Behavior Research Methods
https://doi.org/10.3758/s13428-023-02289-w

ORIGINAL MANUSCRIPT

Using a cognitive model to understand crowdsourced data from citizen scientists

Alex Thorpe[1] · Oliver Kelly[2] · Alex Callen[2] · Andrea S. Griffin[2] · Scott D. Brown[1]

Accepted: 3 November 2023

Abstract
Threatened species monitoring can produce enormous quantities of acoustic and visual recordings which must be searched for animal detections. Data coding is extremely time-consuming for humans and even though machine algorithms are emerging as useful tools to tackle this task, they too require large amounts of known detections for training. Citizen scientists are often recruited via crowd-sourcing to assist. However, the results of their coding can be difficult to interpret because citizen scientists lack comprehensive training and typically each codes only a small fraction of the full dataset. Competence may vary between citizen scientists, but without knowing the ground truth of the dataset, it is difficult to identify which citizen scientists are most competent. We used a quantitative cognitive model, cultural consensus theory, to analyze both empirical and simulated data from a crowdsourced analysis of audio recordings of Australian frogs. Several hundred citizen scientists were asked whether the calls of nine frog species were present on 1260 brief audio recordings, though most only coded a fraction of these recordings. Through modeling, characteristics of both the citizen scientist cohort and the recordings were estimated. We then compared the model's output to expert coding of the recordings and found agreement between the cohort's consensus and the expert evaluation. This finding adds to the evidence that crowdsourced analyses can be utilized to understand large-scale datasets, even when the ground truth of the dataset is unknown. The model-based analysis provides a promising tool to screen large datasets prior to investing expert time and resources.

Keywords Citizen science · Crowdsourcing · Data aggregation · Data analysis

Introduction

The Information Revolution of the 21st century has improved our data-gathering ability via crowdsourcing – technologies such as the Internet and smartphone now connect people across time and space to create, streamline, and innovate. For example, Google includes a collaborative tool for users to add road information to the Google Maps app as a self-improving service. Crowdsourcing breaks big tasks into small chunks for diverse groups of people to tackle. Crowdsourcing expedites the process and improves scalability through spreading workload, reduces running costs by engaging people outside of a company payroll, and can generate self-sustaining promotion. While companies have been quick to realize the commercial benefits of crowdsourcing (Doritos, for example, are famous for their Super Bowl ad campaign because it includes inspiration from everyday people), the knowledge contribution of the global community as co-producers of public services has only recently been realized (Liu, 2021). Psychological researchers were early adopters of crowdsourcing as a means of data collection, e.g., through online marketplaces such as Prolific Academic and Amazon's Mechanical Turk. More recently, psychological researchers have also adopted crowdsourcing to tackle traditionally labor-intensive data analysis problems, such as the manual coding of videos and images (e.g., Lasecki et al. 2015; Root-Gutteridge et al. 2021; Vondrick et al. 2013).

Researchers in ecological sciences also have a history of using crowdsourcing for similar human-centric data analysis tasks. Threatened species monitoring is crucial for conservation, as it identifies how species abundance and activity fluctuate over time, their response to threats, and where to best focus resources to secure their survival (Martin et

✉ Scott D. Brown
scott.brown@newcastle.edu.au

[1] School of Psychological Sciences, University of Newcastle, Callaghan, Australia

[2] School of Environmental and Life Sciences, University of Newcastle, Callaghan, Australia

Published online: 29 November 2023

al., 2007). Understanding these ecological trends requires data sets that are spatially and temporally extensive. Unfortunately, traditional ecological surveying methods can be labor-intensive, time-consuming, and often result in few positive detections of highly seasonal threatened species (Cristescu et al., 2012; Kristensen & Kovach, 2018; Penman et al., 2006). Moreover, long-term monitoring studies are expensive, and conservation science is a public service already struggling with inadequate resourcing (Bakker et al., 2010). In a bid to overcome these hurdles, research groups and conservation organizations have increasingly turned to citizen scientists with their smartphones and Internet to collect knowledge that will assist in solving the wicked problem of biodiversity extinctions in the Anthropocene (Irwin, 2018; McKinley et al., 2017; Trouille et al., 2019). The current in-kind contribution of citizen scientists to science globally is estimated to be well over US $600 million annually, and the data sets this workforce yields are increasingly published in the scientific literature (Theobald et al., 2015). More than 60% of projects offered by Earthwatch, one of the first organizations to offer opportunities for citizens to partake in conservation science, are now published in peer-reviewed scientific journals (Chandler et al., 2017).

Although citizen science has a long history dating back to the 18th century (Silvertown, 2009), the increase in public participation in biodiversity conservation is commensurate with the technological advances of the information revolution and the birth of several crowdsourced digital projects since the turn of the century (Feldman et al., 2021; Kullenberg & Kasperowski, 2016). For example, iNaturalist identifies itself as an "online social network of people sharing biodiversity information to help each other learn about nature"[1]. The platform uses an online database and mobile app to help the community identify and record biodiversity observations and to create an online space where a community of experts provide consensus on identification. Since its formal launch by the California Academy of Sciences in 2008, iNaturalist has generated over 115 million observations of plants and animals worldwide collected by some 350,000 citizens and transformed them into data that are validated and subsequently used by the scientific community (Altrudi, 2021). Similarly, a global network of over 500,000 birdwatching citizens and scientists use the freely available application "e-bird" to record their observations. This now constitutes the world's largest scientific data collection effort with over 100 million bird sightings globally. Birdlife International has used these crowdsourced data to identify which birds are at greatest risk of extinction and should be added to the IUCN's Red List Index (Thibaut et al., 2010).

[1] http://inaturalist.org/pages/what+is+it

While technology and community seems like a match made in heaven, the sheer volume of data now being collected presents a raft of new problems for scientists – data quality, validity, rigor, and confidence (e.g., Brown et al., 2018; Leibovici et al., 2017; Lovett et al., 2018; Weeser et al., 2018). This is a sticking point despite recent advances in artificial intelligence (AI) to support information processing. While computers can learn to know what humans know, their computational algorithms still rely in large part on human-generated data for training, validation, and testing (Ellwood et al., 2015). It can take an AI algorithm more than 10,000 human-labeled images or sounds to reliably recognize and classify visual or acoustic stimuli when they are embedded in a background of noisy information, such as a small animal obscured by vegetation caught on a camera trap, or one frog species calling among multiple others. The same training needs arise when the stimuli are highly variable, such as geographically distinct song dialects of birds and whales, the acoustic mimicry of many birds, or when the background information is highly variable, such as animal sounds amongst forest, urban, and ocean soundscapes. In the psychological domain, to label the emotional content of parent–child interaction videos, training an AI algorithm may take more than 10,000 labeled examples – that is 10,000 images or sounds reliably identified by humans for accurate automated image or audio recognition by computers. Such expert coding would require weeks of work by researchers, and with profound differences between the visual and acoustic presentation of different species, this work would need to be repeated for each new study until a sufficiently large database of tagged examples could effectively train an AI model. This time investment is prohibitive for many researchers.

Data quality is particularly problematic when only subsets of data are compared with expert derived data (Clare et al., 2019), where consensus (agreement by multiple volunteers) is a measure of data quality without understanding how individuals make decisions (Kutlu et al., 2020), and where psychological biases can lead to labeling errors (Callaghan et al., 2021). Group-level consensus can be established with simple counting methods, such as a majority vote or a ranked-choice voting method such as Borda counts, and post hoc statistical methods can be applied to improve the quality of crowdsourced data, such as automated screening of inaccurate data or statistical models that account for individual and demographic differences (Clare et al., 2019; Kosmala et al., 2016), but these methods do not take into account the psychological processes that affect how human analysts make decisions. Advances in this area may both increase the confidence that crowdsourced data are valid and improve the efficiency of data-labeling processes on digital platforms.

Aggregating crowdsourced data

We address a problem in the use of citizen science data known as the "data aggregation" problem. This problem is caused by fundamental individual differences between people, and arises when crowdsourcing is used to label images or sounds manually and the results from many different citizen scientists are combined. Recent reviews of the citizen science approach have identified data aggregation as a key barrier to using crowdsourced data analyses (e.g., Garcia-Molina et al., 2016; Lukyanenko et al., 2020). There exist several statistical approaches to addressing the data aggregation problem which are not based on psychological theory. For example, a ranked-choice voting method such as Borda counts can be used to aggregate conflicting rankings provided by different people and find the most preferred option (Emerson, 2013). Bird et al. (2014), Isaac et al. (2014), and Jones et al. (2018) have all used atheoretical statistical methods to combine manually coded responses in crowdsourced ecological data analyses. While these approaches are statistically efficient, they make no attempt to address the well-known biases in human judgements, or the variation in skill and motivation of the citizen scientists. This results in the problem that aggregating data, without psychological theory, leaves researchers vulnerable to confusing patterns produced by human bias or incompetence with genuine, ground-truth patterns.

Psychological researchers have addressed related issues from a different perspective, focusing on the cognitive processes used by human analysts. Research into the "wisdom of the crowd" has identified applications in which solutions aggregated across a crowd of non-experts out-perform solutions provided by even the best experts, human or machine (for case studies, see Surowiecki, 2005). In the simplest cases, aggregating crowdsourced predictions or estimates using simple statistical averages is sufficient. In more complex tasks, such as betting on sporting events or solving a traveling salesman problem, researchers use cognitive models to aggregate across individuals (Lee et al., 2014; Yi et al., 2012). There are important differences between quantifying the wisdom of the crowd and solving data aggregation problem within crowdsourced data analyses. Research on wisdom of the crowd relies on researchers having access to the "ground truth", so that the accuracy of individual and aggregated data can be evaluated. For example, the actual outcome of a sporting event is eventually known, and so are the prediction outcomes.

