06, \sigma^2 = 1.47$, N=100). The 2018 participants' belief that they were interested in learning more about wildlife after exposure to the games rose slightly ($\mu = 4.15$, $\sigma^2 = 1.30$, N=69). Cohort comparison yielded $p = 0.27$, examining whether play changes population belief for the second cohort yielded $p = 0.26$; while the two cohorts presented with the same question

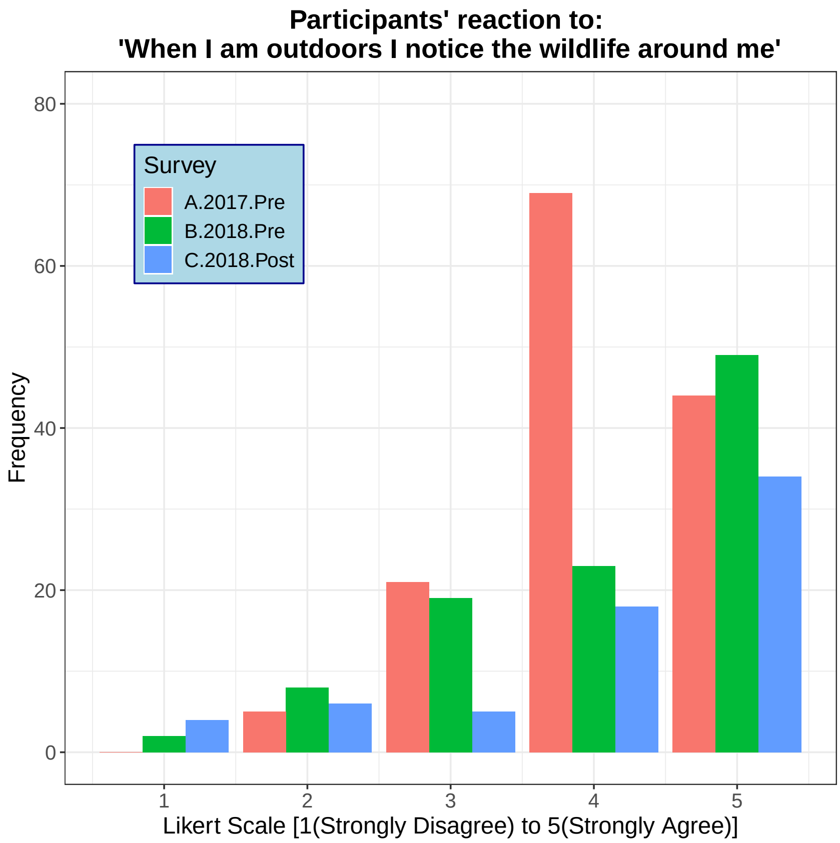

Figure B.II.2: Histogram of Likert responses to participants noticing wildlife when outdoors. Cohorts appears to be drawn from the same population and while affinity for nature is unaffected by play, it remains high.

appear to be drawn from the same population, there is no significant shift for the second cohort post game exposure.

B.IV Results for 6.3.2.2

Medians for both cohorts post-play were $\tilde{x} = 4$; this remained unchanged from pre-play for 2018 participants. The 2017 post-survey results ($\mu = 3.72, \sigma^2 = 0.86$, N=100) are slightly higher, albeit less varied than the 2018 post-survey results ($\mu = 3.71, \sigma^2 = 2.03$, N=69); mean results show weak positive *a priori* desire for engagement with nature. Before game exposure, the 2018 participants' belief that they wanted to spend more time outdoors listening for birds had been higher, albeit slightly less varied ($\mu = 3.86$, $\sigma^2 = 1.91$, N=97). The two cohorts can be assumed to have been randomly drawn from the same population ($p = 0.98$). Second cohort pre- and post-play analysis saw mean desire for engagement with nature fall, albeit insignificantly ($p = 0.18$).

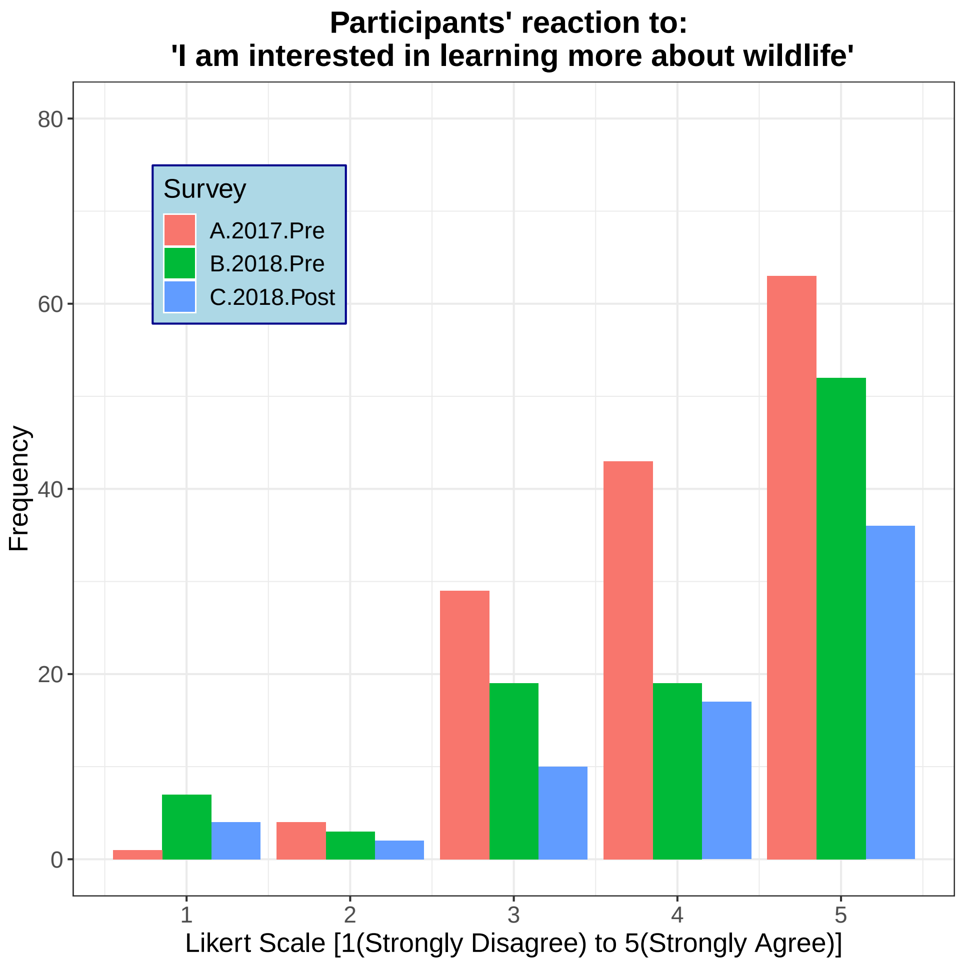

Figure B.III.3: Histogram of Likert responses to the premise that the participants are interested in learning about wildlife. While initially positive, responses diverge with play.

