Consensus Building in On-Line Citizen Science.

NIRWAN SHARMA, Knowledge Media Institute, The Open University, United Kingdom
LAURA COLUCCI-GRAY, Moray House School of Education, University of Edinburgh, United Kingdom
RENÉ VAN DER WAL, Department of Ecology, Swedish University of Agricultural Sciences (SLU), Sweden
ADVAITH SIDDHARTHAN, Knowledge Media Institute, The Open University, United Kingdom

A number of initiatives invite members of the public to perform online classification tasks such as identifying objects in images. These tasks are crucial to numerous large-scale Citizen Science projects in different disciplines, with volunteers using their knowledge and online support tools to, for example, identify species of wildlife or classify galaxies by their shapes. However, for complex classification tasks, such as this case study on identifying species of bumblebee, reaching an agreement between volunteers - or even between experts - may require consensus-building processes. Collaboration and teamwork approaches to problem solving and decision-making have been widely documented to improve both task performance and user learning in the real world. Most of these processes and projects are mediated online through feedback delivered in an asynchronous manner, and this article thus addresses a central research question: How do participants involved in species identification tasks respond to different forms of feedback provided in online collaboration, designed to support peer-learning and improve task performance? We tested four different approaches to feedback within a collaboration task, where participants reviewed their previously annotated data based on information curated from their peers on a long running online citizen science initiative. The selected interfaces have a strong foundation in social science and psychology literature and can be applied to citizen science practices as well as other online communities. Results showed that while all four approaches increased accuracy, there were differences based on the types of consensus that existed before collaboration. Such differences highlight the usefulness of different forms of feedback during collaboration for increasing data accuracy of identification and furthering users' expertise on identification tasks. We found that anonymised and goal-directed free text comments posted on social learning interfaces were most effective in improving data accuracy as well as creating opportunities for peer-learning, particularly where the species identification task was more difficult. This study has significant implications for extending the practice of citizen science across formal and informal learning environments and reaching out to a variety of users.

CCS Concepts: • **Human-centered computing** → **Computer supported cooperative work**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Consensus building, Artificial Intelligence, Citizen Science, Species Identification, Expert ratings

Authors' addresses: Nirwan Sharma, nirwan.sharma@open.ac.uk, Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom; Laura Colucci-Gray, Laura.Colucci-Gray@ed.ac.uk, Moray House School of Education, University of Edinburgh, Edinburgh, United Kingdom; René Van Der Wal, rene.van.der.wal@slu.se, Department of Ecology, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden; Advaith Siddharthan, advaith.siddharthan@open.ac.uk, Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom.

1 INTRODUCTION

Citizen science and crowdsourcing projects focus on using the capabilities of paid or unpaid volunteers for data collection and annotation [3, 79, 94]. The internet provides the opportunity to collect or annotate data on a large scale by soliciting volunteers online. However, this raises concerns regarding data quality [44, 47, 54]. To provide safeguards on data annotations sourced online, typically multiple annotations are requested, and a minimum level of consensus is expected for an annotation to be accepted; i.e. data is validated by other volunteers [79]. A common crowdsourcing task is the classification of data captured through images [39] using project resources such as identification guides and keys [77, 83]. Annotations are either accepted independently where users submit their classifications without any collaboration or through a collaborative interface. Furthermore, most of these platforms provide their users a means to share their knowledge and expertise, either openly on the classification task (for collaborative interfaces) or through open discussion forums (for independent classification projects) to support community building and peer learning [43]. However, there is a lack of understanding on how collaborative design techniques affect users' online behaviour and whether these techniques can be effectively utilised to improve data quality as well as engagement and learning around the task. This study explores the design of interfaces that allow volunteers to collaborate on classification tasks with their peers. Specifically, we consider differences in the types of feedback and learning processes generated, and how these can impact on the formation of more stable citizen science communities. We focus on asynchronous collaboration, where there is no expectation that online volunteers will all be available at the same time.

1.1 Background

The usefulness of citizen science for research has been widely highlighted [19, 48, 71] and citizen science practice has benefitted from advances in digital technologies, such as internet and mobile communication [4, 5, 44]. Its growing contribution to research is visible in diverse ways, from monitoring the environment and biodiversity, to promoting question-driven research and statistical innovations in the handling of variable data sets [11, 25, 38, 42]. Web based citizen science projects have successfully employed volunteer capabilities to accomplish a wide variety of tasks such as digitising biological records [28], predicting protein structures [23], and classifying shapes of galaxies [8]. The success of these and numerous other crowdsourcing projects highlights the important roles that volunteers can play in scientific research as well as for the common good [48, 84]. Data quality is a common concern for these projects and most include safeguarding mechanisms for data validation [44], for example, comprehensive training of volunteers; providing guides, protocols and tools to support data collection; validating collected data samples by experts and building statistical consensus models for classification or object identification tasks [8, 14, 15, 77, 82, 87].

Human collaboration has been a topic of interest across many disciplines such as psychology, social sciences, organisational behaviour, education and more recently human-computer interaction [3, 9, 35, 55, 58, 63, 66, 81, 97]. With the growth of the internet and subsequent rise of online communities such as Wikis, social media websites, citizen science and crowdsourcing platforms, online collaboration is utilised for a variety of purposes such as problem solving, user learning and engagement, consensus building, and decision-making. Performance of a group is, in general, qualitatively and quantitatively superior to the average individual [35, 55], but typically lower than the best member [45]. Identifying the best member or utilising the capabilities of high performing individuals can help increase the group performance as they can guide a group of inexperienced members towards better decision-making [26]. It has also been reported that groups can achieve better performance than even the best individuals for several problem-solving tasks [51].

Research has also highlighted benefits of collaboration in terms of learning, motivation and engagement which, in turn, can lead to sustained focus and deeper learning [9, 58]. Collaboration and teamwork have been a regular practice in the scientific domain, and scientific discoveries have often been made through collaboration [65]. In this article, we focus on collaboration amongst non-expert users for consensus building on a classification task and how digital interfaces can be designed to support this. A variety of feedback techniques have been explored for online collaboration to support problem solving, decision making and learning. In this article, we study some of the main techniques and affordances derived from the literature and from citizen science practice, with a particular focus on improving learning and data quality.

1.2 Feedback processes at the core of collaborative learning

Dialogism, a framework for research into computer-supported collaboration [86], provides an effective method for studying interaction and communication between participants mediated by computers, especially for tasks which require debate, negotiation and coordination among a group [6, 86]. Supporting collaboration for citizen science activities that require learning new skills and knowledge creation can also be examined using the dialogical framework, to understand in the first instance how digital platforms can mediate learning through providing feedback for helping to construct meaning or make sense of a new concept [81]. More widely, its use can make visible how feedback supporting collaboration in citizen science activities fosters social learning and civic participation, enabling a wider range of contexts and experiences to contribute to shaping research agendas [10, 31, 67]. We utilise the concept of feedback [87] in the context of a consensus building task, which may progress through divergence and convergence of multiple viewpoints or 'voices' through debate between the participants for problem solving and reaching consensus. For successful dialogue to happen it is important to design a dialogic (interactional) space for presenting multiple viewpoints in the contexts of the collaboration. To design and understand such an interactional space we utilise multiple design strategies by drawing on literature on collaboration and feedback in social sciences and education and their applications across citizen science practice.

1.2.1 Social persuasion. Within an interactive space, goal setting has been shown to be a particularly effective strategy for increasing contributions and motivation [9]. Setting individual and group goals can have a positive effect on group performance by motivating volunteers in accomplishing tasks important to the success of the group [9, 40, 99]. However, monitoring own and peer activities via feedback is core to supporting collaborative learning environments. For example, prompts and visualisations are often used in online communities for problem-solving, learning and collective decision-making for monitoring progress and activities [40]. To build consensus in a group, highlighting the level of agreement within the members of the group may act as a persuasive method towards taking a particular course of action, as individuals use that information to narrow down on a set of options. This feedback is used with success in the commercial sector to influence consumer choices, and literature suggests that revealing majority 'votes' [63] and levels of consensus [60] can influence other group members in problem-solving contexts. For instance, Project Discovery, utilizes this method, providing community consensus as a feedback for classification tasks without an expert annotation [53]. However, it may be limiting to equate consensus simply with agreement, uniformity or homogeneity [74]. Consensus-building that relies heavily on individuals' dispositions and drivers towards social conformity reinforces habits and behaviours whereby the learning goal is determined a priori; it limits the function of collaboration to the transmission and confirmation of existing ideas while discounting evidence that may not fit with the expectations of the individual or the group [98].

1.2.2 Expertise-driven consensus. Expertise plays an important role for collaboration as members of
the group may possess different skills, have variable levels of knowledge and experience, and show
different interests. Highlighting individual uniqueness and difference can increase contributions
from people collaborating online, while identifying the expertise of the individuals in the group
can be an effective strategy to persuade other members in decision-making tasks [9, 46, 88]. User
expertise ratings, a common method for highlighting individual expertise, are ubiquitous in online
communities, whether it is for ecommerce, tourism, expert reviews, social media or even citizen
science [49]. More specifically, citizen science projects such as iSpot [2] and iNaturalist [1] make
use of user expertise through the use of reputation scores (in iSpot) and leaderboards (in iNaturalist)
to highlight 'best performing' members of the community. Both social persuasion (see 1.2.1) and
expertise-led consensus are widely utilized in online communities for building consensus; however,
both these methods may generate a conformity effect whereby members of the group may agree
with a majority position that may sometimes be incorrect [61, 93] due to the power of influence
exerted by the group or by one of its members [61].

Online communities also have means to enable anonymous collaboration, and this technique
is commonly utilised by many users, for example, in social networking websites to maintain
privacy when collaborating on sensitive issues. Anonymity has shown to be effective for increasing
contributions, but may also have negative effects such as sharing incorrect information, uncivil
behaviour or loss of reputation for contributing users [17, 30, 75].

1.2.3 Social learning. In problem-solving contexts such as citizen science and scientific research, it
is not only the performance on a problem-solving task that may be affected by the level of expertise
of an individual [50] but also how problems may be approached [18, 76]. In such contexts, a first
level of social learning may occur through modelling, followed by reproduction and apprenticeship
of a particular way to frame a problem or execute a task. For example, as novices gain expertise
over time or through training they tend to approach the problem more like experts [20, 76]. But a
second level of social learning may also entail increased levels of self-regulation and self-efficacy in
learners [100]. For example, any form of collaboration which enables communication and sharing
of resources (cognitive or technological) among members of a group (even through chat boxes) can
positively affect attention and engagement. Expanding the number of possible feedbacks on a task
increases the possibility for individuals to observe the effects on a product or a course of action, thus
improving the quality of the work [40, 96]. For instance, a form of sequential task editing which
enables dialogic interaction, where subsequent users edit the input of previous users, has been
shown to be effective over creative tasks [3, 97]). Due to the nature of online communication, which
is largely asynchronous and sequential, this technique can be effective for supporting collaboration,
as each member of a group can utilise the shared inputs of the previous users while providing their
contributions.