We propose to apply the cognitive-model-based approaches of Lee et al. (2014), Yi et al. (2012), and others within psychological research to aggregate manually coded responses obtained through crowdsourcing in ecological research. Our solution addresses the issues of sparse data and lack of ground truth. To illustrate how cognitive modeling can be applied to an ecological dataset and to highlight its merits, we apply this approach to the problem of detecting frog calls in audio field recordings using citizen scientists' coding data collected from two online projects, as well as in simulated data. In many animal detection projects, there are too many recordings to make it practical for an expert to listen to them all, which is why the assistance of citizen scientists is crowdsourced. The audio recordings are split into 30-s clips, which are posted online. Citizen scientists are offered training in the form of reading materials and sample calls. As in most citizen science projects, there is no prerequisite in skill or knowledge for participation, and volunteers are encouraged to learn as they go. A similar case is seen in the Zooniverse platform, where citizen scientists are not required to complete training before participating, although it is advised and offered by project developers. Once ready, citizen scientists can then listen to the uploaded clips and answer questions like "Do you hear Littlejohn's Tree Frog calling in this audio clip?" (we refer to this as "coding" the audio clip in question). In a typical analysis, each clip will be coded independently by a small number (5–10) of citizen scientists, before being hidden from future participants. It is important in citizen scientists to limit the possible barriers of participation for volunteers, as such citizen scientists can freely jump in and out of the project, coding as many or as few clips as they please. Therefore, there is great variability in the number of clips coded, with some people coding just a handful of clips while a few people code dozens or even hundreds.

Within this example, the data aggregation problem lies with the interpretation of different responses given by different citizen scientists for the same audio clip. For example, what should one conclude if amongst ten people who code a particular audio clip, three say "yes", the call is present, while seven say "no", it is not present? In more general terms, how should we identify the group consensus for each audio clip? The simplest approach is to take the majority response. This treats all responses identically, which is almost certainly wrong because some citizen scientists may be much better at identifying frog calls than others. In addition, each citizen scientist is forced to answer "yes" or "no" to each audio clip they code (there is no option to answer "don't know" for example), which poses the question of what strategies they use when they are uncertain if the call is present and resort to guessing. Some people may default to "no", others to "yes", yet others to random mixtures of the two responses. Therefore, a good solution to the data aggregation problem would be to identify the skill level of each citizen scientist, along with their guessing strategy, and to take these factors into account when identifying the group consensus for each audio clip.

Cultural consensus theory

Anthropologists, sociologists, and psychologists studying cross-cultural phenomena have long struggled with closely related problems. In cross-cultural studies, a researcher may ask a group of individuals a series of questions related to their cultural knowledge. Some individuals will know a lot more about the topic than others, and some questions will be much easier to answer than others, but the researcher cannot predict this variation in advance. For example, Weller (1984) asked 24 women from the USA and another 24 women from Guatemala about the nature of diseases and treatments. The research goal was to quantify the women's cultural understanding of disease. Naturally, some respondents knew more about disease than others, and some questions were easier to answer than others.

Cultural consensus theory (CCT) was developed to solve this problem (see: Batchelder and Anders, 2012; Oravecz et al., 2014; Romney et al., 1987). CCT is a mathematical framework which combines separate estimates of item difficulty, *cultural truths* (which, as opposed to "ground truths", represent the beliefs and values of the group), and responders' skills (see also Dawid & Skene, 1979). Modern versions of CCT use a cognitive model to tease apart the effects of these different underlying influences on observed responses. The cognitive model most often used in CCT is a multinomial processing tree, which assumes that the observed responses are the end result of a series of choices. Multinomial processing trees have been used to model memory processes and other psychological phenomena (Batchelder, 2009), as well as in many applications outside psychology, notably in the "random forest" machine learning algorithm of Breiman (2001).

Figure 1 illustrates how a multinomial processing tree can be applied to the problem of identifying frog calls in audio recordings. The tree represents the sequential decision-making process that takes place when a single citizen scientist listens to a single audio clip, and asks themselves the question "Is a particular species of frog calling in this clip?". After listening to the audio clip, one of two outcomes occurs: either the citizen accurately knows the answer (with sufficient certainty), or they do not know the answer (with sufficient certainty). This dichotomous decision is illustrated by the first branching node on the left in Fig. 1, and, ignoring the subscripts and superscripts on the letters for now, the branches carry probabilities d and $1 - d$, for the outcomes of knowing the answer and not knowing the answer, respectively. For some citizen scientists, d will be a high probability because they are highly skilled at the task, but for others it will be low because they are not so skilled. The second dichotomous node (top right) reflects what response the citizen scientist makes, in the cases where they accurately know the answer: they either respond "yes", that is, they know a frog call is present (top right branch of Fig. 1) or "no", that is, they know a call is not present (bottom branch). The probability f and $1 - f$ represent the true probability of a frog call being present on a given audio clip, and not being present, respectively. If the citizen scientist does not know whether a frog is calling (lower initial branch, with probability $1 - d$), they must resort to guessing. The guessing process is shown by the

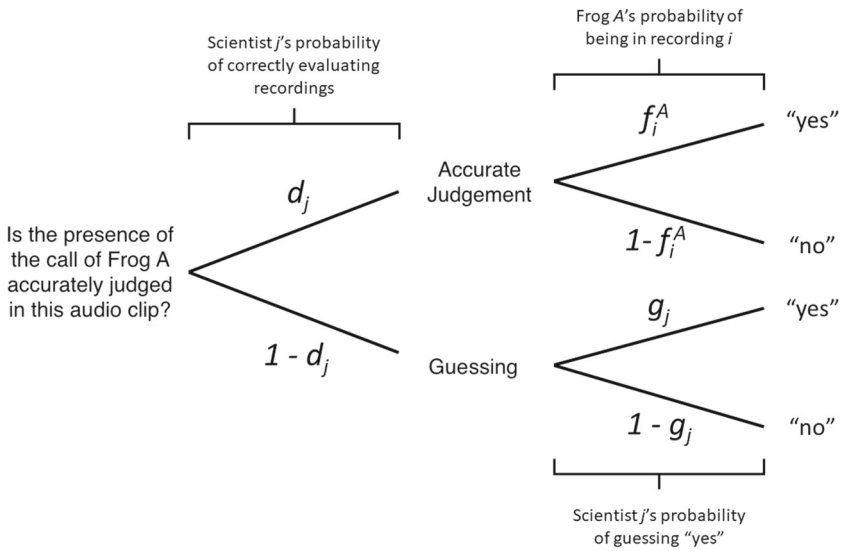

Fig. 1 Multinomial processing tree for the evaluation of element A_{ij}. The probability of each response is the sum of the probabilities of the branches leading to that response. Note that a second, independent tree applies to Frog *B*

third branching node (bottom right) with g being the probability of responding "yes" and $1-g$ being the probability of responding "no".

By encompassing probabilities that can vary both among different citizen scientists (d, g) and among different audio clips (f), this model can help estimate the consensus about which clips are likely to have frog calls (high values of f). The estimation process accounts for the fact that some citizen scientists are not as good as others in detecting frog calls (person-specific parameter d), and also differ in their tendency to guess "yes" vs. "no" (person-specific parameter g) when they do not know. A response of "yes" can be explained under this model in two ways – either because a (competent) scientist accurately identified the call of a frog, or because a (less competent) scientist was unsure about the clip, and guessed "yes". Similarly, "no" responses can be explained in two corresponding ways. These different explanations can be disentangled by estimating the best fitting values of d, g and f for each citizen scientist and each audio clip, and examining how well they predict responses among audio clips within citizens, and how well they predict the responses among citizens within audio clips.

Modern developments for CCT

Karabatsos and Batchelder (2003), Oravecz et al. (2014), and Matzke et al. (2015) all show how CCT models can be estimated in a Bayesian framework. We extend the approach of Oravecz et al. (2014) to allow for hierarchical structure. Where the General Condorcet Model, the model utilized by Oravecz, et al., estimates parameters for individual participants and questions, the current approach also estimates parameters for the distributions of these parameters, and constrains estimates at the individual level. This approach is akin to a standard mixed-effects model using multi-level general linear models (see, e.g., Pinheiro & Bates, 2000). We also extend the modeling approach to match a new feature of the data: each citizen scientist actually answered multiple questions for each audio clip, not one. The questions were about the presence or absence of frog calls related to several different species. In this sense, the current approach can be considered an adaptation of Oravecz, et al.'s model to address the issues present in the current data set, rather than a novel model. Finally, we also outline a simple Markov chain Monte-Carlo (MCMC) scheme for sampling from the marginal posterior distribution over the model parameters.