B.V Results for 6.3.2.4

Medians for all responses were $\tilde{x} = 4$. The 2017 pre-survey results ($\mu = 4.08$, $\sigma^2 = 0.80$, N=133) are slightly higher but more varied than the 2017 post-survey results ($\mu = 3.98$, $\sigma^2 = 0.77$, N=103), while the 2018 pre-survey results ($\mu = 3.85, \sigma^2 = 2.07$, N=98) are slightly lower and more varied than the 2018 post-survey results ($\mu = 3.88, \sigma^2 = 1.84$, N=68). Neither pre- ($p = 0.22$)nor post-play ($p = 0.97$) are the two cohorts significantly different from each other. Nor, however, are either cohort's shifts upon play significant; first cohort $p = 0.32$, second cohort $p = 0.32$, so it cannot be claimed that play changes support for this concept.

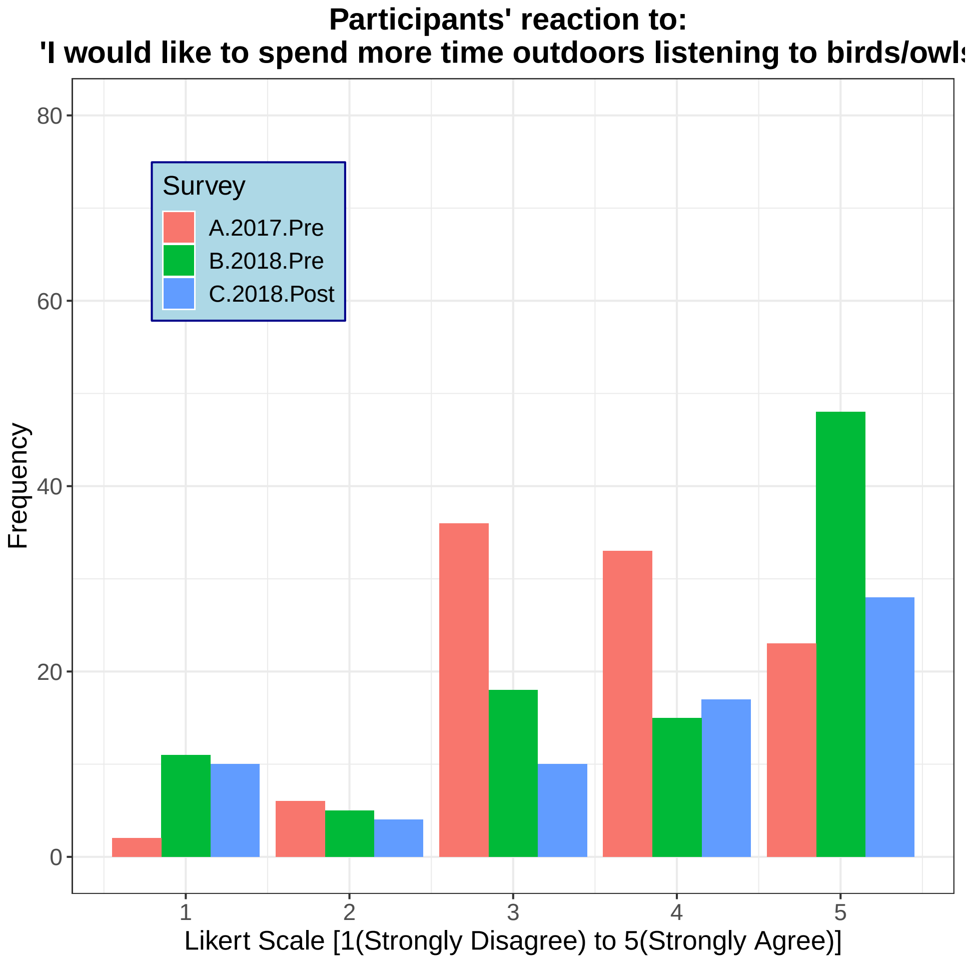

Figure B.IV.4: Histogram of Likert-scaled responses to the question of whether participants wanted to spend more time in nature listening for avian utterances, prerequisite for citizen scientist data collection. This falls, although not significantly, post-play.