Following Rose et al. [1995] we can distinguish between (1) the level of generalised consensus in
a scientific community, which makes understanding possible, and (2) the level of immediate social
interaction, which draws upon difference of opinions and relies on evidence and argumentation.
While it is accepted that these two levels are integral to one another both in social and in scientific
practice, this distinction between levels is particularly useful to citizen science practices, as it points
to the possibility to overcome the idea of scientific information as a series of mental representations
that can be processed and replicated in the heads of individuals. Such an approach would – in
fact- limit the scope of the citizen science inquiry to well-known species. Instead, the ability to
identify unknown or difficult species may be a quality and feature of a diverse community, which
incorporates local peoples' experiences and could include machines as part of a third level of
social learning processes, such as those occurring in extended communities of socio-material

practice [41, 73]. This understanding of social learning is most closely related to the ideas of situated learning [52, 91], distributed cognition [69, 70], and activity theory [29]. Lave and Wenger specifically called out the problematic assumption that treats technology as a given instead of focusing on its interrelations with other aspects of a community of practice [52].

1.2.4 Citizen science practice. Citizen science projects such as iNaturalist [64] and iSpot [80] rely on creating communities of nature enthusiasts uploading photographs of plant and animal species as well as identifying specimens on photos shared by other members of the community. More specifically, they i) require multiple annotations for producing reliable data, ii) highlight expertise of the members using user ratings (a method to highlight individual uniqueness) and iii) use free text commenting to capture opinions as well as scientific information. Annotations are usually provided by members of the community using their expertise while additional members can agree or even improve the existing level of annotations [80]. However, the feedback effect of these techniques (individually or in combination) on data quality, engagement and citizen science learning is largely unknown.

Other citizen science platforms such as Zooniverse, Eyewire, Project Discovery and BeeWatch recruit volunteers for online tasks that primarily concern processing of data [53, 59, 85, 87], with "independent classifications" being solicited. They utilise the principles of goal setting, providing shared learning resources such as tools and visualisations and level of agreement to enable collaboration among community members for consensus building. Zooniverse and Eyewire also provide forums for discussion and dialogue, which in case of Eyewire is in real-time, for community building and peer learning [59]. Zooniverse users utilize a social interface, where members of the community can discuss classifications tasks enabling them to learn to identify through 'practice proxy', a peripheral participation strategy that provides feedback to newcomers within a community of practice [52, 62]. However, these forums are not directly linked to the classification task and thus might limit opportunities for social agency [41] through collaboration and peer-learning, the latter being documented as an important dimension for tasks such as learning to identify species as part of a community of practice [27]. Additionally, opportunities for collaboration among members may also help improve scientific data quality, user-learning and engagement – dimensions, important for Citizen Science practice [13, 37, 92]. Hence, the objective here is to look more closely at how identification tasks derive their meaning as 'social practices' for the people involved, by taking into account their dependence on the affordances and design of the interfaces for their meaning-making.

1.3 Contributions

The overall research aim of this article is to understand the role of feedback strategies for collaboration to support user learning and performance on an on-line (asynchronous) consensus building task. We developed online collaboration interfaces operationalising four feedback techniques which were then used by the participants to perform species identification tasks in a citizen science context. Three of these techniques, i.e. highlighting level of agreement, displaying user expertise and providing means of communication through text such as chat boxes or commenting, are ubiquitous with respect to the gathering of user data and in supporting online communities dealing with user-generated content (e.g. ratings and feedback for ecommerce; open source programming communities; social media websites and wikis; public forums and question-answer websites). The fourth, a Natural Language Generation (NLG) system, is novel and deploys an AI to mediate the task. In addition to the feedback techniques, we also identified three different situations where there was a lack of consensus on the task, thus necessitating intervention through the collaboration interface.

The citizen science and crowdsourcing literatures have not previously explored the effects of feedback techniques on data quality and citizen science learning in collaborative settings. We are also unaware of any studies that have investigated the different types of disagreements identified and studied in this article, which are a significant dimension when seeking to widen the reach and potential of citizen science to involve citizens in important issues related to science and society [31]. This article adds to the literature by addressing these significant gaps.

2 MATERIALS AND METHODS

We investigated the role of different asynchronous online collaboration techniques in impacting performance, learning and engagement within online citizen science communities.

2.1 Dataset

We used data from the citizen science platform BeeWatch (recently relaunched as https://www.plantingforpollinators.org), which provided participants with the option to submit images of bumblebees online as well as to independently identify images submitted by fellow BeeWatch participants through crowdsourcing [87].

The UK-based platform is designed to help users identify photographed specimens to species level, as one of 22 possible bumblebee species. In general, the species can be differentiated on the basis of colour pattern and morphological features (e.g. colour band pattern on their bodies, presence/absence of pollen baskets on their hind legs, size of the face). There are considerable differences in identification difficulty between species [79], with some being readily identified even by novices, and others requiring considerable expertise. Additionally, features may not be visible or harder to detect in photographs, adding to the difficulty of accurate species identification.

Specifically, we used photos submitted to BeeWatch for which multiple independent species identifications have previously been obtained from participants, but without those leading to agreement. The crowdsourcing component of BeeWatch has received more than 25,000 individual identifications for 6,500 images submitted. The independent species identifications submitted by BeeWatch participants were used to calculate the level of consensus for each image; and when a consensus threshold was reached [79], the species identification was accepted, and the original submitter was sent feedback on the species identification. Each image could accept a maximum of 10 independent identifications from the crowd. If there was still a lack of consensus, the image needed to be sent to a bumblebee expert for identification, a time and effort intensive step that would be useful to minimise. Such images, for which there was lack of consensus within the crowd, provided us with a dataset for studying the effects of collaboration for consensus building.

2.2 Types of Consensus Encountered

From this dataset, three different situations were identified where crowdsourcing did not provide an identification that met the consensus threshold for acceptance. All of these situations led to an expert identifier being solicited for authoritative identification of the species. We label these situations as three different consensus types:

- Consensus Correct (CC): If there is an existing majority tending towards the correct identification (i.e. if at least 5 out of the maximum 10 identifiers have identified the image as a single species and that species is the correct answer), but not reaching the required threshold for acceptance (determined by a Bayesian consensus model). We label these images as Consensus Correct (CC) images.
- Consensus Not Correct (CNC): If there is an existing majority tending towards an incorrect identification (i.e. if at least 5 out of 10 identifiers have identified the image as a single

species and that species is not the correct answer), but not reaching the required threshold for acceptance. We label these images as Consensus Not Correct (CNC) images.

- No Consensus (NC): If there does not exist a majority of at least 5 out of 10 for any single species, we label these images as No Consensus (NC) images.

We introduced a collaboration step in such situations, whereby the participants could review their identifications in the light of new information generated from independent peer annotations. We focused on four different types of on-line collaboration interfaces described below, drawn from both literature and practice, as summarised in Section 1.2.

2.3 Collaboration Interfaces

We designed four different collaboration interfaces to investigate their effects on consensus building (Fig. 1 to Fig. 4). The interfaces were co-designed with regular inputs from two bumblebee experts who tested them iteratively to improve the design and workflow of the interfaces to be used in our experiment. Each interface implements one of four types of feedback, which impacts the process of collaboration.

2.3.1 Distribution Interface. To determine whether information on the existing consensus distribution would influence participants into reviewing their identifications - possibly towards the majority opinion - 'Distribution' was used as one of the techniques for persuasion [53, 60, 63]. The first design intervention uses pie chart visualisations of the "Distribution" over species identifications to understand its effect during a consensus building task (Fig. 1). In computer-mediated task-based scenarios, due to the nature of communication (asynchronous and anonymous), the social pressures which are reported in face-to-face communication may be less influential [95]. Yet, the level of agreement on a task is often utilised in online communities for problem solving and decision-making. We are interested in whether online participants are persuaded to modify or change their opinion solely based on what other anonymous participants say. In this instance, feedback will not include the possibility to incorporate specific guidance from others to reduce the number of options and increase self-efficacy as per the second level of social learning that we identified in the literature, but will be largely reinforcing existing knowledge and beliefs [98].
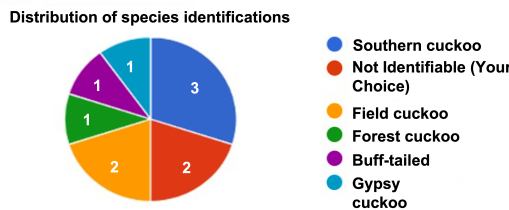

Fig. 1. The Distribution Interface. This interface shows the existing consensus information using a pie chart.

2.3.2 User ratings Interface. The second interface was developed to assess the effect of expertise of other participants. All 10 species identifications were shown as a list with ratings for the participants that provided them. Participants' own ratings were not displayed in order to prevent comparison of their own expertise with others. The identities of participants were anonymized and two user ratings were shown as icons (Fig. 2): one to represent experience (blue bars), specifically the number of previous identifications by that participant on BeeWatch; and another (golden stars) to represent skill level, constructed on the basis of a user's historical identification accuracy (where one star represents <35% accuracy, two stars indicate accuracy between 35-55% and three stars indicate

User	Identification	Submission	Accuracy
1	Ruderal		★★☆
2	Heath		★★☆
3	White-tailed		★★☆
4	Barbut's cuckoo		★★☆
5	Heath		★★☆
6	Heath		★★☆
7	Barbut's cuckoo		★★☆
8	Barbut's cuckoo		★☆☆
9	White-tailed		★☆☆
10	White-tailed		★☆☆

Fig. 2. User Ratings interface showing the existing species identifications together with the user rating of each user. The column headings represent users (User), their identification (Identification), total number of previous BeeWatch identifications (Submission) and accuracy on the previous identifications (Accuracy). Each row represents a single user with their existing identification, number of previous identifications that the user submitted (blue bars) and the existing accuracy of the user on previous identifications (golden stars).

accuracy>55%). The list was ordered in decreasing order of experience. Icons rather than real numbers of submissions or percentage accuracy were used to enhance communication values, using icons commonly used on digital interfaces representing information (i.e. golden stars for ratings, indicator levels for e.g. sound/mobile signal).