The superscript "A" on the parameter f^A in Fig. 1 indicates one particular species of frog. This two-step process of responding to a target species was theorized to occur independently for each target species, with a different f parameter at the upper branch of each tree. This is practically helpful because each citizen scientist in our studies answered multiple questions about multiple species, for each audio clip. For an example with two species of frog, we use the superscripts A and B to distinguish between calls from the two frog species. We use the subscript j to distinguish between different citizen scientists, and the subscript i to distinguish between different audio clips. We allow random effects for the competency (d_j) and guessing strategy (g_j) parameters across citizen scientists, and for the probability of frog calls across clips for each species (f_i^A and f_i^B). The latter represent another point of difference from the General Condorcet Model, which estimates an answer key of dichotomous values, rather than a probabilistic estimate. This reflects a difference in the type of question participants are posed in the current study, and the uncertainty of both the ground truth of the data and the cultural truth of the group's beliefs about the data. The use of random effects provides a principled way of aggregating information across people without making all-or-none assumptions about treating people identically or excluding their data (Pinheiro & Bates, 2000). Participant-level parameters d and g were fixed across audio clips and species. Unlike the example presented in Oravecz et al. (2014), we do not estimate question difficulty, instead assuming all questions relating to a specific target species have equivalent difficulty. Consequently, we do not estimate separate parameters for participant ability and item-specific competence as the General Condorcet Model does. Each of the random effects was assumed to follow a beta distribution (across clips or across people). The beta distribution is a suitable choice because it is flexible enough to capture many reasonable shapes for the distributions of random effects, including peaked, skewed, inverted-U, and flat distributions. We used a uniform distribution over the interval (0–10) as the prior for the parameters of the beta distribution, to represent an uninformed prior. We estimated the model from data via a simple Markov chain Monte-Carlo (MCMC) algorithm, in two ways: once, using a simple custom algorithm with block-wise Metropolis steps, and once using the general-purpose sampling software, JAGS (Just Another Gibbs Sampler Plummer, 2003). The mathematical details of the model and the statistical details of the estimation algorithm are provided in the Appendix. Open-source code for the model and associated analyses can be freely downloaded from https://osf.io/wqmkf. The results from the two different samplers were identical, and so we report here the outcomes from the Gibbs sampler (JAGS) which was more computationally efficient and will also likely be more convenient for other users to adapt and extend.

Simulation studies

A key feature of the citizen science data is the large proportion of missing responses. In the present data set, each citizen scientist only coded a small percentage of the available audio

Table 1 Data generating parameters used for the simulation study

Parameter	α	β
f^A	4	1.5
f^B	1.5	4
g	0.2	1
d	0.15	1.5

clips. This feature is ubiquitous in data aggregation for citizen science because it is inherent to the crowdsourcing approach to large scale data tagging. There can be too many audio clips for any one scientist to code, so clips are shared amongst citizens so that each one has a small fraction of the data that is a manageable load. Model parameters are identified less accurately when there are only a few responses for each person or for each clip. This uncertainty is captured in our estimation framework by a wider posterior distribution.

We ran a simulation study to investigate the magnitude of the uncertainty caused by missing data. To generate the data, we sampled random effects for the model parameters (d_j, g_j, f_i^A, and f_i^B) from beta distributions whose parameters were chosen to approximately match those estimated from the data of Study 2; these parameters are shown in Table 1. We sampled random effects for 375 people and 1260 audio clips, to match the second of the two real data sets analyzed below. Using these random effects, we generated synthetic yes/no tagging responses using the multinomial processing tree shown in Fig. 1 for every (simulated) citizen scientist and audio clip. From this complete data set, in which every simulated citizen scientist listened to every audio clip, we generated data sets with 50, 85, 95, and 98.5% missing data. Missing data were produced in a way that emulated the missingness in the real data, in which some people provided many responses and others provided very few. The distribution of the number of audio clips coded by each person in the real data was well approximated by a geometric distribution, truncated below at the minimum number of audio clips per person used in analysis (3), and with about 98.5% missing data. For the simulation studies, we generated the number of audio clips coded by each person from a truncated geometric distribution with the distribution's parameter adjusted to match the required level of missingness (50, 85, 95, or 98.5% across the whole data set).

Figure 2 shows the key results from the simulation study. Each dot represents a single simulated audio clip (top two

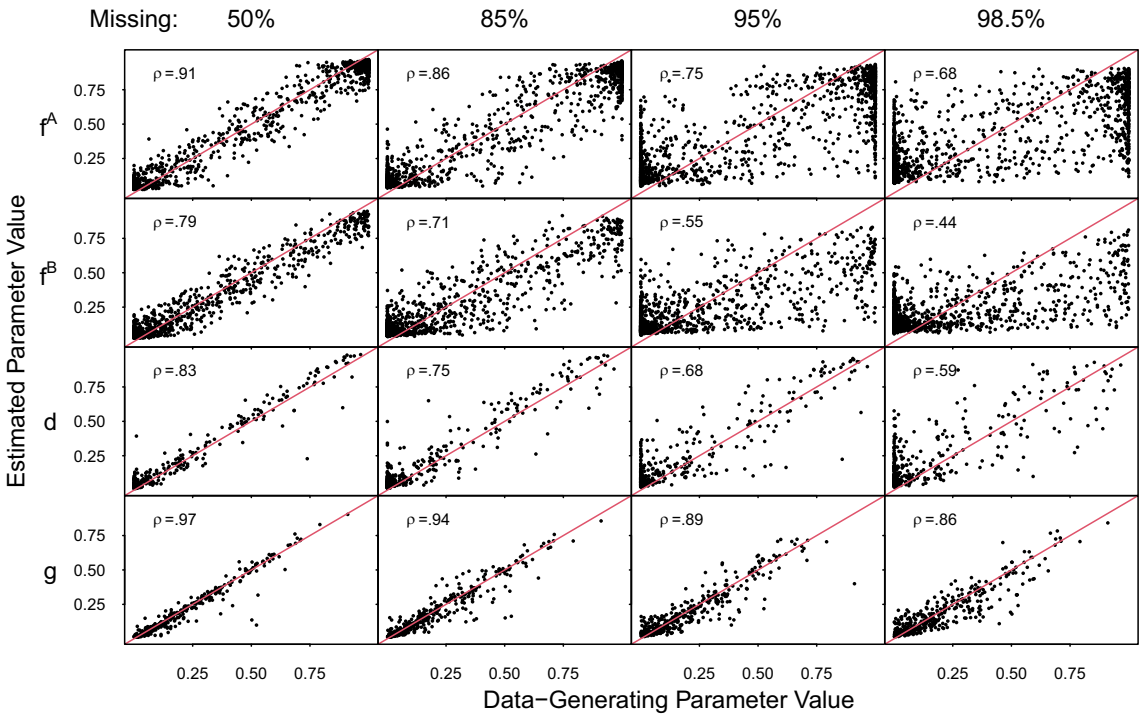

Fig. 2 Parameter recovery studies with 50, 85, 95, and 98.5% of missing data (*columns*). Each panel shows how the random effects for either frog call probabilities in each clip (*top two rows*) or person-specific parameters (*bottom two rows*) are recovered. Each panel is a scatterplot of the data-generating values (x-axis) against estimated parameters, represented as the mean of the posterior distribution (y-axis). All panels show some shrinkage to the mean; the slopes of the scatterplots are shallower than the red diagonal lines. Even in the most difficult cases, the ranks of the random effects are somewhat preserved. This is measured by Spearman rank correlation (ρ) written in the top right corner of each panel

rows) or a single simulated citizen scientist (bottom two rows). The scatterplots illustrate how the parameter values used to generate the simulated data compare with those estimated in the model recovery. Person-specific parameters relating to guessing tendencies (g) and competence (d) were identified more accurately than item-specific parameters (f). This is because there is more information in the data, on average, for each person than each clip because each person answers questions about two different frog species (A, and B). When half of the data are missing (left column, 50%), the parameters frog call probabilities (f^A) are recovered well in rank. That is, the audio clips where the ground truth is a high probability of including a frog call are appropriately identified as such, and conversely for low probability clips. The rank correlation between the ground truth value of f^A and the model-recovered estimate of f^A was $\rho = .91$. This result is of practical importance because it provides assurance that those audio clips identified as highest-ranked in the estimation process are really likely to be those with highest-ranked data-generating random effects as long as there are no more than 50% of missing data. Recovery is a little less good for the frog-specific probability parameters for the second species (f^B, the probability of a call from species B). This is because, to reflect Species B appearing less often in the real data analyzed in Study 2 below, we generated the f^B parameters based on very few "frog-present" clips in the simulated data set.

As we reduced the number of audio clips that each simulated citizen scientist listened to (middle and right columns), the ability of the model to recover the data-generating parameters was compromised. This reduction is to be expected, given the sparsity of data. In the simulation shown in the right-hand column, 98.5% of the data were missing, which approximates the real data set. In this case, the model-based analysis provides some information about which audio clips are likely to have higher vs. lower probability of frog calls (parameters f^A and f^B), but this information is noisy. To make this clearer, we took the simulation with the smallest number of audio clips per scientist (98.5% missing data) and further examined the estimates for the probability of frog calls in each clip. For species A (f^A), this corresponds to the top-right panel of Fig. 2. We applied median splits to the ground truth frog call parameters and to the point estimates of those parameters. Approximately four fifths of the clips that were identified by model estimation as having above-median probability of a call were in fact those that did have an above-median probability of a frog call (i.e., the true-positive rate was 79%). Similarly, about four fifths of the clips that were identified by the model estimation as having below-median call probability were in fact those that did have a below-median probability of a frog call in the data generating parameters (i.e., the true-negative rate was 80%). The results for species B (f^B) were less clear, due to the difference in the parameters used to generate data (Table 1). For species B, the true-positive rate was 65% and the true negative rate was 63%. Thus, even in this very sparse data set, the model provides practically useful information to filter out those audio clips which are more likely to have a frog call than those which are not.