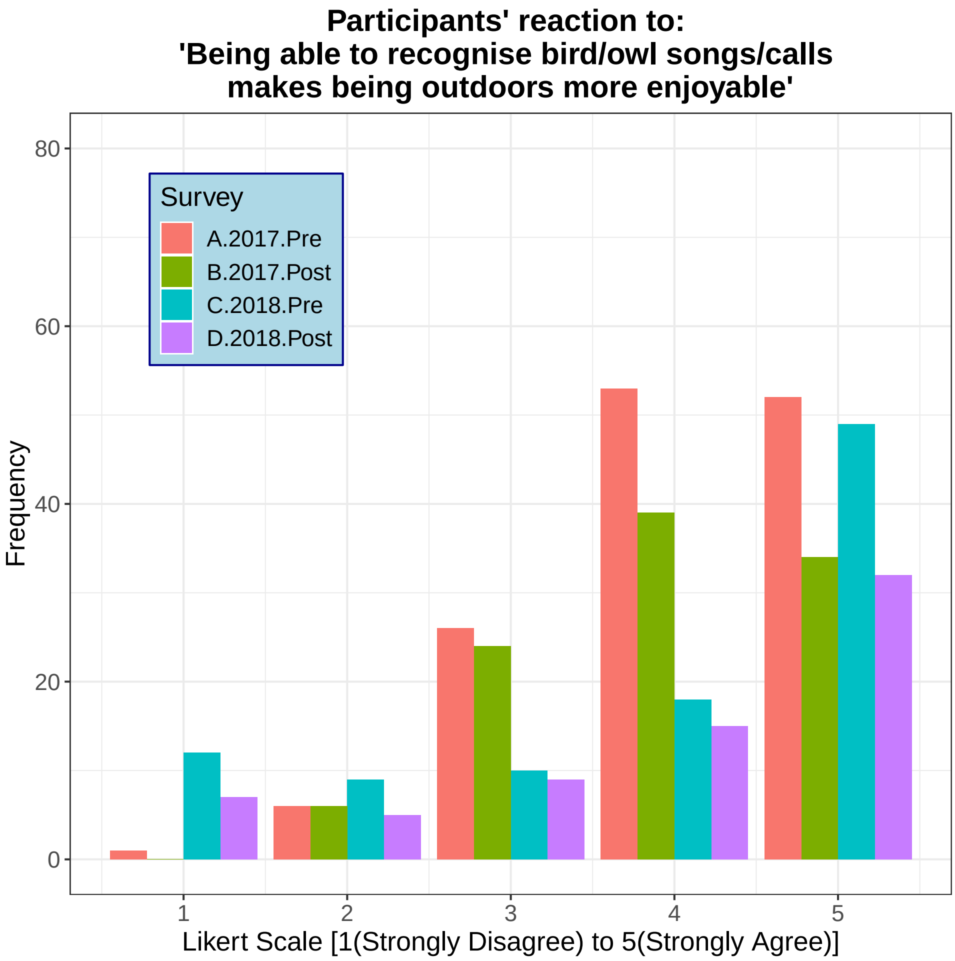

Figure B.V.5: Histogram of Likert-scaled responses to whether participants found having knowledge of avian utterances contributed to enjoyment of nature. Consensus was that participants in both cohorts were comfortable associating avian utterance familiarity with enjoyment of nature, however this marginally decreased with game play.

APPENDIX C

Match game results

Results begin with tables and figures providing complete summary statistics for the first three iterations of play, or fewer if insufficient participants played to a given iteration, for each data-representation/difficulty variant. This is followed by summaries of mean and median play success evolution scores across the first two transitions between game iterations.

C.I Summary statistics and plots

Results from the first three iterations of play, for which sufficient games were played, for each data-representation, are summarised by difficulty in the following sections.

C.I.1 Difficulty Easy, Both Cohorts, All Modes, L_s & L_e Results

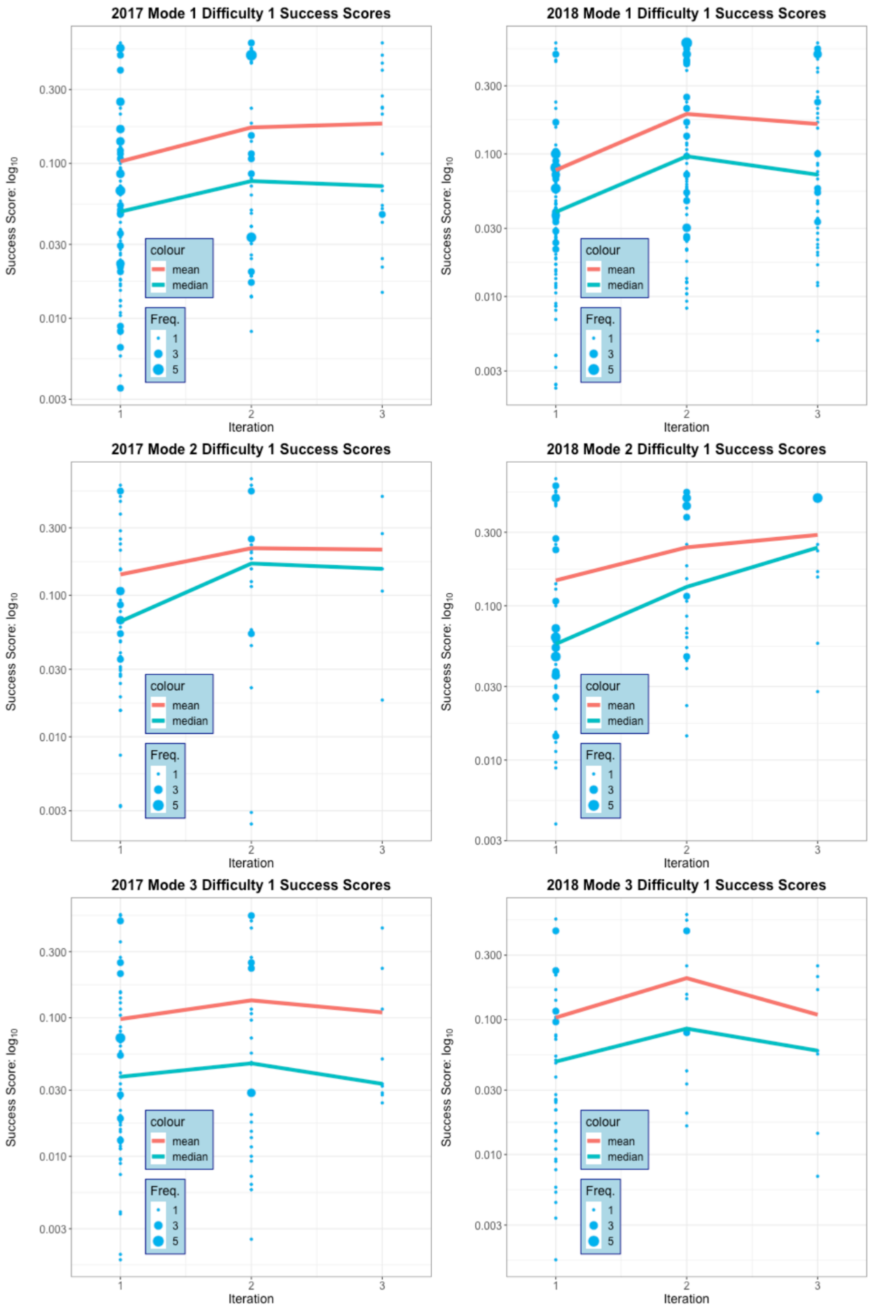

Figure C.I.1: L_s rates across each mode over the first three iterations of play for both cohorts when difficulty is set to easy.