Ratings of user expertise are widely utilised in both online and citizen science communities, and the literature has suggested they are effective for engagement and for improving the performance of a group [9, 46, 49, 80, 88]. We use "User ratings" as the second design intervention, whereby we highlight the expertise and experience of individual members of the group and study its effect for building consensus in a group. For this interface, participants' own self-reflection is supported by feedback pointing to specific areas of expertise which would to some extent support apprenticeship as per the first level of social learning [20, 76].

2.3.3 Social Interface. Sharing of resources and knowledge through communication is important for building communities of practice; providing a forum for members of a group to communicate may thus influence task accuracy and consensus [3, 40, 96, 97]. Additionally, a knowledge sharing forum may support peer learning, an important outcome of citizen science. Social communities-based projects use this method to capture expertise and to provide opportunities for high expertise individuals to guide others [64, 80]. We designed a third "Social" interface whereby members of the group could effectively communicate their knowledge and expertise using anonymous goal-directed comments. In the third interface (Fig. 3) the user was provided with the option to share views, motivations underpinning an identification and further relevant experience or contextual information related to an image (such as image quality or angle of the specimen) with others through free text comments. In this interface, the user was first given textual information, highlighting their identification and the alternative identifications provided by their peers. The user was then encouraged to leave comments, specifically focusing on the features that may help their peers with the identification. It was also mentioned that they could read comments left by their peers to see if they might have identified the species incorrectly. All comments from users were anonymised for the platform to enable greater scope for dialogic feedback, by incorporating the detail of the specific

You identified the bee as Southern cuckoo bumblebee. Other users identified this bee as Forest cuckoo bumblebee, Southern cuckoo bumblebee and Not identifiable

Please comment below to help other users by highlighting features that helped you in the identification. If there are comments from other users you can use this information to see if you missed something.

Leave a Comment

Post

6369 — Dark wings indicate cuckoo; single dark yellow band on abdomen just above tail suggests female Southern cuckoo
September 18, 2016

5841 — The hind leg is roughly hairy all over. along with dark wings and the yellow patch above the white stripe all point to a Southern cuckoo bumblebee.
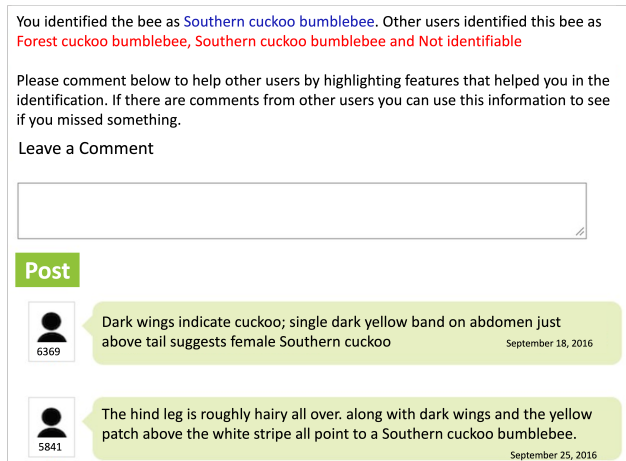September 25, 2016

Fig. 3. Social interface with chat boxes. This interface firstly highlights the user's own submission (blue font) and provides the alternate submissions by the other users. Then the interface encourages the user to leave comments that may help other users. The interface also encourages the user to read any existing comments, which might help in building consensus.

features that participants considered significant for identification as well as their justifications for the identification made, in line with second and third level of social learning [31, 41, 100].

2.3.4 Natural Language Generation (NLG) Interface. Artificial Intelligence and machine learning algorithms are increasingly being researched and utilised in ecological and citizen science projects [56, 57, 87, 90]. As more and more projects utilise these technologies, human-AI interaction in citizen science becomes an important domain of research and practice [16]. To add to this relatively new area of research we utilise a Natural Language Generation system to study the role of AI for supporting online collaboration. More specifically, we use a system which provides machine generated texts highlighting any differences in visual features between the different species independently identified by participants. This machine-generated text concerning comparison was adapted from an existing implementation used to provide feedback to citizens on their submissions on the BeeWatch platform by explaining what features to focus on [12, 87]. In short, when a participant's identification is found to be incorrect by an expert, the NLG system uses the identification key to identify the visual features that differ between the species selected by the participant and expert, and organises these differences into a formative feedback message that explains why the identification is incorrect and what features to focus on to make the correct identification. In this work, we automatically generated NLG text comparing the participant's identification to the existing consensus identification as feedback for re-appraising their original identification. This enables participants to consider alternatives and support consensus-building. The use of machine-generated texts for consensus building may be a suitable method for creating sustainable online citizen science communities where expert knowledge can be presented in a user-friendly manner to promote collaborative learning between community members.

This interface was designed to determine whether the formulated differences in visual features between bumblebee species, which are machine-generated and presented in the form of natural language texts, would be useful to build consensus. This interface (Fig. 4) used automatically generated texts to identify the features that distinguish the user's identification from those of other users. We first automatically generated text that highlighted the features of a user's existing identification,

You identified the bee as Buff-tailed bumblebee. Other users have reached a different conclusion. See the comparison text below which highlights differences and enables you to confirm or review your identification

Your choice: Buff-tailed bumblebee

Queens and males of this species have a buff coloured tail. Workers have a white tail, which makes it difficult to separate them from White-tailed bumblebees. A narrow fringe of buff-coloured hairs at the top margin of the tail, when seen, can identify Buff-tailed workers. The two yellow bands are golden in this species and more of a lemon-yellow in the White-tailed bumblebee.

Other Choices

White-tailed bumblebee vs Buff-tailed bumblebee(Your Choice)

Males and queens differ by tail colour, Buff-tailed are buff and White-tailed are white. The workers of these two species are very hard to separate, but the tail of the White-tailed bumblebee is pure white, with a complete absence of any buff-coloured hairs. The colour of the two yellow bands is brighter and more lemon on the White-tailed bumblebee than that seen in the Buff-tailed bumblebee. The male White-tailed has yellow facial hairs, whereas the male Buff-tailed has black facial hairs.

Forest cuckoo bumblebee vs Buff-tailed bumblebee(Your Choice)

The Forest cuckoo does not have a pollen basket whereas the Buff-tailed bumblebee has a pollen basket. The Forest cuckoo's wings are smokey dark whereas the Buff-tailed bumblebee's wings are clear. The Forest cuckoo's abdomen is either black with one yellow band near the top of it and a white to greyish white tip or black with either a grey to yellow or ginger tip whereas the Buff-tailed bumblebee's abdomen is black with one yellow band around the middle of it and a white to buff tip.

Fig. 4. Natural Language Generation interface showing machine generated texts. This interface firstly highlights the user's own identification (in this case Buff-tailed bumblebee) together with NLG text which highlights specific features of that species. Then the interface displays comparison NLG texts for all other species submitted by other users.

so that these could be compared against the submitted image(s). Next, we automatically generated text that reported the differences between the users' identification and each of the conflicting identifications. Notably, with reference to our theoretical framework focused on feedback in social learning, this fourth interface focused more on the affordances of the machine rather than the strength of the social environment that was generated, thus further decoupling the social from the material in the identification task. As we will discuss in the data analysis that follows, this aspect was important in order to probe our understanding of the effect of feedback on learning in socio-material interactions [41].

2.4 Procedure

Based on previous experience with employing BeeWatch data for user performance studies, we sought at least 50 participants for the study and 15 images for each of the four interfaces. We eventually decided upon working with 72 photographs from BeeWatch, which were randomly selected from a total of 497 images that had not reached the required threshold for acceptance (see Siddharthan et al 2016 for more detail). Of the 72 images, 36 had consensus of at least 5/10 identifications for the correct species (CC condition), 16 had consensus of at least 5/10 identifications, but for a species that was not correct (CNC) and 20 had no consensus of at least 5/10 identifications for any species (NC). All 72 images had an expert identification, which was used for evaluating accuracy of participants' identifications before and after the collaboration step. Participants who had provided an identification for any of the selected photographs were contacted via email (114 in total) and invited to participate in a study to review their previously submitted identifications. Each participant viewed different numbers as well as types of interfaces depending upon the number of

images in this study's sample which they had previously identified. Hence, some of the participants saw all interfaces while others saw only one, two or three.

The email contained the information about the study as well as the link to a webpage (see Appendix A.1). After clicking on the link provided, participants were shown the consent form for participation and – upon agreeing to participate – they were shown a list of images allocated to them for review (see Appendix A.2)). When selecting an image, the user was directed to the collaboration interface associated with that image, together with "Review" and "Do Not Review" buttons at the bottom of the page (see Appendix A.2 for an example workflow).

In addition to comparing accuracy before and after collaboration, information regarding the reasons why the participants clicked on "Review" or "Do Not Review" during the process was also collected (see Fig. 5). Participants were not obliged to respond. In the 'not reviewing' popup window, a radio button was provided saying "My existing identification is correct", as well as a free text option. When clicking on "Review", users were directed to a page where they were shown the image and collaboration interface together with the guide used previously to derive at the identification (Appendix A.2). On this page, they could submit their new identification, which could be the same as the original one or a different identification. After submitting the new identification, participants could also provide their reasoning for reviewing their identification (Fig. 5).

Fig. 5. Popups for collecting reasoning information when a user clicked "Review" (top) or "Do Not Review" (bottom). The top popup box provides a text box for the user to provide specific information on why they reviewed their existing identification. If the user selected the same species after reviewing the popup displayed a note to the user as shown in the figure. The bottom popup was shown if the user did not review their identification after seeing one of the collaboration interface. This popup provides two options: one where the user can indicate that their existing identification was correct and another to make clear any other reason (through free text) why they did not review their identification.

Through the above processes of reviewing and reasoning, participants were engaged in dialogic interactions (visual and text-based) with information provided from other participants, interaction with the interfaces and process of reviewing. These interactions may have supported divergence and convergence of participants opinions, and the reasoning information (of either their previous or new annotations) from these interactions were hence utilised to assess how participants may have engaged with the collaborative interfaces for consensus building. The qualitative results thus (see 3.3) provide a summary of engagement across each interface assessed from the reasoning provided by the participants.

2.4.1 Participants and responses. A total of 61 out of 114 invited BeeWatch users (53.5% response rate) participated in the study and completed a total of 373 out of 720 (51.8%) possible identifications.