Like all parameter recovery simulations, ours used the same assumptions to generate the simulated data as used in the model applied to the data. An interesting consequence of this is that the presence of a frog call in any audio clip is represented in a probabilistic manner, by parameters f^A and f^B, which can take on any values between 0 and 1. An alternative assumption would be to have each simulated audio clip coded as either containing a frog call ($f = 1$) or not ($f = 0$). While that assumption aligns with intuitive ideas of frog pond occupancy (frogs either call, or do not call) we adopted the graded assumption because it better aligns with the psychological decision task faced by citizen scientists with real data. In real audio clips, frog calls can be very clear, or they can be buried amongst background noise, or be atypical calls, or come from a great distance from the recorder, or be very weak for any number of other reasons. Our simulation assumptions capture the graded nature of frog calls. It is an open question whether the simpler, binary assumption could lead to more robust or efficient estimation methods, which we leave to future studies.

Applications to real data

Study 1: Listening for many species

Study 1: Methods

We applied our analysis method based on cultural consensus theory to an ecological study of endangered frog species. We used acoustic data recorded from data loggers deployed to Ingar Pool, in the Blue Mountains National Park of New South Wales, Australia. Ingar Pool is a potential habitat for two threatened frog species, the Northern Heath Tree Frog (also called Littlejohn's Tree Frog) and the Giant Burrowing Frog. Data loggers were set to record 5 minutes of ambient noise in every hour between 6:00 pm and 2:00 am, as this was the expected daily activity period of the target frogs. In total, 10.5 h of audio were used for analysis. The audio recordings were split into 1260 clips each 30 s in duration, and these were posted on the citizen science platform "Zooniverse" (http://zooniverse.org/ Simpson et al. 2014). Zooniverse is a long-standing and successful platform for citizen science, with more than two million registered users who have contributed to hundreds of scientific publications. For example, the "Penguin Watch" project on Zooniverse had users count penguins in camera photos from Antarctica (Jones et al., 2018), while

Table 2 Target frog species with scientific names and abbreviations

Common name	Scientific name	Abbreviation
Tree frog species		
Blue Mountains Tree Frog	*Litoria citropa*	BMTF
Northern Heath Frog	*Litoria littlejohni*	LLJ
Peron's Tree Frog	*Litoria peronii*	PTF
Green Stream Frog	*Litoria phyllochroa*	GSF
Whistling Tree Frog	*Litoria verreauxii*	WTF
Ground Frog Species		
Common Eastern Froglet	*Crinia signifera*	CEF
Eastern Banjo Frog	*Limnodynastes dumerilii*	EBF
Giant Burrowing Frog	*Heleioporus australiacus*	GBF
Striped Marsh Frog	*Limnodynastes peronii*	SMF

other projects have investigated areas as diverse as botany, astronomy, and history, amongst others.

Citizen scientists listened to the clips online and answered questions about which frog species they heard. Participants had the option of completing a training module and quiz about the frog calls, at any time. The study was divided in two sub-projects, with citizen scientists in one sub-project identifying calls from five tree frog species and in the other section calls from four ground frog species. Table 2 gives the frog species' common names, scientific names, and abbreviations. For each audio clip the citizen scientist was asked to indicate which species of either ground frog or tree frog they heard call. There was also an option to indicate that no target species frog calls were present, which we used to check for internal consistency in the responses.

Separate from the crowdsourced data coding, one of the authors (Oliver Kelly) coded all 1260 audio clips and provided expert decisions about frog calls for all species. To assess the reliability of the expert coding, a second expert (author Alex Callen) coded a subset ($n = 20$) of the audio clips with 98.7% similarity (74/75 possible classifications). Expert decisions indicated that four frog species called in a large proportion of recordings (SMF 94%; CEF 93%; EBF 71%; PTF 62%), three species were rare (GBF 10%; BMTF 3.6%; GSF 0.7%), and two species were absent from all recordings (WTF and LLJ).

Study 1: Results

A total of 309 citizen scientists participated in tagging the tree frog species and 358 in coding the ground frog species (only 23 people participated in both). We excluded data from 35 people who coded for tree frog species and 52 people who coded for ground frog species because these people did not meet our criterion for internally consistent responding. They used the "no frog calls" option inconsistently in more than 25% of the audio clips they rated. That is, they indicated that one or more target species was present, while simultaneously selecting the option "no frog calls"; alternatively, they failed to identify any species as present, but did not select the option "no frog calls". This was inconsistent as they were instructed to select the "no frog calls" option if no target species were present, even if the calls of other species were present. For the remaining participants, the median number of audio clips coded per person was 4 or 5, for the tree and ground frog species, respectively. The corresponding means were 17.4 and 18.7, indicating substantial positive skew in the relative contributions of different citizen scientists.

We conducted an exploratory correlation analysis between participants' accuracy (defined here as agreement with the expert coder) and the number of clips evaluated. If there was a relationship between engagement and task competence, this analysis would provide us with *a priori* information about which participants contributed better-quality data. However, no such relationship was found, and a Bayesian correlation found moderate evidence against a relationship, $BF_{10} = 0.18$.

We estimated the CCT model from the data using the sampling algorithm described in the Appendix. We ran three independent chains in parallel, thinned by keeping samples from only every fifth iteration, and discarded 20,000 samples (100,000 iterations) for burn-in. We kept 5000 posterior samples. Visual inspection of the Markov chains corresponding to group-level parameters showed good convergence and mixing, confirmed by Gelman diag ($\hat{R}$) values between 1.00 and 1.04 for all group-level parameters. Inspection of some of the chains corresponding to random effects showed good convergence, and while some had less efficient mixing, they were still adequate.

Figure 3 shows the estimated person-specific and audio-clip-specific random effect model parameters. The darker gray histograms show distributions of the f for each of the nine species. These are the parameters that estimate the consensus for each audio clip about whether the given frog species is calling in any given audio clip. There are clear differences between species. For example, most audio clips had high estimates (near $f = 0.8$) for species SMF, indicating that this species is common, and the group consensus is that calls were audible in most clips. Other species were estimated as very rarely calling, such as GBF and LLJ, for which most audio clips had $f < 0.2$. Overall, species SMF, BMTF, and CEF were identified as calling very frequently. Species PTF, EBF, and WTF were identified as calling moderately frequently. Species LLJ, GSF, and GBF were identified as calling very rarely or never. Broadly, these outcomes agree well with the expert ratings of the data described above. The expert identified species SMF and CEF as very common, species EBF and PTF as quite common, species GBF, BMTF and GSF as rare, and species WTF and LLJ as absent. It seems that the citizen scientists' consensus, as identified

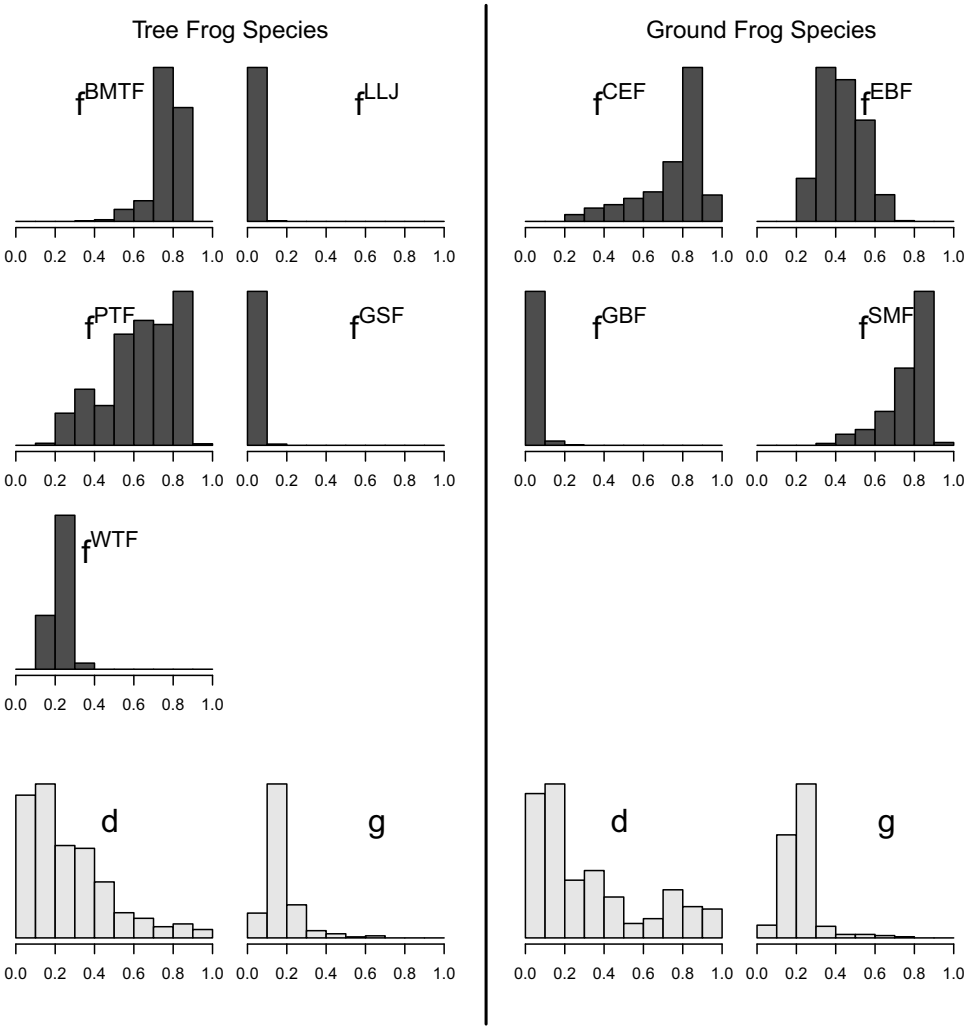

Fig. 3 Estimated distributions for random effects of the model parameters in Study 1. The *left side* of the figure shows outcomes for the tree frog species and the *right side* for the ground frog species. The *top nine panels* show the distribution of frog call probability parameters across audio clips for different species. The *bottom row of four panels* show the distribution of competence (d) and guessing bias (g) parameters across citizen scientists in each sub-project. In each panel, histograms show the distribution of point estimates (means of the posterior distributions) across people or audio clips

by the model, was accurate except for the over identification of species BMTF and WTF. This result may be due to citizen scientists confusing their calls with highly frequent co-occurring species. For example, species BMTF and PTF both employ grumbles in their mating calls, which may have led to the over identification of BMTF and under identification of PTF. As for species WTF, the most similar call (in this study) would be that of species CEF. Both species call in quick and repetitive pulses.