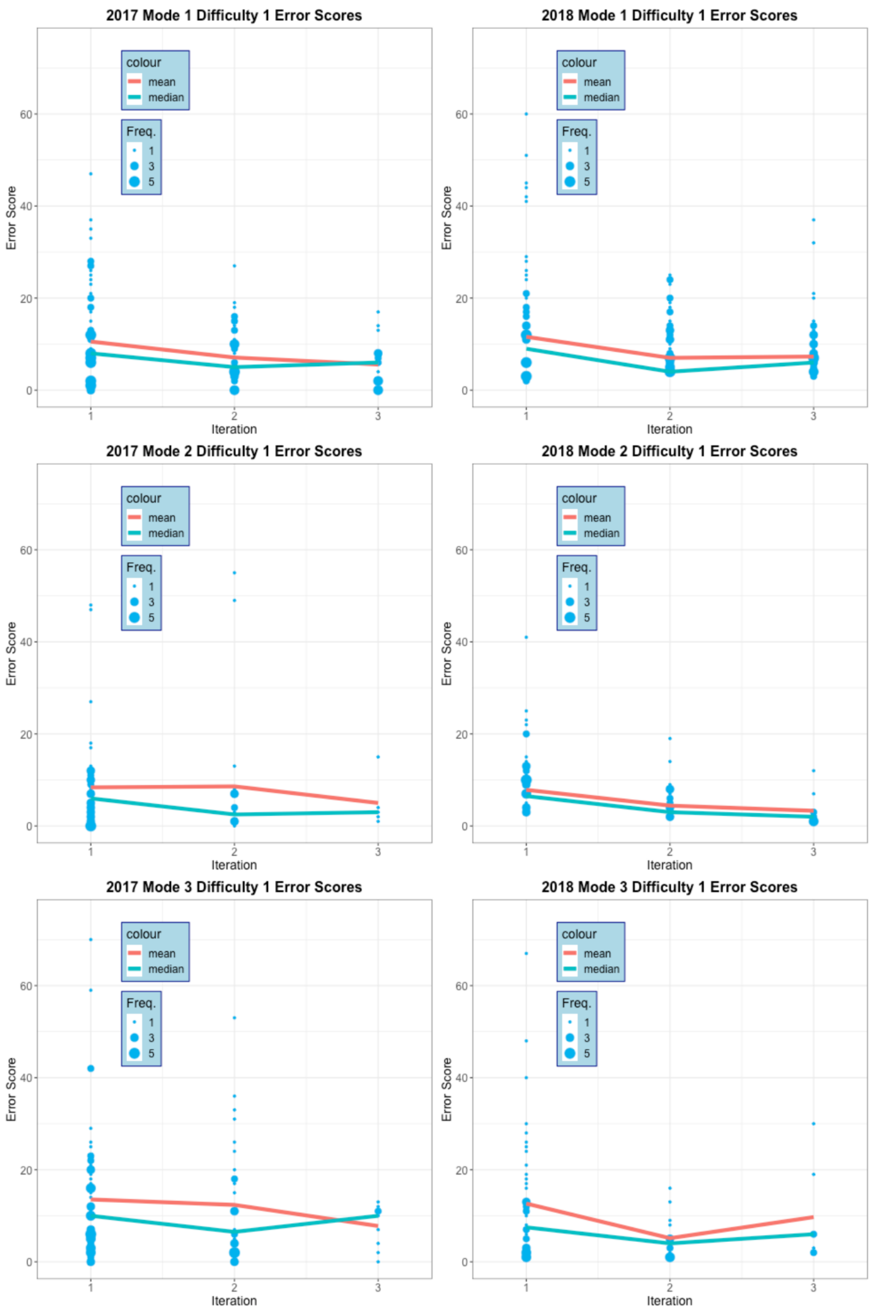

Figure C.I.2: L_e rates across each mode over the first three iterations of play for both cohorts when difficulty is set to easy.