Of the total of 72 images the mean number of responses for each image was 5.8 with a minimum of 2 and a maximum of 9 responses per image. The interfaces received a similar number of responses (94, 96, 91 and 92 responses for the Distribution, User Ratings, Social and NLG interfaces respectively). Response distribution was also rather even across interfaces with respect to consensus type, with Consensus Correct images having 50, 49, 50 and 49 responses; Consensus Not Correct images having 18, 21, 17 and 19 responses and No Consensus images with 26, 26, 24 and 24 responses for Distribution, User Ratings, Social and NLG interfaces respectively.

2.4.2 Statistical procedure. All statistics was performed using R version 4.1.0 [72]. New accuracy was fitted using R's base glm function with old accuracy, interface type, consensus type, number of interfaces used (by the participant during the experiment) and the interaction of interface type and consensus type; and the ANOVA function from the car package was then used to test for significance. To test interface effect by consensus type (CC, CNC and NC), hypothesis testing of count data (change in accuracy by interface) was done using the Fischer test due to the presence of cells with low counts.

3 RESULTS

We focused on how participants responded to different forms of online collaboration techniques, which were designed to support peer-learning and improve user performance during species identification tasks. We compared the four different collaboration interfaces, in terms of their effectiveness for consensus building on the task of species identification and their potential to introduce or reinforce bias. For the latter, we considered whether there was already a level of consensus (5/10 identifications) for any species and if so whether it was already for the correct species. Our expectations were that: (a) user accuracy would improve through reviewing their classification with any interface; (b) where there was an existing consensus, the Distribution and User Rating interfaces would persuade participants to revise their identification to that consensus, whether or not it was correct; (c) the Social and NLG interfaces, by focusing on the identification skills rather than the peer responses, would outperform the two majoritarian interfaces for images where there was no existing consensus, and also where the existing consensus was for an incorrect identification.

3.1 Change in consensus

The majority of images, 96 out of 117 reviewed (82%), for which participants clicked "Review" were initially incorrectly identified by that participant (expert identification different from participant identification); and the majority of images, 151 out of 256 not reviewed (59%), for which participants clicked "Not Review" were initially correctly identified by the participant (expert identification same as participant's). This indicated that incorrect identifications were more likely to be reviewed during collaboration. However, for the incorrect images which were not reviewed (105 out of 256), participants mostly selected 'My identification is correct' as the reasoning. This suggested that the information provided through the interfaces may have been either lacking or not persuasive enough to review the original identifications. Figure 6 shows the percentage of identifications changed and reviewed across interfaces, highlighting that User Ratings interface (35.4% reviewed, 32.3% changed) may have been the most persuasive, followed by Social (32.9% reviewed, 28.6% changed) and NLG (29.3% reviewed, 26.1% changed). Using the Distribution interface (27.6% reviewed, 24.4% changed) resulted into lowest percentage of identifications reviewed and changed.

The results in Fig. 7 show that the level of consensus changed from before to after collaboration for many of the 72 images worked with. Consensus was defined as proportion of the group (between 0 and 1) that selected the most-selected species, and was different from accuracy, where
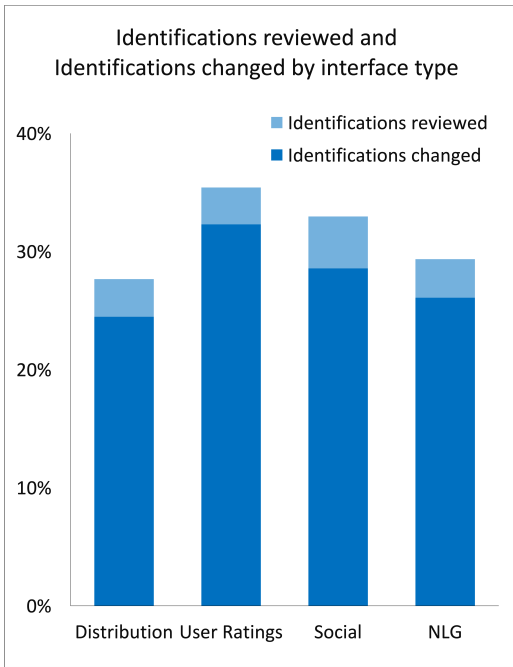
Fig. 6. Figure showing the percentage of identifications that were reviewed and then eventually changed by the volunteers across each Interface type.
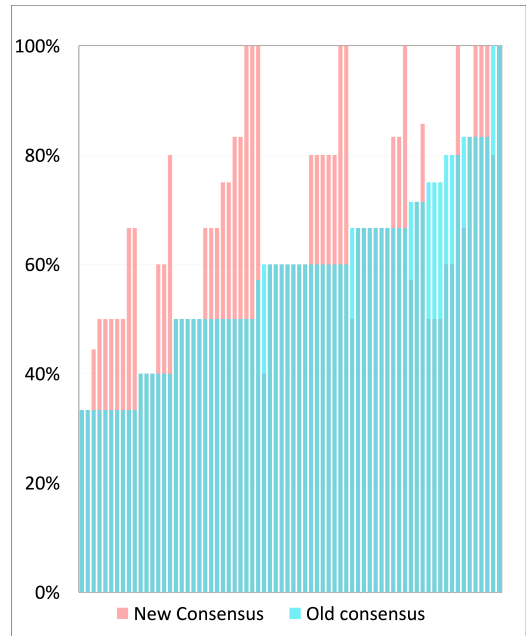

Fig. 7. Figure showing the percentage agreement for all 72 images before (blue) and after (pink) collaboration, ranked on the basis of 'original (i.e. old) consensus values'

each individual identification was evaluated as correct or wrong by comparing to a gold standard expert identification (0 or 1) and then averaged. The results illustrate that there was an increase in consensus after reviewing as the average consensus increased from 58.2% to 67.6% overall. According to consensus type when the original consensus was correct the increase was from 63.4% to 75.9%, for incorrect consensus images it increased from 62.9% to 63.9% and for no consensus images from 45.1% to 55.6%. Moreover, the greatest gains were when consensus was initially relatively low, and where consensus was already high, reviewing frequently reduced the level of consensus.

3.2 Effect of interface and consensus type

Significant variation in new accuracy was explained by the effects of old accuracy (χ^2= 158.4, df = 1, p < 0.001), Interface type (χ^2= 7.67, df = 3, p = 0.053) and Consensus type (χ^2= 27.26, df = 2, p < 0.001), while the number of interfaces used by each participant and the interaction of Interface type and Consensus type did not significantly affect new accuracy. Average increase in accuracy after reviewing was 4.2% for the Distribution interface, 12.5% for the User Ratings interface, 18.6% for the Social interface and 11.9% for the NLG interface (Fig. 8D), indicating that the Social interface design led to the largest average increase in accuracy.

When taking Consensus type into consideration, we found that when the initial consensus was towards the correct species (Fig. 8A), the increase in accuracy differed significantly by interface (p < 0.05), suggesting the usefulness of the User Ratings interface for this category of images to enable participants to review their identifications. Where consensus was towards the incorrect species however, no significant difference of interface type was found (Fig. 8B). Finally, for images where there was no consensus towards a particular species (Fig. 8C), differences across interface type
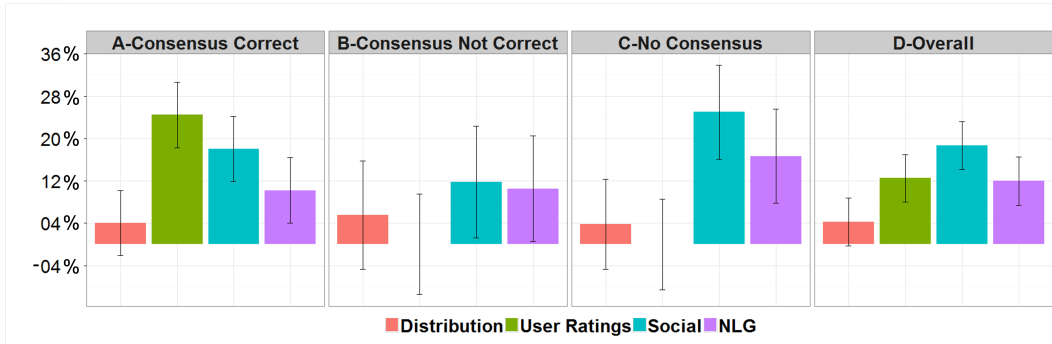
Fig. 8. Bar Plots showing mean percent changes in accuracy across Interface type and Consensus type in the first three columns and overall mean change in accuracy in the last column (Error bars represent standard error of the mean). The missing bars for User Ratings in (B) and (C) highlight no overall change in accuracy for these cases as the mean increase in accuracy was equal to the mean decrease in accuracy for these cases.

were again significant (p < 0.01). Of all interfaces, User ratings were most sensitive to Consensus type, improving accuracy markedly (more than any other feedback interface) for the Consensus correct images but leading to no improvements in accuracy for other image types, highlighting the potential negative effect it may have in such cases. Across all consensus types, the Social type achieved greatest accuracy, followed by the NLG interface, both contrasting sharply with the low level effect of the Distribution interface.

3.3 Qualitative results on engagement

3.3.1 Distribution. For the Distribution interface, participants who did not review their identifications mostly selected the "My identification is correct" option as previously highlighted, although they might have been unsure if indeed their identification was correct due to the limited information provided by the interface. Comments showing the reasoning from some of the participants highlighted that the distribution information may have been insufficient for them to either change their existing identification, for example because they were not given any new information about how to identify tricky photographs ("Not an adequate image to identify with sufficient confidence {4900}"; "Unable to identify from this picture {1400}"). This was evident also for consensus not correct and no consensus images ("I can't be sure so not going to change my view {6301}" ; "I think it may well be a buff tailed, but I am not able to see the tail colour, so I cannot be sure, so I think it is better to say not sure. {3091}"). Participants who did review their identifications did not mention influence of the interface information explicitly and confidence in their new identifications appeared often rather limited ("I'm not sure about the wings, they may be smokey and I don't think there's a pollen basket {4438}"; "Looking again, I think that this is .. have pollen baskets … round face with a yellow 'moustache' … yellow bands … the second look has helped me come to a decision. Though I am still not very confident of this id. {3091}").

The Distribution interface was least effective in making participants review their incorrect identifications across all consensus types (cf. Fig. 8). While this interface gives information about the existing disagreement within a group, it offers limited opportunities to engage in a dialogic interaction with the group/interface [6, 86]. In line with the theorisation of feedback via persuasion [61, 93], the interface relies on the participant's self-motivation and own prior knowledge to carefully review their existing identification. Though some of the participants were motivated to review, this was a small number compared to other interfaces, in line with it offering limited

impetus to participants to review their identification. For those who did modify their identification, their comments highlighted the limited use of the interface in helping them revisit their species identification, thus preventing self-regulation and self-efficacy.