The bottom row of four panels in Fig. 3 shows the subject-level parameters. The parameters d measure the estimated proficiency of different citizen scientists in identifying frog calls, and parameters g show the bias towards guessing "no" (g near zero) or "yes" (g near one) when the citizen scientist is uncertain of the truth. The guessing parameters were mostly small, indicating that citizen scientists default towards answering "no" when asked about recordings they are unsure of. The d parameters were distributed unevenly across people, with a few citizen scientists being mostly consistent with the group consensus ($d > 0.5$) and most citizens being quite inconsistent with the consensus ($d < 0.2$). The estimated d parameters were quite consistent between groups who coded for tree frog species (mean $d = 0.27$) and ground frog species (mean $d = 0.33$).

We used the expert coding to support a more fine-grained analysis and compare which audio clips were identified as likely to include a call by citizen scientists and by the expert. This comparison is complicated by the need to choose a

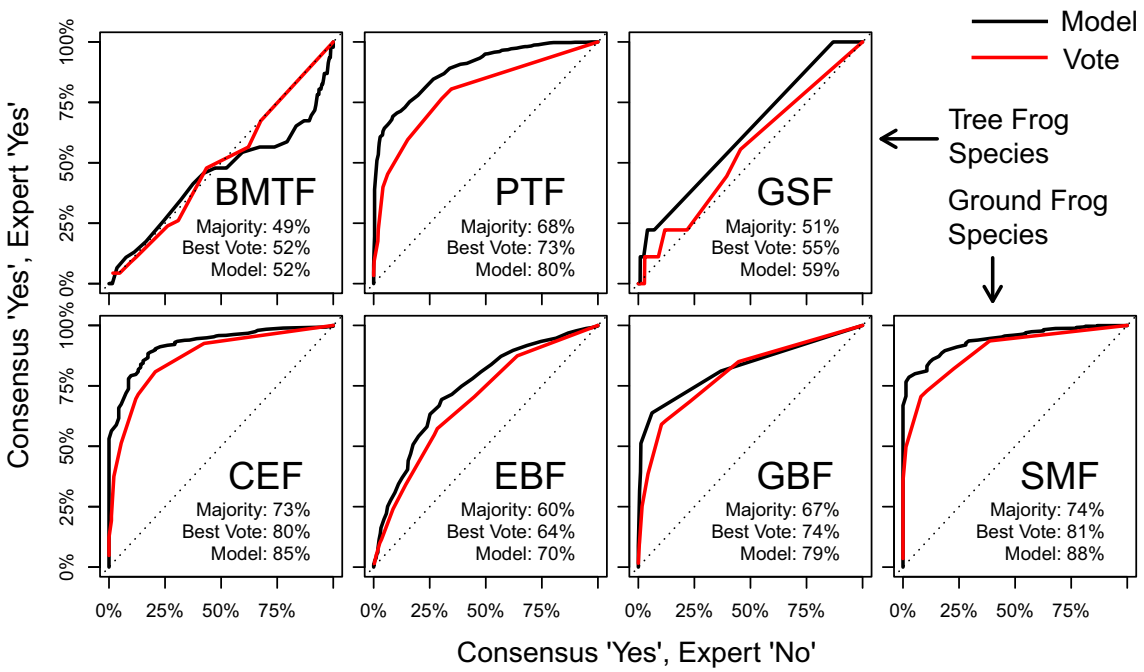

Fig. 4 Receiver operating characteristic (ROC) for agreement between model output and expert coding. The *black curves* shows how the proportion of "hits" (*y*-axes) and "false alarms" (*x*-axes) change with different cutoff thresholds used to transform frog call probabilities into yes/no outcomes. "Hits" are audio clips for which the model output and the expert both agree that a frog call is present; "False alarms" are clips where the model output suggests that a frog call is present, but the expert says that is wrong. *Red curves* represent model-free ROC analyses based on simple vote-counting. The *top row* shows tree frog species, and the *bottom row* shows ground frog species. In each panel, statistics show the overall accuracy (across hits and correct rejections) for three methods of analyzing the data: using simply majority voting; using vote counting but with the post hoc calculated best possible threshold for identifying a "hit"; and using the model, but with the post hoc calculated best possible threshold for identifying a "hit". Note that no ROC can be calculated for species LLJ and WTF because there were no audio recordings that included calls from those species, according to expert coding

threshold for what is meant by "likely" in the model outputs. For example, we may define a cutoff for the mean of the posterior distribution over f, say $f = 0.9$, meaning that we will identify an audio clip as likely to contain a call from a particular frog species if the estimated mean posterior probability of a frog calling in that clip is $f > 0.9$, but we could equally choose $f > 0.5$ or any other value. Figure 4 shows receiver operating characteristic (ROC) curves for each species[2]. These show how agreement between consensus and expert changes for different cutoffs by comparing the rate of true positives (*y*-axis in Fig. 4) to the rate of false positives (*x*-axis). As the threshold for a consensus "yes" response moves from extremely strict (bottom-left) to extremely permissive (top-right), the rate of both true and false positives increase. The diagonal dotted line on each plot represent a one-to-one relationship in this increase, indicating no greater sensitivity than random noise. ROC curves which are above the diagonal dotted lines indicate a better-than-noise relationship between the consensus and expert responses, with higher curves on the *y*-axis indicating better agreement. The ROC curves show that citizen scientists' consensus agreed well with the expert coding for some species, particularly CEF, SMF, PTF, and GBF. Agreement was moderate for species EBF and GSF[3]. For the BMTF, there was no agreement at all: the ROC even dips below diagonal line, indicating that for some thresholds, the consensus of the citizen scientists was negatively correlated with the expert coding.

To evaluate the performance of the model against a model-free benchmark, we compared the model-based ROCs in Fig. 4 against ROCs calculated by simple vote-counting procedure (shown by the red curves in Fig. 4). For the vote-counting ROCs, we calculated the mean proportion of "yes" votes for each audio clip and classified the clip as containing the corresponding frog call if this mean proportion exceeded some threshold. Using a threshold of 0.5 corresponds to the common data aggregation approach of simply choosing the modal response for each audio clip. The red curves are produced by making this calculation for all thresholds from zero

[2] ROC curves could not be calculated for species LLJ and WTF because there were no audio clips which the expert identified as including those calls

[3] The ROC curve has "steps" for GSF because the expert identified GSF calls in only nine out of 1260 audio clips, leading to low resolution on the *y*-axis.

to one, in increments of 0.01. For all species except BMTF, the black curves are above the red curves, indicating that the model-identified consensus is more closely aligned with the expert coding than is the simple vote-counting consensus. For some species, the benefit is substantial, as shown by the overall accuracy values reported in each panel (e.g., for species SMF, majority vote yields 74% accuracy, the best possible vote-counting method yields 81% accuracy, and the model-identified consensus yields 88% accuracy). The difference between the vote-counting and model-based approaches for this species are due to the CCT model down-weighting votes from citizen scientists who were identified as weaker at this task, and up-weighting votes from citizen scientists identified as stronger. For the species BMTF and GSF, both methods failed, which indicates that the citizen scientists in this sample simply could not accurately detect those calls.

An example of how the model-based analysis outputs might be used in practice can help make potential benefits clearer. Suppose an ecology researcher wants to find evidence of "occupancy" for species GBF in the audio clips – that is, to positively identify the call of a GBF individual in an audio recording. Without the benefit of citizen science, an expert listener would have to repeatedly listen to randomly chosen audio clips, and stop when they identified the call of the GBF. Since our expert data revealed that GBF calls occur in 10.1% of audio recordings, that approach would require on average that the ecologist listened to around ten clips before they randomly hit on a clip that contains a GBF call. On the other hand, if the expert ecologist had the benefit of the citizen science data and our CCT analysis, they could select clips to listen in a systematic, non-random manner. The ecologist could begin with the audio clip judged most likely (by CCT posterior estimates) to contain a GBF call, then proceed to the second most likely clip, and so on. By this method, they would listen to just one clip – the very first audio clip contained a GBF call according to our expert rating. This was true for all the ground frog species where calls were present in the data: in each case, the audio clip rated by CCT analysis as most likely to contain a call of that species *did* in fact have a call of that species. For tree frog species where calls were present, the situation was not quite as ideal, but there was still an advantage in screening clips according to the CCT results compared with randomly. Random clip selection would take on average 27 clips to find a BMTF call, but CCT-guided clip selection would reduce this slightly, to 24 clips. Species PTF was very common, and so random clip selection would take on average just two audio clips identify a call, but CCT-guided clip selection would find a PTF call on the very first clip. There was a very large advantage for the model with species GSF, which was rare: random clip selection would require, on average, listening to 140 clips, but CCT-guided clip selection would find a GSF call after listening to just 11 clips.