C.I.2 Difficulty Medium, Both Cohorts, All Modes, L_s & L_e Results

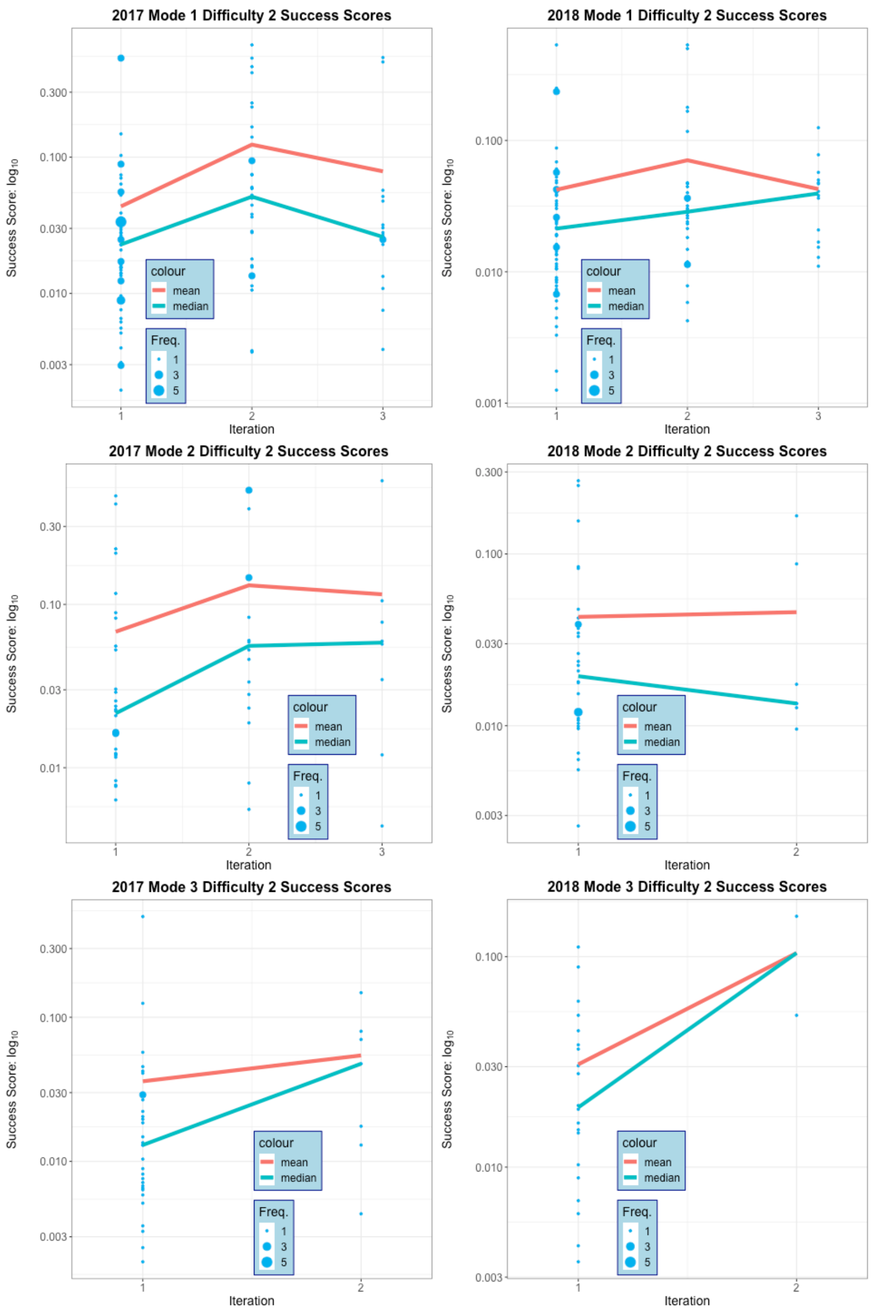

Figure C.I.3: L_s rates across each mode over the first three iterations of play, as possible, for both cohorts when difficulty is set to medium.

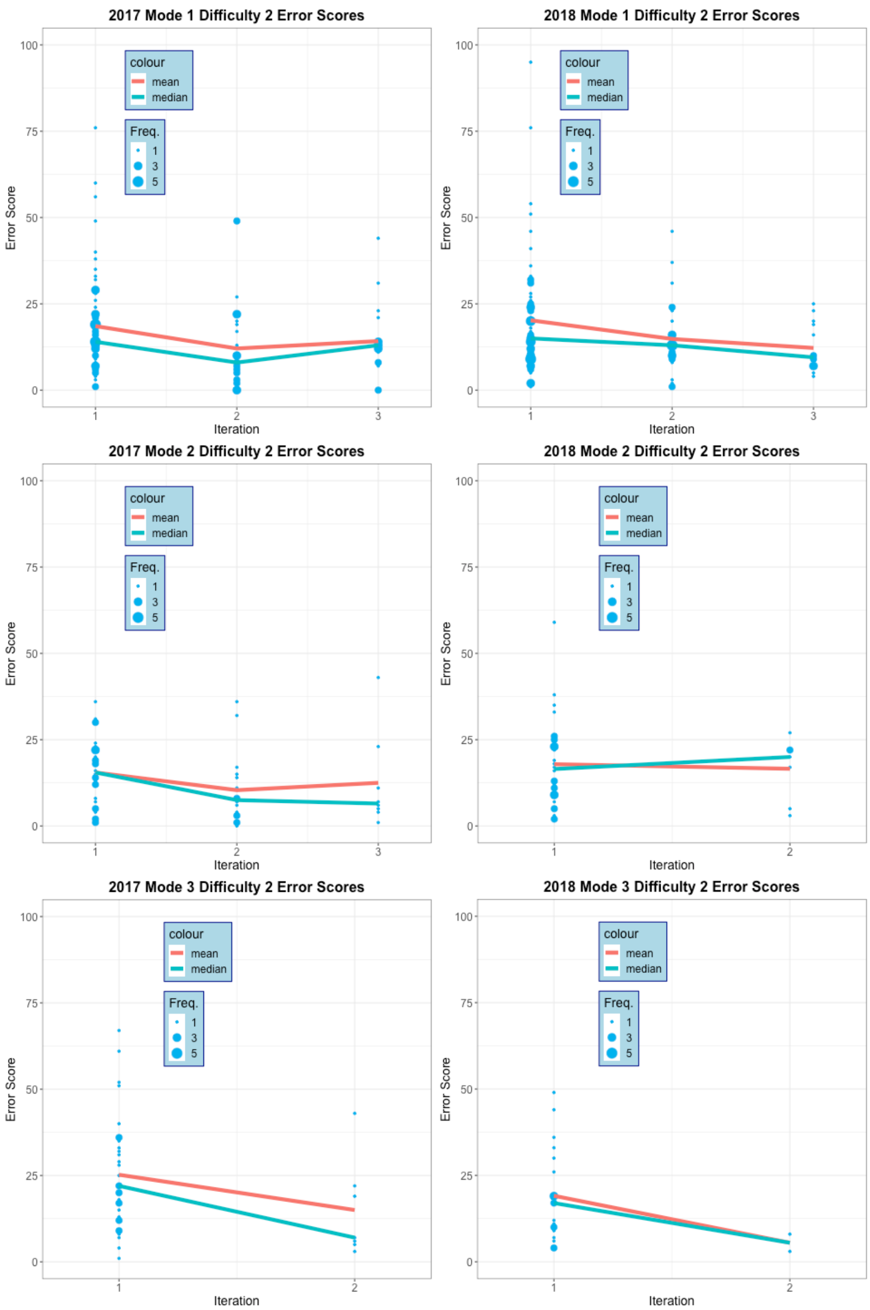

Figure C.I.4: L_e rates across each mode over the first three iterations of play, as possible, for both cohorts when difficulty is set to medium.