3.3.2 User Ratings. For the User Ratings interface, some of the participants who reviewed their identifications explicitly mentioned that they were persuaded by other users ("I think the others are correct, I must have clicked the wrong button {6269}"). Analysing the reviewed identifications, we found that more experienced users were indeed usually correct. More importantly, the reasons for reviewing indicated that some participants were not just mimicking or agreeing with the top user but also validated the features, which they might have missed previously ("Buff margins to white tail. Not observed in previous viewing {6691}"; "margin between the white tail and black has a hint of colour {1670}"; "What I think is a long tongue was missed by me originally … the yellow banding at back of thorax and front of abdomen is not quite correct .. {2939}").

These observations highlight that firstly the user ratings were persuasive for participants, which led them to review their identifications using the identification tool. As the participants were not aware of their own ratings the interface emphasises how the perception of expertise can be used to improve engagement. Although feature-based information was not provided on this interface, participants tried to pay attention to features different from those considered for their original identification. However, this may be argued to be only true if the existing consensus was correct as the interface may lead to negative performance when the consensus is incorrect or in the cases where there is no consensus (cf. Fig. 8).

3.3.3 Social. The Social interface led to a consistent increase in accuracy across each consensus type and overall outperformed the others (cf. Fig. 8D). In this interface, the types of comments left mainly concerned reasons behind participants changing their identification. In total, there were 28 comments left for the 18 Social images and only 2 out of them did not have any comment. However, the new information was only available after someone had commented on this interface. The types of comments centered around two themes.

a). Key features. This theme included comments mainly highlighting features that could help other participants with identifying the specimen (e.g. "Dark wings indicate cuckoo; single dark yellow band on abdomen just above tail suggests female Southern cuckoo {6369}"; "Clearly a cuckoo bumblebee due to absence of pollen baskets. Bombus vestalis due to yellow patch above white on abdomen {6880}."). Some participants also commented on the reasons why their identification may be correct compared to other options ("Two dark yellow bands and a dirty white tail means this is B. terrestris. If it were B. pratorum the tail would be much more orange/red. {6272} "; "…The three yellow bands seem too thick to be other than those of Garden or Ruderal bumblebee. It's not a Field Cuckoo - they don't have this pattern of three yellow bands and a white tail. I don't think it's a Barbut's as the bands are too thick. I went for Garden as Ruderal are rather rarer. I think the legs might be showing a pollen basket but can't really see. {532}"). The highlighting of features as well as comparison with other species was persuasive for other users for reviewing the identifications as was evident from the reasoning provided after using the Social interface through reasoning pop-ups (Fig. 5) ("Based on other's views - seems likely {1414}"; "Changed mind - agree with comment added by other user {6353}").

b). Contextual information. This theme included comments related to the contextual information which might affect the identification, such as image quality ("I thought the photo was poorly illuminated but could just see a band on the thorax and abdomen and orangey tail. {5347}"), visibility of features ("To identify this, I'd need to get a better view of the abdomen and the bottom of the thorax." {532}) and angle ("The face does look quite long at this angle, so I did wonder about it being a Garden bumblebee. However, at this angle, it is tricky to be sure if the face is long, and also

where the second yellow band is. Without a side or top view I am not sure of the id, so I am going with Not identifiable. {3091}").

In some cases, others seem to agree with the reasoning information that was present in the Social interface ("Dark head and thorax, slim build, reddish abdomen - this is a Red Mason bee, not a bumblebee {6369}"); "this is not a bumblebee, a discoloured (bleeched by the sun) red mason bee is correct." {6269}), while in other cases participants did not agree with the reasoning information ("To identify this, I'd need to get a better view." {532}; "The photo does not show it so Just looked at the main features available on the photo/photos I.e strips, colour and antenae - then look at the standard pictures given - and just see what fits best" {4082}).

Hence, using the Social interface, participants could provide reasoning behind their identifications as well as learn from others, including if they had overlooked certain features that might have led to different identifications. The comments left also suggested higher engagement with the overall process of identification. Using this interface, participants were willing to comment on the contextual information, such as image quality or angle of the bumblebee, that would be difficult to capture using other interfaces and may be relevant for identification. The interface promotes dialogic interactions between the participants by adding information and/or possibilities to view the same object from different angles; the information provided during these interactions enabled convergence of thinking, thus leading to improvements in accuracy and consensus building. However, lack of comments for some images as well information only available after someone left a comment may have limited the collaboration and interactions between participants. Nevertheless, the interface consistently enhanced engagement on the task as compared to other interfaces, with participants willing to revisit their existing identifications and utilising the new information to improve accuracy.

3.3.4 Natural Language Generation (NLG). The NLG interface improved consensus across all image types. The comments from participants also indicated how they may have used the information provided by the interface. The comments often focused on the features that were clearly visible in the image and may have directed attention of other participants towards those features, contributing to increase in accuracy. For example, Participant 5347, while identifying a Consensus Correct image commented: "Hairy ab. and that probably is pollen baskets and not just long hairs. Cannot tell if wings are clear or dark "; which highlights information from the NLG texts to focus on the relevant feature (pollen baskets) to change the identification. Similarly, for Consensus Not Correct image, Participant 6269 commented "Either garden or heath, but in one picture it looks like the face is not that long, so heath "; again describing the feature information that was used to change the identification.

The NLG texts were designed to provide a comparison of identification features of species, and the texts may arguably simulate a dialogic interaction that assisted participants to derive at better species identifications.

4 DISCUSSION

We investigated how participants involved in a citizen science project concerning on-line species identification responded to different forms of feedback, in order to better understand online collaboration and inform the design of toots that support peer-learning and improve task performance. To address this, we designed four collaboration interfaces implementing different feedback mechanisms within asynchronous, collaborative online interactions to support consensus-building. The interfaces, underpinned by collaboration literature and citizen science practice, allowed for studying the role of majority vote, user expertise, communication and sharing of resources through social interactions, and automatically generated texts representing expert knowledge.

Highlighting "level of agreement" in the form of visualisations is a widely utilised technique in online communities and is suggested as a persuasive method in literature [60] for supporting collaboration; however, we found it to have little effect on user accuracy and engagement in our study. Although the visualisation to communicate "level of agreement" may have helped users in monitoring and progressing on the identification task [40], it generally lacked persuasiveness due to the absence of any species identification knowledge or connection therewith (i.e. authority), thus preventing to reconsider the initial species identification made.

Validating prior literature [9, 46, 49, 88], user expertise influenced participants to carefully reconsider their choices, making it a persuasive method to support collaboration, even though this interface also lacked communication of species identification knowledge. The conformity effect [93] discussed in Section 1.2 may have been elicited by this interface, leading to increased accuracy when the consensus was correct but no enhanced accuracy (unlike for the three other interfaces) due to participants being driven towards the incorrect answer [61]. The fact that there was no decrease in accuracy when the majority vote was incorrect or when there was no consensus indicates that the conformity effect of the interface was not fully blinding: participants were more likely to conform with outcomes that seemed right than with those that seemed wrong or ambiguous. Thus, as participants utilised the identification key while reviewing their identifications, they also looked at reasoning behind reviewing their identifications, which limited social conforming effects and bias. This may not be the case in online communities where such keys or learning resources are not provided or readily accessible. Hence, usage of expert ratings to support collaboration should be approached with caution.

The Social interface, which promotes knowledge sharing and communication between members of the group, was most effective in improving user performance. The results provided evidence of participants being persuaded to review their observations in light of the comments posted by others, facilitating peer-learning [24]. Comments posted were anonymised to prevent any effects associated with the user's personal profile [17]. The interface with the goal setting "instructions" may have acted as a 'rules for interaction' for the group members [40, 99], leading to comments with detailed reasoning and preventing "general commenting" behaviour as is the case with many online communities. As this method of online collaboration is widely used in citizen science, be it in various ways, the results of this study may further inform the design of social collaboration for citizen science [59, 80, 85]. Results from the analysis of the Social interface also indicated that this kind of task-focused annotation appeared the most effective in motivating participants to review their initial submissions, deliberately compare features across species and types of images and thus simulate the experiential feed-back loops underpinning learning in the field and as part of a community of practice [27, 53].

The effect of AI in mediating the collaboration task using the NLG interface revealed that it was on average as effective as the user ratings interface in improving user performance. More importantly, however, the effect was consistent across all consensus types, showing that this technology can be a useful intervention for supporting social collaboration in citizen science. This is an important finding as this technology provides the same amount of information for every image and is not dependent on the participating user. The Social and NLG interfaces were found to also help with creating consensus in situations where there is no consensus, which make them suitable methods for building consensus in web-based citizen science. The distinct capabilities of the interfaces also highlight their potential for use in combination with supporting online collaboration. Some of these are already utilised across online communities such as Reddit and Stack overflow, where some of the comments and posts are uploaded by members of the community [24]. Additionally, citizen science platforms such as iSpot show the level of agreement on a species, the user ratings (called 'reputation' on iSpot), and allow for comments around the identifications, thus using a combination

of social, user ratings and distribution interfaces [80]. We haven't explored these dimensions within our study, but our results highlight the potential for utilizing machine-generated content to support and promote contributions from community members, for example, the NLG texts can be used to support and strengthen discussions initiated by citizens on a social platform.

All four feedback techniques provided insights into how collaboration can assist consensus building online while supporting peer-learning for citizen science, strengthening some of the findings from the collaboration literature [9, 26, 40, 51, 55]. The study also provides some novel insights, such as the limited effects of the Distribution interface, context specific effect of the User Ratings interface, and value of machine-generated texts for consensus building. Projects on citizen science platforms rely on independent validation of datasets from multiple users, which is an important criterion for scientific analysis and removing biases. An arguably yet more important finding, given how central data quality, learning and engagement are to online citizen science [43, 44, 47, 54], is what online collaboration can bring more generally. Strong and persistent concerns about data quality have driven many citizen science projects to seek independent validation routes [53, 59, 83, 87]. Whilst sensible, and required for scientific data analysis, this may lead to missed opportunities for collaboration and peer-learning. The results from our study show that careful collaboration design, such as the Social interface which provides a platform for knowledge sharing and communication, may help to improve scientific data quality as well as foster user-learning and engagement [59]. This is an important finding that (1) can support greater integration of citizen science in formal and informal science education contexts; and (2) can enable members of the public not only to contribute but also to potentially influence scientific research agendas with novel targets and questions that will emerge from shared experiences in their local contexts. This is a notable contribution, particularly for citizen science and policy-making practice in contexts which demand greater sensitivity to historical, linguistic and contextual dimensions of specific environmental or developmental problems [34].