Study 2: Listening for fewer target species

In Study 1, citizen scientists had to listen for calls of four or five species of frogs simultaneously, depending on whether they were listening for ground or tree frog species, respectively. We set the crowdsourced data analysis up in this way in order to ensure that citizens searched for all frog species that might be present at Ingar Pool.

In reality, in the region of interest there is one species of tree frog and one species of ground frog which are under severe ecological threat. This increases the scientific value of finding calls from those two species, and at the same time makes those calls harder to find – because individual frogs of those species are unlikely to be recorded calling in the audio data. Identifying the calls of rare frog species is not just important, it can also be difficult, even for experts. Many Australian frogs chorus concurrently during the breeding season, and identifying the calls of threatened species amongst a cacophony of calls is a bioacoustic challenge for experts, citizen scientists, and AI alike. We focused on this challenge in Study 2.

From a psychological design perspective, there are positive and negative aspects to the design we used in Study 1. On the positive side, each citizen scientist contributed quite a lot of data in Study 1 because they gave decisions about four or five frog species on each audio clip. This amount of data makes estimation of the CCT model more efficient and more stable because the model has greater opportunity to learn person-level parameters (d and g). On the negative side, human information processing capacity has a notoriously small limit, and it is possible that the number of different target species' audio calls that a human listener can keep in mind is less than four. For example, auditory versions of the n-back working memory task (Monk et al., 2011) find substantial loss of performance over a span of 2–4 items. Citizen scientists on Zooniverse could avoid these limitations by repeatedly listening to the same audio clip, each time keeping in memory just one or two target frog call species, but such investment of time is not likely for all participants.

We investigated these issues in Study 2 by limiting the target questions to just the two threatened species (LLJ and GBF). That is, for each audio clip, citizen scientists were asked whether or not they heard calls from species LLJ and whether or not they heard calls from species GBF. They were asked no other questions, including the "no frog calls" question. This was dropped because of the lighter working memory load.

Study 2: Methods

Study 2 used the same 1260 audio clips as Study 1. These were posted on Zooniverse in a different project. For each

clip, citizen scientists were asked only whether the clip included calls from species LLJ and GBF.

Study 2: Results

A total of 587 people participated in Study 2. We removed data from people who provided ratings for only one or two audio clips, on the basis that these people did not sufficiently engage with the project. For the remaining 375 people, the median number of clips rated per person was 7, and the mean was 18.3. We estimated the same model as for Study 1, but with only two f parameters, for the two species. We used the same estimation procedure as for Study 1. The raw responses from the citizen scientists were different in Study 2 as compared with questions about the same target species from Study 1. Most noticeably, in Study 2, citizen scientists exhibited a greater tendency to respond "yes" compared with Study 1. This was most pronounced for species GBF, with almost twice as many "yes" responses in Study 2 (mean 27% "yes") than Study 1 (mean 15% "yes").

Figure 5 shows the estimated parameters of the model using a similar format to Fig. 3. The bottom row of the figure shows person-specific parameters related to competence, d, and guessing bias, g. The top row shows audio-clip-specific parameters for the probability of frog calls in the audio clips, for the two frog species (f^{GBF} and f^{LLJ}). As for Study 1, the person-level parameters show that most people were not very competent at identifying the group consensus (low values of d), and that the default guessing response was typically "no" (low values of g). The model identified a small subset of citizen scientists whose responses were closer to the consensus, with $d > 0.5$. In the framework of cultural consensus theory, these citizen scientists were quite reliable at identifying the consensus response for each audio clip that they coded.

The distributions of estimated f^{LLJ} parameters across audio clips was much the same for Study 2 as for Study 1. In both data sets, the model identified a consensus response of "no" for almost all audio clips (values of f^{LLJ} near zero). On the other hand, the distribution of f^{GBF} parameters across audio clips was noticeably different in Study 2 than Study 1, with higher values of f^{GBF} estimated, likely driven by the higher proportion of "yes" responses for this species observed in the data for Study 2.

Once again, we compared the estimated model-based consensus against expert coding using ROC analysis (just for species GBF, as expert coding did not identify any calls from LLJ). Figure 6 plots the ROC using the same format as Fig. 4. Comparing the two figures, it is apparent that the model-based analysis provided benefit over the simpler model-free aggregation methods in Study 1, but not in Study 2. In Study 2, model-based data aggregation based on CCT did not improve agreement with expert ratings, although it did also not hurt (73% accuracy overall for the model-based method, 74% accuracy overall for majority voting). This likely reflects

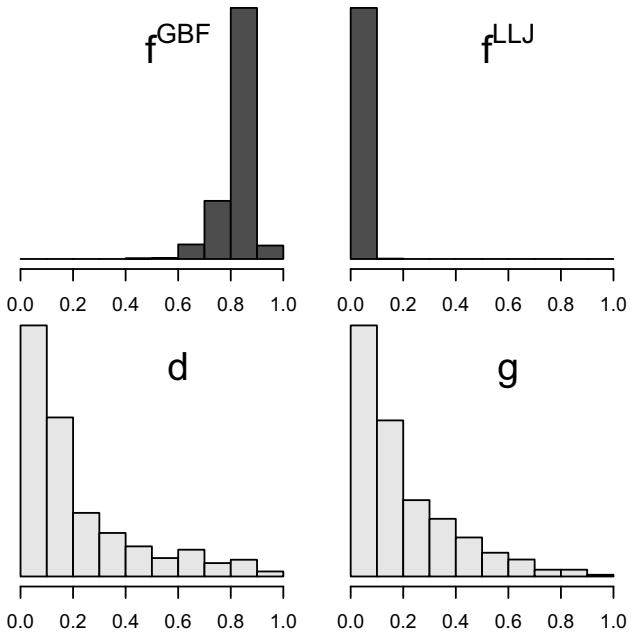

Fig. 5 Estimated distributions for random effects of the model parameters in Study 2. The *top two panels* show the distribution of audio-clip-specific frog call probability parameters (f^{GBF} and f^{LLJ}) across audio clips. The *bottom two panels* show the distribution of person-specific competence (d) and guess bias (g) parameters across citizen scientists. In each panel, the histograms show the distribution of point estimates (means of the posterior distributions) across people or audio clips

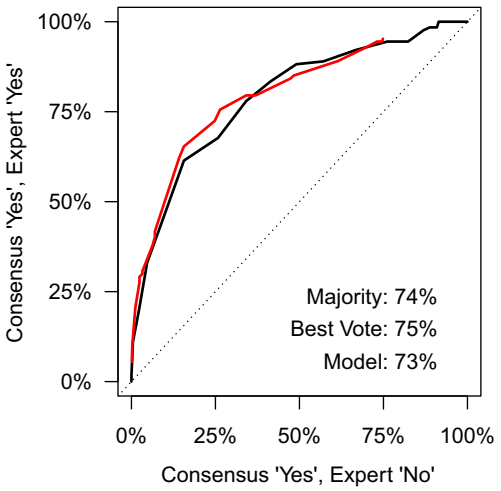

Fig. 6 Receiver operating characteristic (ROC) for agreement between consensus model output and expert coding of GBF calls. The *black curve* shows how the proportion of hits and false alarms change with different cutoff thresholds. "Hits" are audio clips for which the citizen scientist consensus and the expert both agree that a frog call is present; "False alarms" are clips where the citizen scientist consensus is that a frog call is present, but the expert says that is wrong

the reduced opportunity for the model to learn person-specific information in Study 2 than in Study 1. With less information about each citizen scientist, the model is less able to leverage differences between citizen scientists in identifying consensus. This conclusion is further supported by an analysis corresponding to that reported for Study 1, estimating how many audio clips an ecologist would have to listen to in order to find a call from species GBF. If the clips were selected randomly, it would take on average ten clips before one contained a GBF call, but if the clips were selected in order of model-estimated probability, it would take slightly fewer: seven clips.

Discussion

Crowdsourcing provides a promising avenue for engaging citizen scientists' help with large-scale data coding. This approach can make it possible to code larger and richer data sets than ever before, and more quickly, by sharing the work across a large group of volunteers. This poses the problem of how to aggregate different responses from different citizen scientists. For example, Jones et al. (2018) used crowd sourcing on the Zooniverse platform (http://zooniverse.org/ to count and locate penguins in more than 70,000 images. For each image, different citizen scientists pinpointed different locations for the penguins, and sometimes different numbers of penguins in total (e.g., if one penguin was only partially visible, or difficult to spot). These differences between citizen scientists can be very hard to resolve because the ground truth is unknown. These problems are compounded by psychological differences between citizen scientists in their skill, their motivation, and what response strategy they use when they are uncertain.

We have used a data aggregation approach based on quantitative cognitive modeling using cultural consensus theory (CCT). This approach directly models some psychological aspects of the data analysis process, as well as differences between citizen scientists. We developed a hierarchical Bayesian estimation framework and demonstrated via simulation that the model provided usable information even in the presence of substantial missing data.