C.I.3 Difficulty Hard, Both Cohorts, All Modes, L_s & L_e Results

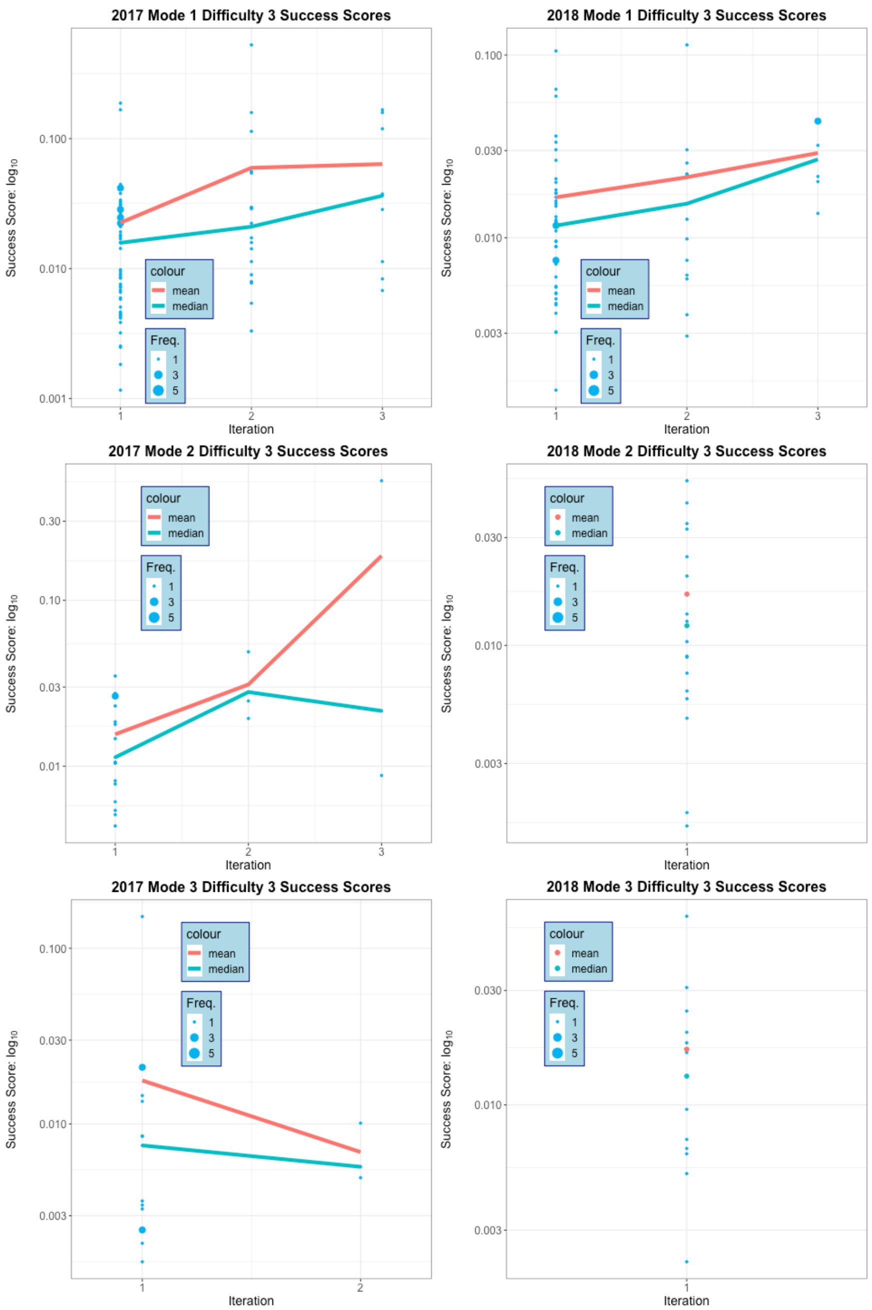

Figure C.I.5: L_s rates across each mode over the first three iterations of play, as possible, for both cohorts when difficulty is set to hard.