Machine learning and AI algorithms are being explored increasingly for automated species identification, and for some species groups the algorithms are very efficient [22, 57]. Nevertheless, for noisy data and difficult species such as bumblebees AI performance is still inadequate, therefore requiring human expertise for data validation and emphasizing the need for training of volunteers through learning resources, such as the collaborative technologies explored in our study [33, 36]. More importantly, developing identification technologies may help in engaging the wider public around environmental issues, such as climate change and biodiversity loss, through positive citizen action [78]. Finally, our research has wider implications in multiple disciplines, including taxonomic research into developing novel species identification technologies vital for ecological and conservation activities [89] and the domains of HCI and Human-AI interaction for the development of AI-mediated collaborative learning environments. Future research can investigate how annotations from automated image identification can be incorporated into such online environments to support collaborative learning.

This study shows that collaborative interfaces can be used to help novices perform complex species identification tasks. Therefore, citizen science projects that provide such interfaces can facilitate novices in contributing valuable scientific information as well as acquiring scientific skills [77]. Importantly, our study corroborates the value of the socio-material frame to make sense of 'learning through feedback' [7, 68], within a system or assemblage which may include humans and non-human expertise. This is an important finding which suggests that the value of digital interfaces lies beyond their use as a novelty or repository of factual information, indeed by shifting emphasis from the passive acquisition of 'expert knowledge' to generating interest and motivation amongst participants. Significantly, a kind of 'hot function' that is related to social-affective engagement appears to be present due to the newly established social interactions and collaboration

(such as by commenting, supporting peers) on a cognitively demanding task (reviewing, utilising new information) which the participants (as citizen scientists) were intrinsically motivated to perform, leading to a common goal (building consensus) [32]. This emotional, social and cognitive engagement emerges as intrinsic to the process of learning, enabling participants to achieve the more immediate and practical goal of reaching the 'correct' identification. In addition, our findings also showed that for successful outcomes, computational devices supported together two aspects of the identification process: representational practices and relational practices. Representational practices included writing verbal descriptions, including contextual information and estimating shapes and sizes, converting one form of information (e.g. observation) into another (comparison with another known species), which were made available to other users as textual 'sketches'. Relational practices were visible as interactional exchanges mediated by simple language: the way in which team members communicated with one another, often re-elaborating complex information into accessible descriptions, influenced the level of engagement and collaboration.

This finding opens up exciting new avenues in citizen science research, looking at the integration of technology in the development of hybrid communities of practice, which can help bring together the 'visual precision' of the expert with more varied forms of encounter with nature [27], including contextual, aesthetic, affective and embodied features underpinning environmental conscious-ness [21]. Further research could focus on the design of social learning interfaces, supporting more extended field-based investigations, with the potential to widen participation and inclusion in citizen science initiatives in different cultural and geographical contexts, and by a variety of different groups.

5 CONCLUSIONS

The primary aim of the study reported in this article was to explore the use of different types of feedback within collaborative interfaces for building consensus on species identification tasks. Collaborative interfaces such as the ones studied here are ubiquitous in online communities. We report that interfaces which support logical reasoning for problem solving, such as the developed Social and NLG interfaces, are more effective than the ones which only display consensus and user expertise, and that the latter is context specific. We found the Social interface to be most effective, however, the user-comments may need to be goal-directed to foster meaningful outcomes. Additionally, we also found machine-mediated consensus building using NLG to bring value across different consensus types, highlighting the potential of this technology for consensus building.

6 ACKNOWLEDGEMENTS

REFERENCES

[1] 2021. iNaturalist. https://www.inaturalist.org/
[2] 2021. iSpot. https://www.ispotnature.org/
[3] Paul André, Robert E. Kraut, and Aniket Kittur. 2014. Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 139–148. https://doi.org/10.1145/2556288.2557158

[4] Koen Arts, René Van der Wal, and William M. Adams. 2015. Digital technology and the conservation of nature. *Ambio* 44, S4 (27 11 2015), 661–673. https://doi.org/10.1007/s13280-015-0705-1

[5] Tom August, Martin Harvey, Paula Lightfoot, David Kilbey, Timos Papadopoulos, and Paul Jepson. 2015. Emerging technologies for biological recording. *Biological Journal of the Linnean Society* 115, 3 (7 2015), 731–749. https://doi.org/10.1111/bij.12534

[6] Mikhail Bakhtin. 1984. *Problems of Dostoevsky's poetics (C. Emerson, Trans. C. Emerson Ed.).* University of Minnesota Press, Minneapolis.

[7] Gregory Bateson. 1979. *Mind and nature : a necessary unity.* Hampton Press, London.

[8] Melanie R Beck, Claudia Scarlata, Lucy F Fortson, Chris J Lintott, B D Simmons, Melanie A Galloway, Kyle W Willett, Hugh Dickinson, Karen L Masters, Philip J Marshall, and Darryl Wright. 2018. Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society* 476, 4 (1 6 2018), 5516–5534. https://doi.org/10.1093/MNRAS/STY503

[9] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E Kraut. 2004. Using Social Psychology to Motivate Contributions to Online Communities General Terms. *Proceedings of the 2004 ACM conference on Computer supported cooperative work - CSCW '04*, 212–221. https://doi.org/10.1145/1031607

[10] Györgyi Bela, Taru Peltola, Juliette C. Young, Bálint Balázs, Isabelle Arpin, György Pataki, Jennifer Hauck, Eszter Kelemen, Leena Kopperoinen, Ann Van Herzele, Hans Keune, Susanne Hecker, Monika Suškevičs, Helen E. Roy, Pekka Itkonen, Mart Külvik, Miklós László, Corina Basnou, Joan Pino, and Aletta Bonn. 2016. Learning and the transformative potential of citizen science. *Conservation Biology* 30, 5 (2016), 990–999. https://doi.org/10.1111/cobi.12762

[11] Tomas J. Bird, Amanda E. Bates, Jonathan S. Lefcheck, Nicole A. Hill, Russell J. Thomson, Graham J. Edgar, Rick D. Stuart-Smith, Simon Wotherspoon, Martin Krkosek, Jemina F. Stuart-Smith, Gretta T. Pecl, Neville Barrett, and Stewart Frusher. 2014. Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation* 173 (2014), 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

[12] S. Blake, A. Siddharthan, H. Nguyen, N. Sharma, A.-M. Robinson, E. O'mahony, B. Darvill, C. Mellish, and R. Van Der Wal. 2012. Natural language generation for nature conservation: Automating feedback to help volunteers identify bumblebee species. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers.*

[13] Rick Bonney, Caren B. Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, V. Kenneth Rosenberg, and Jennifer Shirk. 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience* 59, 11 (12 2009), 977–984. https://doi.org/10.1525/bio.2009.59.11.9

[14] Eleanor D. Brown and Byron K. Williams. 2019. The potential for citizen science to produce reliable and useful information in ecology. *Conservation Biology* 33, 3 (1 6 2019), 561–569. https://doi.org/10.1111/COBI.13223

[15] Matthias Budde, Andrea Schankin, Julien Hoffmann, Marcel Danz, Till Riedel, and Michael Beigl. 2017. Participatory Sensing or Participatory Nonsense? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (11 9 2017), 1–23. https://doi.org/10.1145/3131900

[16] Luigi Ceccaroni, James Bibby, Erin Roger, Paul Flemons, Katina Michael, Laura Fagan, and Jessica L. Oliver. 2019. Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. *Citizen Science: Theory and Practice* 4, 1 (28 11 2019). https://doi.org/10.5334/cstp.241

[17] Andrea Chester and Gillian Gwynne. 1998. Online Teaching: Encouraging Collaboration through Anonymity. *Journal of Computer-Mediated Communication* 4, 2 (1 12 1998). https://doi.org/10.1111/J.1083-6101.1998.TB00096.X

[18] Michelene T. H. Chi, Paul J. Feltovich, and Robert Glaser. 1981. Categorization and Representation of Physics Problems by Experts and Novices*. *Cognitive Science* 5, 2 (4 1981), 121–152. https://doi.org/10.1207/s15516709cog0502_2

[19] Jeffrey P. JP Cohn. 2008. Citizen Science: Can Volunteers Do Real Research? *BioScience* 58, 3 (2008), 192. https://doi.org/10.1641/B580303

[20] HM Collins and R Evans. 2002. The third wave of science studies: Studies of expertise and experience. *Social studies of science* (2002).

[21] Laura Colucci-Gray, Pamela Burnard, Donald Gray, and Carolyn Cooke. 2019. A Critical Review of STEAM (Science, Technology, Engineering, Arts, and Mathematics). *Oxford Research Encyclopedia of Education* (26 3 2019). https://doi.org/10.1093/ACREFORE/9780190264093.013.398

[22] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. 2018. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (14 12 2018), 4109–4118. https://doi.org/10.1109/CVPR.2018.00432

[23] Vickie Curtis. 2015. Motivation to Participate in an Online Citizen Science Game: A Study of Foldit. *Science Communication* 37, 6 (16 10 2015), 723–746. https://doi.org/10.1177/1075547015609322

[24] Marc Esteve Del Valle, Anatoliy Gruzd, Priya Kumar, and Sarah Gilbert. 2020. *Learning in the Wild: Understanding Networked Ties in Reddit.* Springer International Publishing, Cham, 51–68. https://doi.org/10.1007/978-3-030-36911-8_4

[25] Janis L. Dickinson, Benjamin Zuckerberg, and David N. Bonter. 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics* 41, 1 (12 2010), 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

[26] John R.G. Dyer, Christos C. Ioannou, Lesley J. Morrell, Darren P. Croft, Iain D. Couzin, Dean A. Waters, and Jens Krause. 2008. Consensus decision making in human crowds. *Animal Behaviour* 75, 2 (2 2008), 461–470. https://doi.org/10.1016/j.anbehav.2007.05.010

[27] Rebecca Ellis. 2011. Jizz and the joy of pattern recognition: Virtuosity, discipline and the agency of insight in UK naturalists' arts of seeing. *Social Studies of Science* 41, 6 (12 2011), 769–790. https://doi.org/10.1177/0306312711423432

[28] Elizabeth R. Ellwood, Betty A. Dunckel, Paul Flemons, Robert Guralnick, Gil Nelson, Greg Newman, Sarah Newman, Deborah Paul, Greg Riccardi, Nelson Rios, Katja C. Seltmann, and Austin R. Mast. 2015. Accelerating the Digitization of Biodiversity Research Specimens through Online Public Participation. *BioScience* 65, 4 (1 4 2015), 383–396. https://doi.org/10.1093/biosci/biv005

[29] Yrjö Engeström et al. 1999. Activity theory and individual and social transformation. *Perspectives on activity theory* 19, 38 (1999), 19–30.