We demonstrated the cognitive modeling approach in an ecological coding application, using http://zooniverse.org, 375 citizen scientists who listened for the presence of calls from two target frog species in 1260 audio recordings. Each audio clip was coded by several citizen scientists, and these codings were aggregated by model estimation. Comparison with expert coding of the same audio clips demonstrated that the model-based aggregation of the citizen scientists' responses agreed with the expert's responses when the optimal threshold for a "yes" response was applied, with the model identifying the presence of calls from one frog species and the absence of calls from the other. The information in the data from the citizen scientists was very scarce; for example, many citizen scientists *never* gave a "yes" response to any question, and more than one third of them coded fewer than five audio clips. Despite this scarcity, the model-based data aggregation approach provided useful indications about the performance of individual citizen scientists and which audio clips were most likely to include frog calls. This information is limited in its accuracy due to the limitations of the citizen scientists who volunteered, and the inherent difficulty of identifying different frog calls, but it is nevertheless useful. For example, it can be used to identify a small subset of audio clips that should be coded by more intensive or expensive means, such as by expert raters. For example, a common research goal in ecological work is to identify whether a species occurs in a location ("occupancy"). This is equivalent to asking whether the calls of this species occur in a given location. For an expert rater, inferring occupancy requires coding audio clips until one (or a few) calls are identified, or alternatively to check all clips to confirm that the species is never heard. The model-based analysis has potential to increase the efficiency of this work by accurately sorting clips according to their likelihood of including the relevant call, potentially reducing the number of clips that an expert has to code.

We hope that the approach we have presented here may be useful in a very broad range of citizen science projects because the data aggregation problem is common to almost all crowdsourced data analysis efforts. Computational models based on cultural consensus theory naturally address this problem whenever the answers that the citizen scientists provide can be framed as choosing between a relatively small set of category labels. These labels were "yes"/"no" in our case, but could well be more complex such as ratings on a scale ("low", "medium", "high"), as in the Latent Truth Rater Model (Anders & Batchelder, 2015), or category identifications ("target", "non-target", "don't know") as in the variable-response model (Aßfalg, 2018). There is potential to apply these models to crowdsourced data that better suit these response methods. Future extensions of the work could also involve tracking citizen scientists across projects. For example, some citizen scientists in our data set also contributed to other related projects, such as coding audio clips from different recording experiments for the same frog species. A limitation in our application was that the small amount of data available for each citizen scientist made it difficult to estimate person-specific variables precisely. This limitation could be eased by remembering users across multiple projects, and the Bayesian estimation framework provides a natural way to include such data. For example, competence in responding to one species may generalize to the target species of other studies, assuming a single participant-level competence parameter d can be used. Future work could also aim to

improve the computational speed of the algorithm. Analyses for the current data set required several hours on a personal computer. This speed could become a problem for the analysis of much larger projects, where the number of citizen scientists and the number of audio clips are both orders of magnitude larger.

An important limitation of the citizen science approaches we have addressed is that crowd-sourced coding identifies *consensus*, not necessarily *truth*. Our analyses have demonstrated that a cognitive-model-based approach, using CCT can help to more efficiently aggregate data and identify consensus, but of course this cannot change the fact that the group consensus may be different from truth. This difference can be important in some applications, such as when the crowd-sourced consensus is used to guide limited resources to subsets of the data. For example, if the crowd-sourced consensus was used to screen audio clips so that a subset of them could be analyzed by an expert, it is possible that an incorrect group consensus could mis-lead the expert. This is not a limitation of our CCT-based approach, or of any data aggregation approach in particular, but is a limitation of using crowd sourced coding to guide resource-limited expert coding.

A less fundamental limitation of our approach is that the statistical modeling assumes stationarity. The parameters of each person are assumed to be constant as they code different audio clips, and the parameters of each audio clip are assumed to be constant across people. It is especially plausible that person-level parameters might change across audio clips, for example the competence parameter (d) might increase as a person learns more about the task and frog calls, or it might decrease with fatigue. Like all other existing methods of aggregating citizen scientists' data, our CCT approach treats the data as independent and identically distributed, conditional on the fixed parameters. In ongoing work, we are investigating the effects of this limitation using more tightly-controlled data collection and more relaxed modeling assumptions.

Conclusion

Citizen science methods have the potential to fill an important gap in research where data coding is time-intensive. This gap will become increasingly important as technological change leads to ever greater data collection capacity which rapidly exceeds the capacity of experts to code the data in practical time. Our analyses demonstrate that model-based analysis using cultural consensus theory provides a practical method for aggregating coding responses across citizen scientists while taking into account the different abilities of the citizens. Comparison with a simpler model-free aggregation method (vote counting) showed that the extra complexity of the model-based approach does not incur a cost in the accuracy of results for easy-to-classify problems, and can provide substantial improvements in difficult-to-classify problems. Future research could improve results further by considering the individual characteristics of the citizen scientists, such as their training and performance histories and psychological measurements. The model-based aggregation approach provides a psychologically-motivated and coherent framework in which to include those variables, opening up new possibilities for improving the efficiency and accuracy of citizen science by better matching individuals to tasks.

Funding This research was supported by the NSW government through a partnership between the Saving our Species program and the Environmental Trust.

Open Practices Statement The data and materials for the analyses presented in this article are available and may be freely accessed on OSF from https://osf.io/wqmkf.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Appendix

Suppose A and B are sparse matrices of responses from citizen scientists about the presence or absence of calls from frog species GBF and LLJ, respectively. The rows index different audio periods, $i = 1 \ldots P$, and the columns index different citizens, $j = 1 \ldots S$. Entries are coded so that element A_{ij} is: empty if citizen j did not rate audio clip i; zero if citizen j rated GBF as absent in clip i; and one if citizen j rated GBF as present in clip i. The structure is the same for matrix B, with respect to species LLJ.

The multinomial processing tree model defines the probability of responding "present" or "absent" for each citizen and for each clip, conditional on parameters which set the probabilities of the tree branches as depicted in Fig. 1.

The j^{th} citizen scientist is characterized by two parameters. One is for their probability of correctly identifying frog presence vs. absence, d_j, and the other is for their probability of responding "present" as a guess in those cases where correct identification does not happen, g_j. Each audio clip also has two parameters, representing the probabilities of presence of a call from each species: f_i^A and f_i^B. Note that this model assumes that scientists are equally good (or bad) at identifying the presence vs. absence of calls, and also are equally competent across the two frog species. The guessing process is also equal across frog species. Conditional on these parameters, the probability of a "present" response for

scientist j rating audio clip i for the presence of frog species A is given by Eq. 1.

$$p\left(A_{ij}=1|f_i^A,d_j,g_j\right)=f_i^A d_j+g_j\left(1-d_j\right) \tag{1}$$

The corresponding probability for species B is given by the same equation but with f_i^B. The probabilities of responding "absent" are just one minus Eq. 1.

We make the simplifying assumptions that, conditional on the parameters, the responses are independent for each audio clip and each citizen scientist. This is probably reasonable for scientists, but may not be for audio clips (e.g., clips from nearby in time may have positive correlation, if the frogs persist in one location for some time). The simplifying assumptions mean that the probability of the observed response matrices, conditional on the parameter vectors, just the product of the individual values:

$$\begin{aligned}
&p\left(A,B|f^A,f^B,d,g\right)\\
&=\prod_{i=1}^{P}\prod_{j=1}^{S}(p\left(A_{ij}=1|f_i^A,d_j,g_j\right)A_{ij}\\
&\quad+\left(1-p\left(A_{ij}=1|f_i^A,d_j,g_j\right)\left(1-A_{ij}\right)\right))\\
&\quad\times(p\left(B_{ij}=1|f_i^B,d_j,g_j\right)B_{ij}\\
&\quad+\left(1-p\left(B_{ij}=1|f_i^B,d_j,g_j\right)\left(1-B_{ij}\right)\right))
\end{aligned} \tag{2}$$

We define the probability in Eq. 2 to be 1 for missing values of A_{ij} or B_{ij}. This is clearer to read than writing more formal versions using indicator functions.