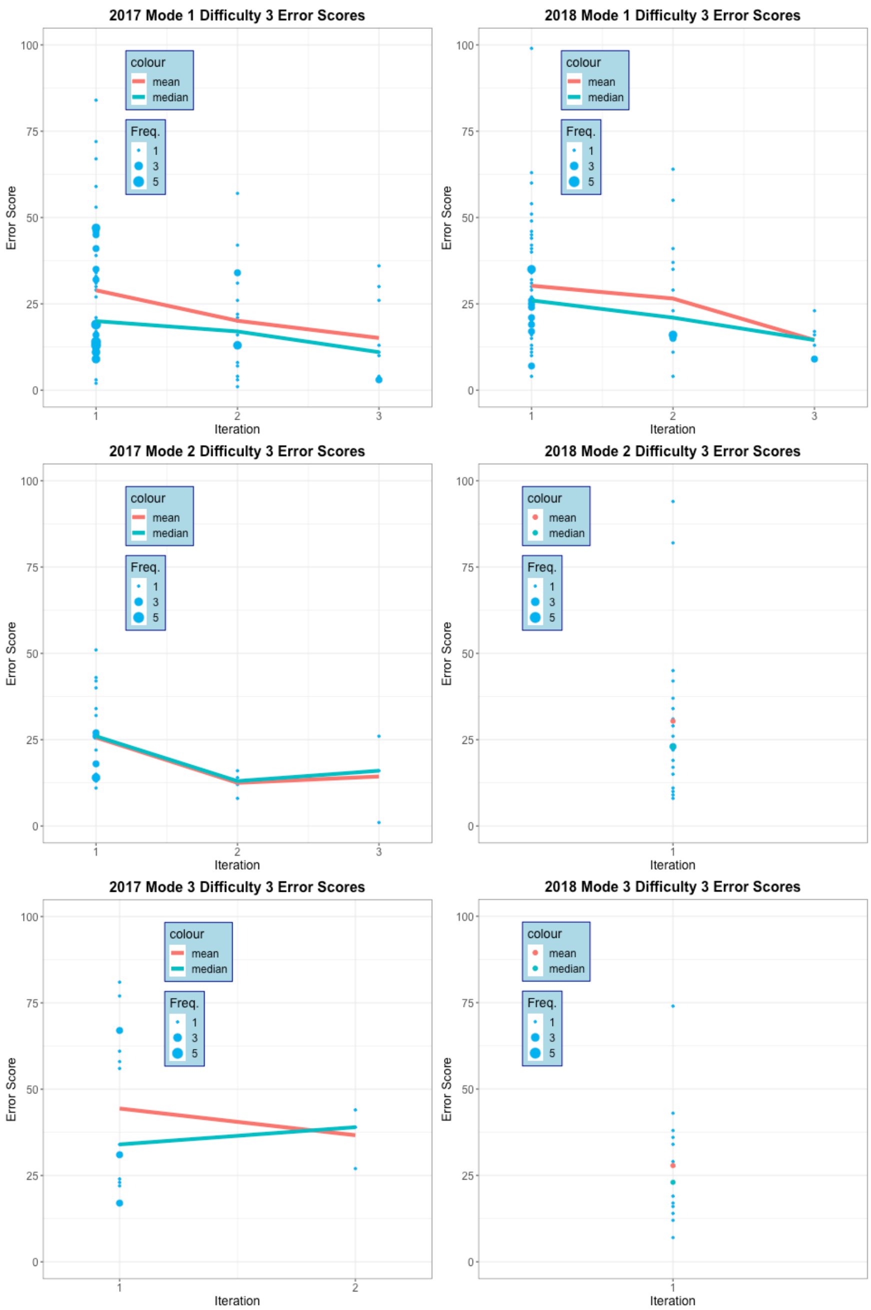

Figure C.I.6: L_e rates across each mode over the first three iterations of play, as possible, for both cohorts when difficulty is set to hard.

C.I.4 Success scores and success evolution

The following tables present the learning trajectories and knowledge evolution over iterations of play resulting from computations described in 6.2.2.3. Median evolutions are considered more likely to reflect population shifts, while means are susceptible to strong individual influence, both are presented here for completeness.

C.I.4.1 First cohort mean play success evolution

2017 Mean play success evolution			
Variant	μ_{μ_α}	μ_{μ_β}	σ_{τ_μ}
Mode 1, Difficulty Easy	1.66	1.06	0.64
Mode 1, Difficulty Medium	2.84	0.64	0.22
Mode 1, Difficulty Hard	2.64	1.07	0.40
Mode 2, Difficulty Easy	1.53	0.98	0.64
Mode 2, Difficulty Medium	1.92	0.88	0.46
Mode 2, Difficulty Hard	2.00	5.97	2.99
Mode 3, Difficulty Easy	1.37	0.82	0.60
Mode 3, Difficulty Medium	1.51	N/A	N/A
Mode 3, Difficulty Hard	0.39	N/A	N/A

Table C.4: Table shows the 2017 mean of success transition means between the first two iterations (α) and the second two iterations (β) of play for each data-representation/difficulty variant along with a measure of play success evolution, σ_τ. Blue denotes a learning effect, yellow, insufficient data. All but the hardest difficulty of the third data-representation saw a learning effect in the first transition.

C.I.4.2 First cohort median play success evolution

2017 Median play success evolution			
Variant	$\mu_{\tilde{x}_\alpha}$	$\mu_{\tilde{x}_\beta}$	$\sigma_{tau_{\tilde{x}}}$
Mode 1, Difficulty Easy	1.58	0.93	0.59
Mode 1, Difficulty Medium	2.24	0.51	0.23
Mode 1, Difficulty Hard	1.33	1.73	1.30
Mode 2, Difficulty Easy	2.57	0.92	0.38
Mode 2, Difficulty Medium	2.59	1.05	0.40
Mode 2, Difficulty Hard	2.47	0.77	0.31
Mode 3, Difficulty Easy	1.25	0.71	0.57
Mode 3, Difficulty Medium	3.67	N/A	N/A
Mode 3, Difficulty Hard	0.77	N/A	N/A

Table C.5: Table shows the 2017 mean of success transition medians between the first two iterations (α) and the second two iterations (β) of play for each data-representation/difficulty variant along with a measure of play success evolution, σ_τ. Blue denotes a learning effect, yellow, insufficient data.

C.I.4.3 Second cohort mean play success evolution

2018 Mean play success evolution			
Variant	μ_{μ_α}	μ_{μ_β}	σ_{τ_μ}
Mode 1, Difficulty Easy	2.48	0.85	0.34
Mode 1, Difficulty Medium	1.68	0.60	0.36
Mode 1, Difficulty Hard	1.29	1.36	1.05
Mode 2, Difficulty Easy	1.64	1.20	0.74
Mode 2, Difficulty Medium	1.07	N/A	N/A
Mode 2, Difficulty Hard	N/A	N/A	N/A
Mode 3, Difficulty Easy	1.96	0.54	0.27
Mode 3, Difficulty Medium	3.38	N/A	N/A
Mode 3, Difficulty Hard	N/A	N/A	N/A

Table C.6: Table shows the 2018 mean of success transition means between the first two iterations (α) and the second two iterations (β) of play for each data-representation/difficulty variant along with a measure of play success evolution, σ_τ. Blue denotes a learning effect, yellow, insufficient data.

C.I.4.4 Second cohort median play success evolution

2018 Median play success evolution			
Variant	$\mu_{\tilde{x}_\alpha}$	$\mu_{\tilde{x}_\beta}$	$\sigma_{tau_{\tilde{x}}}$
Mode 1, Difficulty Easy	2.46	0.74	0.30
Mode 1, Difficulty Medium	1.34	1.38	1.03
Mode 1, Difficulty Hard	1.32	1.75	1.33
Mode 2, Difficulty Easy	2.34	1.80	0.77
Mode 2, Difficulty Medium	0.69	N/A	N/A
Mode 2, Difficulty Hard	N/A	N/A	N/A
Mode 3, Difficulty Easy	1.76	0.69	0.39
Mode 3, Difficulty Medium	5.42	N/A	N/A
Mode 3, Difficulty Hard	N/A	N/A	N/A

Table C.7: Table shows the 2018 mean of success transition medians between the first two iterations (α) and the second two iterations (β) of play for each data-representation/difficulty variant along with a measure of play success evolution, σ_{τ}. Blue denotes a learning effect, yellow, insufficient data.