[30] Andrea Forte, Nazanin Andalibi, and Rachel Greenstadt. 2017. Privacy, anonymity, and perceived risk in open collaboration: A study of tor users and wikipedians. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (25 2 2017), 1800–1811. https://doi.org/10.1145/2998181.2998273

[31] Steffen Fritz, Linda See, Tyler Carlson, Mordechai (Muki) Haklay, Jessie L Oliver, Dilek Fraisl, Rosy Mondardini, Martin Brocklehurst, Lea A Shanley, Sven Schade, Uta Wehn, Tommaso Abrate, Janet Anstee, Stephan Arnold, Matthew Billot, Jillian Campbell, Jessica Espey, Margaret Gold, Gerid Hager, Shan He, Libby Hepburn, Angel Hsu, Deborah Long, Joan Masó, Ian McCallum, Maina Muniafu, Inian Moorthy, Michael Obersteiner, Alison J Parker, Maike Weisspflug, and Sarah West. 2019. Citizen science and the United Nations Sustainable Development Goals. *Nature Sustainability* 2, 10 (2019), 922–930. https://doi.org/10.1038/s41893-019-0390-3

[32] Stuart Iain Gray, Judy Robertson, Andrew Manches, and Gnanathusharan Rajendran. 2019. BrainQuest: The use of motivational design theories to create a cognitive training game supporting hot executive function. *International Journal of Human-Computer Studies* 127 (1 7 2019), 124–149. https://doi.org/10.1016/J.IJHCS.2018.08.004

[33] Oskar L.P. Hansen, Jens Christian Svenning, Kent Olsen, Steen Dupont, Beulah H. Garner, Alexandros Iosifidis, Benjamin W. Price, and Toke T. Høye. 2020. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and Evolution* 10, 2 (jan 2020), 737–747. https://doi.org/10.1002/ECE3.5921

[34] Barbara Heinisch. 2020. Knowledge translation and its interrelation with usability and accessibility. Biocultural diversity translated by means of technology and language—the case of citizen science contributing to the sustainable development goals. *Sustainability* 13, 1 (2020), 54.

[35] Gayle W. Hill. 1982. Group versus individual performance: Are N?+?1 heads better than one? *Psychological Bulletin* 91 (1982), 517–539. https://doi.org/10.1037/0033-2909.91.3.517

[36] Toke T. Høye, Johanna Ärje, Kim Bjerge, Oskar L. P. Hansen, Alexandros Iosifidis, Florian Leese, Hjalte M. R. Mann, Kristian Meissner, Claus Melvad, and Jenni Raitoharju. 2021. Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences* 118, 2 (2021), e2002545117. https://doi.org/10.1073/pnas.2002545117

[37] Alan Irwin. 1995. *Citizen science: a study of people, expertise, and sustainable development.* Routledge.

[38] Nick J. B. Isaac, Arco J. Strien, Tom A. August, Marnix P. Zeeuw, and David B. Roy. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* 5, 10 (1 10 2014), 1052–1060. https://doi.org/10.1111/2041-210X.12254

[39] Corey Brian Jackson, Carsten Østerlund, Kevin Crowston, Mahboobeh Harandi, and Laura Trouille. 2020. Shifting forms of Engagement: Volunteer Learning in Online Citizen Science. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020). https://doi.org/10.1145/3392841

[40] Heisawn Jeong and Cindy E. Hmelo-Silver. 2016. Seven Affordances of Computer-Supported Collaborative Learning: How to Support Collaborative Learning? How Can Technologies Help? *Educational Psychologist* 51, 2 (2 4 2016), 247–265. https://doi.org/10.1080/00461520.2016.1158654

[41] Aditya Johri. 2011. The socio-materiality of learning practices and implications for the field of learning technology. *Research in Learning Technology* 19, 3 (2011), 207–217.

[42] Benjamin L. Jones, Richard K.F. Unsworth, Len J. McKenzie, Rudi L. Yoshida, and Leanne C. Cullen-Unsworth. 2018. Crowdsourcing conservation: The role of citizen science in securing a future for seagrass. *Marine Pollution Bulletin* 134 (1 9 2018), 210–215. https://doi.org/10.1016/J.MARPOLBUL.2017.11.005

[43] Dick Kasperowski and Niclas Hagen. 2022. Making particularity travel: Trust and citizen science data in Swedish environmental governance. *Social studies of science* 52, 3 (apr 2022), 447–462. https://doi.org/10.1177/03063127221085241

[44] Steve Kelling, Daniel Fink, Frank A. La Sorte, Alison Johnston, Nicholas E. Bruns, and Wesley M. Hochachka. 2015. Taking a 'Big Data' approach to data quality in a citizen science project. *Ambio* 44 (1 11 2015), 601–611. https://doi.org/10.1007/S13280-015-0710-4

[45] Norbert L Kerr and R Scott Tindale. 2004. Group performance and decision making. *Annual review of psychology* 55 (2004), 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

[46] Vasily Klucharev, Ale Smidts, and Guillén Fernández. 2008. Brain mechanisms of persuasion: how 'expert power' modulates memory and attitudes. *Social Cognitive and Affective Neuroscience* 3, 4 (12 2008), 353–366. https://doi.org/10.1093/scan/nsn022

[47] Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment* 14, 10 (1 12 2016), 551–560. https://doi.org/10.1002/FEE.1436

[48] Christopher Kullenberg and Dick Kasperowski. 2016. What Is Citizen Science? – A Scientometric Meta-Analysis. *PLOS ONE* 11, 1 (14 1 2016), e0147152. https://doi.org/10.1371/journal.pone.0147152

[49] Joseph Lampel and Ajay Bhalla. 2007. The Role of Status Seeking in Online Communities: Giving the Gift of Experience. *Journal of Computer-Mediated Communication* 12, 2 (1 2007), 434–455. https://doi.org/10.1111/j.1083-6101.2007.00332.x

[50] J Larkin, J McDermott, DP Simon, and HA Simon. 1980. Expert and novice performance in solving physics problems. *Science* (1980).

[51] Patrick R. Laughlin, Bryan L. Bonner, and Andrew G. Miner. 2002. Groups perform better than the best individuals on Letters-to-Numbers problems. *Organizational Behavior and Human Decision Processes* 88, 2 (7 2002), 605–620. https://doi.org/10.1016/S0749-5978(02)00003-1

[52] Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation.* Cambridge university press.

[53] Hjalti Leifsson and Jóhann Örn Bjarkason. 2015. *Project Discovery Advancing scientic research by implementing citizen science in EVE Online Supervisor.* Ph. D. Dissertation. Reykjavik University.

[54] Eva Lewandowski and Hannah Specht. 2015. Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology* 29, 3 (1 6 2015), 713–723. https://doi.org/10.1111/COBI.12481

[55] D. W. Liang, R. Moreland, and L. Argote. 1995. Group Versus Individual Training and Group Performance: The Mediating Role of Transactive Memory. *Personality and Social Psychology Bulletin* 21, 4 (1 4 1995), 384–393. https://doi.org/10.1177/0146167295214009

[56] Yu Pin Lin, Dongpo Deng, Wei Chih Lin, Rob Lemmens, Neville D. Crossman, Klaus Henle, and Dirk S. Schmeller. 2015. Uncertainty analysis of crowd-sourced and professionally collected field data used in species distribution models of Taiwanese moths. *Biological Conservation* 181 (1 1 2015), 102–110. https://doi.org/10.1016/J.BIOCON.2014.11.012

[57] Marcelo T. Lopes, Lucas L. Gioppo, Thiago T. Higushi, Celso A.A. Kaestner, Carlos N. Silla, and Alessandro L. Koerich. 2011. Automatic bird species identification for large number of species. *Proceedings - 2011 IEEE InternationalSymposium on Multimedia, ISM 2011* (2011), 117–122. https://doi.org/10.1109/ISM.2011.27

[58] Y. Lou, P. C. Abrami, J. C. Spence, C. Poulsen, B. Chambers, and S. d'Apollonia. 1996. Within-Class Grouping: A Meta-Analysis. *Review of Educational Research* 66, 4 (1 1 1996), 423–458. https://doi.org/10.3102/00346543066004423

[59] M Luczak-Roesch, R Tinati, E Simperl, and Van M Kleek. 2014. Why Won't Aliens Talk to Us? Content and Community Dynamics in Online Citizen Science. *ICWSM* (2014).

[60] Robin Martin, Antonis Gardikiotis, and Miles Hewstone. 2002. Levels of consensus and majority and minority influence. *European Journal of Social Psychology* 32, 5 (9 2002), 645–665. https://doi.org/10.1002/ejsp.113

[61] Serge Moscovici and Willem Doise. 1994. *Conflict and consensus: A general theory of collective decisions.* Sage.

[62] Gabriel Mugar, Carsten Østerlund, Katie Devries Hassman, Kevin Crowston, and Corey Brian Jackson. 2014. Planet hunters and seafloor explorers: Legitimate peripheral participation through practice proxies in online citizen science. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2014), 109–119. https://doi.org/10.1145/2531602.2531721

[63] CJ Nemeth and J Wachtler. 1983. Creative problem solving as a result of majority vs minority influence. (1983).

[64] Jill Nugent. 2018. INaturalist. *Science Scope* 41, 7 (2018), 12–13.

[65] Takeshi Okada and Herbert A. Simon. 1997. Collaborative Discovery in a Scientific Domain. *Cognitive Science* 21, 2 (11 4 1997), 109–146. https://doi.org/10.1207/s15516709cog2102_1

[66] Victoria Palacin, Maria Angela Ferrario, Gary Hsieh, Antti Knutas, Annika Wolff, and Jari Porras. 2021. Human values and digital citizen science interactions. *International Journal of Human-Computer Studies* 149 (may 2021), 102605. https://doi.org/10.1016/J.IJHCS.2021.102605

[67] Lorenzo Palamenghi, Serena Barello, Stefania Boccia, and Guendalina Graffigna. 2020. Mistrust in biomedical research and vaccine hesitancy: the forefront challenge in the battle against COVID-19 in Italy. *European journal of epidemiology* 35, 8 (2020), 785–788.