We impose a hierarchical structure on the model to constrain the individual person and audio clip parameters:

$$f_i^A\sim\beta\left(\alpha_A,\omega_A\right) \tag{3}$$
$$f_i^B\sim\beta\left(\alpha_B,\omega_B\right) \tag{4}$$
$$d_j\sim\beta\left(\alpha_d,\omega_d\right) \tag{5}$$
$$g_j\sim\beta\left(\alpha_g,\omega_g\right) \tag{6}$$

Here, $\beta(x,y)$ represents a beta distribution with shape parameters x and y. For more compact notation we write α and ω for the vectors $\left(\alpha_A,\alpha_B,\alpha_d,\alpha_g\right)$ and $\left(\omega_A,\omega_B,\omega_d,\omega_g\right)$. We put a relatively uninformed prior on the α and ω parameters of the group-level distributions; uniform priors over the interval (0–10). From Bayes theorem, the posterior distribution over parameters and random effects is:

$$\begin{aligned}
p\left(f^A,f^B,d,g,\alpha,\omega|A,B\right)\propto\ &p\left(A,B|f^A,f^B,d,g\right)\\
&\times p\left(f^A|\alpha_A,\omega_A\right)\\
&\times p\left(f^B|\alpha_B,\omega_B\right)\\
&\times p(d|\alpha_d,\omega_d)\,p\left(g|\alpha_g,\omega_g\right)\\
&\times p\left(\alpha\right)p\left(\omega\right)
\end{aligned} \tag{7}$$

Estimation using Markov chain Monte Carlo (MCMC) can be accomplished easily using JAGS (Plummer, 2003) or a simple Metropolis scheme. For Metropolis, we need to repeatedly compute the ratio between Eq. 7 when calculated with two different sets of parameters. For the multinomial processing tree parameters, f^A, f^B, d, and g, we exploit the factorial structure of Eq. 2. This allows proposals for the vectors f^A and f^B to be accepted/rejected element-wise, for each audio clip, and proposals for d and g to be accepted/rejected element-wise, for each citizen scientist. For example, a new proposal for element i of f^A, say $\hat{f}_i^A$ has the Metropolis-Hastings acceptance probability given by:

$$\begin{aligned}
&\frac{p\left(f^A,f^B,d,g,\alpha,\omega|A,B\right)}{p\left(f^A,f^B,d,g,\alpha,\omega|A,B\right)}\\
&=\frac{\prod_{j=1}^{S}(p\left(A_{ij}=1|\hat{f}_i^A,d_j,g_j\right)A_{ij}+\left(1-p\left(A_{ij}=1|\hat{f}_i^A,d_j,g_j\right)\left(1-A_{ij}\right)\right))}{\prod_{j=1}^{S}(p\left(A_{ij}=1|f_i^A,d_j,g_j\right)A_{ij}+\left(1-p\left(A_{ij}=1|f_i^A,d_j,g_j\right)\left(1-A_{ij}\right)\right))}\\
&\times\frac{p\left(\hat{f}_i^A|\alpha_A,\omega_A\right)}{p\left(f_i^A|\alpha_A,\omega_A\right)}
\end{aligned}$$

Note the products in this ratio only run over citizen scientists, $j=1\ldots S$, for the single audio clip relevant to the proposal, j. Corresponding simplifications occur for the person-wise parameters d and g. Computationally, this means the row and column products are the key elements in the sampling scheme.

References

Altrudi, S. (2021). Connecting to nature through tech? The case of the inaturalist app. *Convergence, 27*(1), 124–141.

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika, 80*. https://doi.org/10.1007/s11336-013-9382-9

Aßfalg, A. (2018). Consensus theory for mixed response formats. *Journal of Mathematical Psychology, 86*, 51–63. https://doi.org/10.1016/j.jmp.2018.08.005. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022249617302110

Bakker, V. J., Baum, J. K., Brodie, J. F., Salomon, A. K., Dickson, B. G., Gibbs, H. K., ... McIntyre, P. B. (2010). The changing landscape of conservation science funding in the united states. *Conservation Letters, 3*(6), 435–444.

Batchelder, W. H. (2009). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring Psychological Constructs: Advances in Model Based Measurement.* American Psychological Association Books.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology, 56*(5), 316–332.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., ... Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation, 173*, 144–154.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Brown, G., McAlpine, C., Rhodes, J., Lunney, D., Goldingay, R., Fielding, K., . . . Vass, L. (2018). Assessing the validity of crowdsourced wildlife observations for conservation using public participatory mapping methods. *Biological Conservation,227*, 141–151.

Callaghan, C. T., Poore, A. G., Hofmann, M., Roberts, C. J., & Pereira, H. M. (2021). Large-bodied birds are over-represented in unstructured citizen science data. *Scientific Reports, 11*(1), 1–11.

Chandler, M., Rullman, S., Cousins, J., Esmail, N., Begin, E., Venicx, G., ... Studer, M. (2017). Contributions to publications and management plans from 7 years of citizen science: Use of a novel evaluation tool on earthwatch-supported projects. *Biological Conservation, 208*, 163–173.

Clare, J. D., Townsend, P. A., Anhalt-Depies, C., Locke, C., Stenglein, J. L., Frett, S., ... Zuckerberg, B. (2019). Making inference with messy (citizen science) data: When are data accurate enough and how can they be improved? *Ecological Applications, 29*(2), e01849.

Cristescu, R. H., Goethals, K., Banks, P. B., Carrick, F. N., & Frere, C. (2012). Experimental evaluation of koala scat persistence and detectability with implications for pellet-based fauna census. *International Journal of Zoology*, 2012.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*(1), 20–28.

Ellwood, E. R., Dunckel, B. A., Flemons, P., Guralnick, R., Nelson, G., Newman, G., ... Mast, A. (2015). Accelerating the digitization of biodiversity research specimens through online public participation. *BioScience, 65*(4), 383–396.

Emerson, P. (2013). The original Borda count and partial voting. *Social Choice and Welfare, 40*(2), 353–358.

Feldman, M. J., Imbeau, L., Marchand, P., Mazerolle, M. J., Darveau, M., & Fenton, N. J. (2021). Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PloS one, 16*(3), e0234587.

Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A., & Verroios, V. (2016). Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering, 28*(4), 901–911.

Irwin, A. (2018). No PHDS needed: How citizen science is transforming research. *Nature, 562*(7726), 480–483.

Isaac, N. J., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution, 5*(10), 1052–1060.

Jones, F. M., Allen, C., Arteta, C., Arthur, J., Black, C., Emmerson, L. M., ... Hart, T. (2018). Time-lapse imagery and volunteer classifications from the zooniverse penguin watch project. *Scientific Data, 5*(1), 1–13.

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika, 68*(3), 373–389.

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment, 14*(10), 551–560.

Kristensen, T. V., & Kovach, A. I. (2018). Spatially explicit abundance estimation of a rare habitat specialist: implications for SECR study design. *Ecosphere, 9*(5), e02217.

Kullenberg, C., & Kasperowski, D. (2016). What is citizen science?-A scientometric meta-analysis. *PloS one, 11*(1), e0147152.

Kutlu, M., McDonnell, T., Elsayed, T., & Lease, M. (2020). Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research, 69*, 143–189.

Lasecki, W. S., Gordon, M., Leung, W., Lim, E., Bigham, J. P., & Dow, S. P. (2015). Exploring privacy and accuracy trade-offs in crowdsourced behavioral video coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1945–1954).

Lee, M. D., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people's rankings. *PloS one, 9*(5), e96431.

Leibovici, D. G., Rosser, J. F., Hodges, C., Evans, B., Jackson, M. J., & Higgins, C. I. (2017). On data quality assurance and conflation entanglement in crowdsourcing for environmental studies. *ISPRS International Journal of Geo-Information, 6*(3), 78.

Liu, H. K. (2021). Crowdsourcing: Citizens as coproducers of public services. *Policy & Internet, 13*(2), 315–331.

Lovett, M., Bajaba, S., Lovett, M., & Simmering, M. J. (2018). Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of amazon's mechanical turk masters. *Applied Psychology, 67*(2), 339–366.

Lukyanenko, R., Wiggins, A., & Rosser, H. K. (2020). Citizen science: An information quality research frontier. *Information Systems Frontiers, 22*(4), 961–983.

Martin, J., Kitchens, W. M., & Hines, J. E. (2007). Importance of well-designed monitoring programs for the conservation of endangered species: Case study of the snail kite. *Conservation Biology, 21*(2), 472–481.

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika, 80*(1), 205–235.

McKinley, D. C., Miller-Rushing, A. J., Ballard, H. L., Bonney, R., Brown, H., Cook-Patton, S. C., ... Soukup, M. (2017). Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation, 208*, 15–28.

Monk, A. F., Jackson, D., Nielsen, D., Jefferies, E., & Olivier, P. (2011). N-backer: An auditory n-back task with automatic scoring of spoken responses. *Behavior Research Methods, 43*(3), 888–896.

Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian cultural consensus theory. *Field Methods, 26*(3), 207–222.

Penman, T. D., Lemckert, F. L., & Mahony, M. J. (2006). Meteorological effects on the activity of the giant burrowing frog (Heleioporus Australiacus) in south-eastern Australia. *Wildlife Research, 33*(1), 35–40.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Plummer, M. (2003). JAGS: A program for the analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist, 31*(2), 163–177.

Root-Gutteridge, H., Brown, L. P., Forman, J., Korzeniowska, A. T., Simner, J., & Reby, D. (2021). Using a new video rating tool to crowd-source analysis of behavioural reaction to stimuli. *Animal cognition, 24*(5), 947–956.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution, 24*(9), 467–471.

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of*

the 23rd International Conference on World Wide Web (pp. 1049–1054).

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., ... Parrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation, 181*, 236–244.

Thibaut, J.-P., French, R., & Vezneva, M. (2010). Cognitive load and semantic analogies: Searching semantic space. *Psychonomic Bulletin & Review, 17*, 569–574.

Trouille, L., Lintott, C. J., & Fortson, L. F. (2019). Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human-machine systems. *Proceedings of the National Academy of Sciences, 116*(6), 1902–1909.

Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision, 101*(1), 184–204.

Weeser, B., Kroese, J. S., Jacobs, S., Njue, N., Kemboi, Z., Ran, A., ... Breuer, L. (2018). Citizen science pioneers in Kenya-a crowdsourced approach for hydrological monitoring. *Science of the Total Environment, 631*, 1590–1599.

Weller, S. C. (1984). Cross-cultural concepts of illness: Variation and validation. *American Anthropologist, 86*(2), 341–351.

Yi, S. K. M., Steyvers, M., & Lee, M. D. (2012). The wisdom of crowds in combinatorial problems. *Cognitive Science*.