C.I.4.5 L_s & L_e α- & β-transition significance

BirdMatch Learning & Error Summary Easy Difficulty, First 3 Iterations, All Data-representations													
Iteration	Year	2017						2018					
	Mode	1		2		3		1		2		3	
	Statistic	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e
1	N	92		46		59		99		66		44	
	μ	0.10	10.6	0.14	8.4	0.10	13.5	0.08	11.6	0.15	7.9	0.10	12.7
	$\tilde{x}$	0.05	8	0.07	6	0.04	10	0.04	9	0.06	6	0.05	7.5
	σ^2	0.02	87.5	0.03	100.3	0.02	185.9	0.01	122.6	0.04	49.8	0.02	188.4
2	N	44		20		30		75		28		15	
	μ	0.17	7.1	0.22	8.6	0.13	12.4	0.19	7	0.24	4.4	0.20	5.1
	$\tilde{x}$	0.08	5	0.17	2.5	0.05	6.5	0.10	4	0.13	3	0.09	4
	σ^2	0.04	40.4	0.04	231.3	0.03	171	0.04	45.9	0.04	19.3	0.04	20.8
3	N	19		5		9		53		10		7	
	μ	0.18	5.6	0.21	5	0.11	7.8	0.16	7.3	0.29	3.3	0.11	9.7
	$\tilde{x}$	0.07	6	0.15	3	0.03	10	0.07	6	0.24	2	0.06	6
	σ^2	0.03	25.4	0.04	32.5	0.02	22.4	0.03	54.8	0.04	12.7	0.01	114.9

Table C.1: Summary statistics for L_s and L_e at easy difficulty for all data-representations over the first three iterations of play; at least 5 participants played each variant.

BirdMatch Learning & Error Summary Statistics Medium Difficulty, First 3 Iterations, All Data-representations													
Iteration	Year	2017						2018					
	Mode	1		2		3		1		2		3	
	Statistic	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e
1	N	67		30		31		63		32		20	
	μ	0.04	18.6	0.07	15.5	0.04	25.2	0.04	20.2	0.04	17.9	0.03	19.1
	$\tilde{x}$	0.02	14	0.02	15.5	0.01	22	0.02	15	0.02	16.5	0.02	17
	σ^2	0.01	193.6	0.01	93	0.01	272.4	0.01	280.4	<0.01	149.7	<0.01	175
2	N	30		16		7		30		7		2	
	μ	0.12	12.1	0.13	10.4	0.05	15	0.07	14.8	0.05	16.6	0.10	5.5
	$\tilde{x}$	0.05	8	0.06	7.5	0.05	7	0.03	13	0.01	20	0.10	5.5
	σ^2	0.03	158.3	0.03	111.9	0.03	206.3	0.02	100.6	0.03	83	0.01	17.5
3	N	19		8		—		14		—		—	
	μ	0.8	14.2	0.12	12.5	–	–	0.04	12.2	–	–	–	–
	$\tilde{x}$	0.03	13	0.06	6.5	–	–	0.04	9.5	–	–	–	–
	σ^2	0.02	103.1	0.04	196.6	–	–	<0.01	48.6	–	–	–	–

Table C.2: Summary statistics for L_s and L_e at medium difficulty for all data-representations over the first three iterations of play for variants played by multiple participants.

BirdMatch Learning & Error Summary Statistics Hard Difficulty, First 3 Iterations, All Data-representations													
Iteration	Year	2017						2018					
	Mode	1		2		3		1		2		3	
	Statistic	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e	L_s	L_e
1	N	55		19		15		43		19		13	
	μ	0.02	28.9	0.02	25.6	0.02	44.4	0.02	30.3	0.02	30.4	0.02	27.9
	$\tilde{x}$	0.02	20	0.01	26	0.01	34	0.01	26	0.01	23	0.01	23
	σ^2	<0.01	434.2	<0.01	142	<0.01	526	<0.01	328.5	<0.01	534	<0.01	317.1
2	N	19		4		3		15		—		—	
	μ	0.06	20.1	0.03	12.5	0.01	36.7	0.02	26.5	–	–	–	–
	$\tilde{x}$	0.02	17	0.03	13	0.01	39	0.02	21	–	–	–	–
	σ^2	0.01	211	<0.01	11.7	<0.01	76.3	<0.01	285.8	–	–	–	–
3	N	9		3		—		6		—		—	
	μ	0.06	15.1	0.19	14.3	–	–	0.03	14.5	–	–	–	–
	$\tilde{x}$	0.04	11	0.02	16	–	–	0.03	14.5	–	–	–	–
	σ^2	<0.01	155.1	0.09	158.3	–	–	<0.01	28.7	–	–	–	–

Table C.3: Summary statistics for L_s and L_e at hard difficulty for all data-representations over the first three iterations of play for variants played by multiple participants

Both cohorts, L_s & L_e transition significance, $p_{\alpha,\beta}$, by representation									
Rep.	Diff.	2017				2018			
		p_α		p_β		p_α		p_β	
		L_e	L_s	L_e	L_s	L_e	L_s	L_e	L_s
1	Easy	0.02	0.96	0.26	0.44	$7.5*10^{-6}$	1.00	0.70	0.13
	Med.	$2.4*10^{-3}$	0.999	0.88	0.01	0.03	0.92	0.10	0.67
	Hard	$3.3*10^{-3}$	0.997	0.21	0.85	0.14	0.83	0.02	0.98
2	Easy	0.17	0.98	0.63	0.58	$1.4*10^{-4}$	0.998	0.27	0.76
	Med.	0.02	0.98	N/A	N/A	0.45	0.48	N/A	N/A
	Hard	$2.9*10^{-4}$	0.999	0.69	0.44	N/A	N/A	N/A	N/A
3	Easy	0.41	0.75	0.12	0.36	$1.9*10^{-3}$	0.99	0.93	0.07
	Med.	0.03	0.99	N/A	N/A	$2.4*10^{-4}$	0.999	N/A	N/A
	Hard	0.11	0.17	N/A	N/A	N/A	N/A	N/A	N/A

Table C.8: α- and β-transition significance for L_e and L_s across both cohorts. Blue denotes significance, yellow denotes insufficient data.