[68] Parva Panahi, Parviz Birjandi, and Behrooz Azabdaftari. 2013. Toward a sociocultural approach to feedback provision in L2 writing classrooms: the alignment of dynamic assessment and teacher error feedback. *Language Testing in Asia*

3, 1 (1 12 2013). https://doi.org/10.1186/2229-0443-3-13

[69] Roy D Pea. 1993. Learning scientific concepts through material and social activities: Conversational analysis meets conceptual change. *Educational psychologist* 28, 3 (1993), 265–277.

[70] Roy D Pea. 1993. Practices of distributed intelligence and designs for education. *Distributed cognitions: Psychological and educational considerations* 11 (1993), 47–87.

[71] Jennifer Preece. 2016. Citizen Science: New Research Challenges for Human–Computer Interaction. *International Journal of Human-Computer Interaction* 32, 8 (2016), 585–612. https://doi.org/10.1080/10447318.2016.1194153

[72] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[73] Mark S. Reed, Anna C. Evely, Georgina Cundill, Ioan Fazey, Jayne Glass, Adele Laing, Jens Newig, Brad Parrish, Christina Prell, Chris Raymond, and Lindsay C. Stringer. 2010. What is Social Learning? *Ecology and Society* 15, 4 (2010). http://www.jstor.org/stable/26268235

[74] Diana Rose, Danielle Efraim, Marie-Claude Gervais, Helene Joffe, Sandra Jovchelovitch, and Nicola Morant. 1995. Questioning consensus in social representation theory. *Papers on social representations* 4 (1995), 150–176.

[75] Arthur D. Santana. 2014. Virtuous or Vitriolic. *Journalism Practice* 8, 1 (2 1 2014), 18–33. https://doi.org/10.1080/17512786.2013.813194

[76] Alan H. Schoenfeld and Douglas J. Herrmann. 1982. Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8, 5 (1982), 484–494. https://doi.org/10.1037/0278-7393.8.5.484

[77] Nirwan Sharma, Laura Colucci-Gray, Advaith Siddharthan, Richard Comont, and René Van Der Wal. 2019. Designing online species identification tools for biological recording: The impact on data quality and citizen science learning. *PeerJ* 2019, 1 (28 1 2019), e5965. https://doi.org/10.7717/peerj.5965

[78] Nirwan Sharma, Sam Greaves, Advaith Siddharthan, Helen B. Anderson, Annie Robinson, Laura Colucci-Gray, Agung Toto Wibowo, Helen Bostock, Andrew Salisbury, Stuart Roberts, David Slawson, and René van der Wal. 2019. From citizen science to citizen action: Analysing the potential for a digital platform to cultivate attachments to nature. *Journal of Science Communication* 18, 1 (2019). https://doi.org/10.22323/2.18010207

[79] Advaith Siddharthan, Christopher Lambin, Anne-Marie Robinson, Nirwan Sharma, Richard Comont, Elaine O'mahony, Chris Mellish, and Van Der René Wal. 2016. Crowdsourcing Without a Crowd. *ACM Transactions on Intelligent Systems and Technology* 7, 4 (5 5 2016), 1–20. https://doi.org/10.1145/2776896

[80] Jonathan Silvertown, Martin Harvey, Richard Greenwood, Mike Dodd, Jon Rosewell, Tony Rebelo, Janice Ansine, and Kevin McConway. 2015. Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys* 480, 480 (2 1 2015), 125–46. https://doi.org/10.3897/zookeys.480.8803

[81] Gerry Stahl, Ulrike Cress, Sten Ludvigsen, Nancy Law, G Stahl, U Cress, S Ludvigsen, and N Law. 2014. Dialogic foundations of CSCL. *International Journal of Computer-Supported Collaborative Learning 2014 9:2* 9, 2 (may 2014), 117–125. https://doi.org/10.1007/S11412-014-9194-7

[82] Jonathan Steinke, van Jacob Etten, and Pablo Mejía Zelan. 2017. The accuracy of farmer-generated data in an agricultural citizen science methodology. *Agronomy for Sustainable Development 2017 37:4* 37, 4 (24 7 2017), 1–12. https://doi.org/10.1007/S13593-017-0441-Y

[83] Alexandra Swanson, Margaret Kosmala, Chris Lintott, and Craig Packer. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology* 30, 3 (6 2016), 520–531. https://doi.org/10.1111/cobi.12695

[84] 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. 2015. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation* 181 (2015), 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

[85] Ramine Tinati, Elena Simperl, and Markus Luczak-Roesch. 2017. To help or hinder: Real-time chat in citizen science. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 270–279.

[86] Stefan Trausan-Matu, Rupert Wegerif, and Louis Major. 2021. Dialogism. (2021), 219–239. https://doi.org/10.1007/978-3-030-65291-3_12

[87] René Van der Wal, Nirwan Sharma, Chris Mellish, Annie Robinson, and Advaith Siddharthan. 2016. The role of automated feedback in training and retaining biological recorders for citizen science. *Conservation Biology* 30, 3 (1 6 2016), 550–561. https://doi.org/10.1111/COBI.12705

[88] Ivar E. Vermeulen and Daphne Seegers. 2009. Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management* 30, 1 (2009), 123–127. https://doi.org/10.1016/j.tourman.2008.04.008

[89] David Evans Walter and Shaun Winterton. 2007. Keys and the crisis in taxonomy: extinction or reinvention? *Annual review of entomology* 52 (12 1 2007), 193–208. https://doi.org/10.1146/annurev.ento.51.110104.151054

[90] Guiming Wang. 2019. Machine learning for inferring animal behavior from location and movement data. *Ecological Informatics* 49 (1 1 2019), 69–76. https://doi.org/10.1016/J.ECOINF.2018.12.002

[91] Etienne Wenger. 1999. *Communities of practice: Learning, meaning, and identity.* Cambridge university press.

[92] Sarah West and Rachel Pateman. 2016. Recruiting and Retaining Participants in Citizen Science: What Can Be Learned from the Volunteering Literature? *Citizen Science: Theory and Practice* 1, 2 (31 12 2016), 15. https://doi.org/10.5334/cstp.8

[93] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Impact of contextual and personal determinants on online social conformity. *Computers in Human Behavior* 108 (1 7 2020), 106302. https://doi.org/10.1016/J.CHB.2020.106302

[94] J C Woodcock, A Greenhill, K Holmes, G Graham, J Cox, E Y Oh, and K Masters. 2017. Crowdsourcing citizen science: exploring the tensions between paid professionals and users. *Journal of Peer Production* 10 (2017).

[95] Gamze Yilmaz and Reef Youngreen. 2016. The Application of Minority Influence Theory in Computer-Mediated Communication Groups. *Small Group Research* 47, 6 (12 2016), 692–719. https://doi.org/10.1177/1046496416661033

[96] Lixiu Yu, Paul André, Aniket Kittur, and Robert Kraut. 2014. A comparison of social, learning, and financial strategies on crowd engagement and output quality. *Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing - CSCW'14* (2014), 967–978. https://doi.org/10.1145/2531602.2531729

[97] Lixiu Yu and V. Jeffrey Nickerson. 2011. Cooks or cobblers? *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 1393. https://doi.org/10.1145/1978942.1979147

[98] Dana L Zeidler. 1997. The central role of fallacious thinking in science education. *Science Education* 81, 4 (1997), 483–496.

[99] Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Organizing without Formal Organization: Group Identification, Goal Setting and Social Modeling in Directing Online Production. *CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 935–944. https://doi.org/10.1145/2145204.2145344

[100] B J Zimmerman. 2001. Social Learning, Cognition, and Personality Development. In *International Encyclopedia of the Social Behavioral Sciences*, Neil J Smelser and Paul B Baltes (Eds.). Pergamon, Oxford, 14341–14345. https://doi.org/10.1016/B0-08-043076-7/01765-4

A APPENDICES

A.1 Email text for participation

Dear BeeWatch user,

We would like to invite you to participate in an online study on the BeeWatch website, which is part of ongoing research which focuses on collaboratively building consensus of species identifications. Participation to this study is voluntary and you can withdraw any time. All the data you provide will be anonymised and your identities won't be disclosed to anyone outside the research team.

We have selected a few images for which there is lack of agreement among BeeWatch users, and our records indicate that one or more of these images were identified by you.

We are investigating ways to improve consensus among users in order to get a reliable identification for the difficult photos. When you click on the link below you will be redirected to an experiment webpage, which shows a list of the image(s) that you had already identified and where a consensus has not been reached.

When you click on the individual images on the experiment webpage, you will be presented with a different interface, which provides information about how other BeeWatch users identified the image as well as options to review your identification.

If you want to change your identification based on the new information, you can click Review and you will be redirected to the identification tool (that you have already used) where you can change your identification. If you want to keep your existing identification, please click Don't Review. If you have any further questions related to this experiment, you can contact us by replying to this email. Experiment Link http://homepages.abdn.ac.uk/wpn003/beewatch/index.php?r=image/identifiedCs

Best wishes
The BeeWatch team
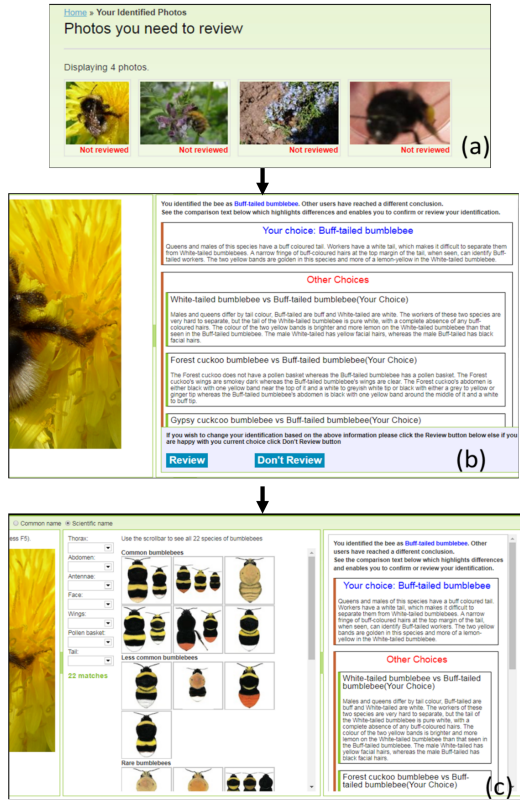
A.2   User workflow for the Review process

Fig. 9. This image shows an example workflow of the user for the review process using one of the interfaces. a) User is shown the images allocated for them to review, b) When the user clicks an image its associated collaboration interface is shown and c) When the user clicks 'Review' the species identification tools is shown together with the collaboration information.