Consensus building in on-line citizen science

Conference or Workshop Item

How to cite:

For guidance on citations see FAQs.

© 2022 Association for Computing Machinery

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Accepted Manuscript

Link(s) to article on publisher’s website:
https://dl.acm.org/journal/pacmhci

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
Consensus building in on-line citizen science.

ANONYMOUS AUTHOR(S)

A number of initiatives invite members of the public to perform online classification tasks such as identifying objects or wildlife 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 the 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 in 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 highlighted 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

ACM Reference Format:

1 INTRODUCTION

Citizen Science and crowdsourcing projects focus on using the capabilities of paid or unpaid volunteers for data collection and annotation [3, 76, 91]. 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 [41, 44, 51]. 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 [76]. A common crowdsourcing task is the classification of data captured through images using project resources [37] such as identification guides or keys [74, 80]. 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. 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 for collaboration that allow volunteers to collaborate on the classification task 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 be all available at the same time.

1.1 Background

The usefulness of citizen science for research has been widely highlighted [19, 45, 68], and citizen science practice has benefitted from advances in digital technologies, such as internet and mobile communication [4, 5, 41]. Its growing contribution to and impact on research in diverse ways is visible, ranging from monitoring the environment and biodiversity, to promoting question-driven research and statistical innovations in the handling of variable data sets [11, 25, 36, 40]. 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 [45, 81]. Data quality is a common concern for these projects and most include safeguarding mechanisms for data validation [41], 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, 74, 79, 84].

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, 34, 52, 55, 60, 63, 78, 94]. With the growth of the internet and subsequent rise of online communities such as Wikis, social media websites, citizen science and crowdsourcing platforms and ecommerce websites, 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 [34, 52], but typically lower than the best member [42]. Identifying the best member or utilising the capabilities of the 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 [48]. Research has also highlighted benefits of collaboration in terms of learning, motivation and engagement which, in turn, lead to sustained focus and deeper learning [9, 55]. Collaboration and teamwork have been a regular practice in the scientific domain, and scientific discoveries have often been made through collaboration [62]. 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, provides an effective method for studying interaction and communication between participants mediated using computers, especially for tasks which require debate, negotiation, and coordination among a group [6, 83]. 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 construct meaning or make sense of a new concept[78]. And more widely, how feedback supporting collaboration in citizen science activities can support social learning and civic participation more broadly, enabling a wider range of contexts and experiences to contribute to shaping research agendas [10, 31, 64]. We utilise the concept of feedback 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, 38, 96]. 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 [38]. In the context of building a group consensus through convergence of opinions, highlighting the level of agreement within a group may act as a persuasive method towards taking a particular course of action as individuals use that information as feedback to narrow down on a set of options. This method is used with success in the commercial sector to influence consumer choices, and literature suggests that revealing majority ‘votes’ [60] and levels of consensus [57] can influence other group members in problem-solving contexts. Project Discovery also utilizes this method, where community consensus is provided as a feedback for classification tasks without an expert annotation [50]. However, it may be limiting to equate consensus simply with agreement, uniformity or homogeneity [71]. 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 in with the expectations of the individual or the group [95].

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, 43, 85]. 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 and even citizen
Science [46]. 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 informing the user ratings method and expertise-led consensus informing the level of agreement are widely utilized in online communities for building consensus; however, both these methods may also lead to a conformity effect whereby members of the group may agree with a majority group who may sometimes be incorrect [58, 90] due to the power of influence exerted by the group or by one of the members perceived as being more knowledgeable [58].

Online communities also have means to enable anonymous collaboration and this technique is commonly utilized by many users, for example, in social networking websites for maintaining 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 the contributing users [17, 30, 72].

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 [47] but also how problems may be approached [18, 73]. 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, 73]. But a second level of social learning may also entail increased levels of self-regulation and self-efficacy in learners [97]. For example, any form of collaboration which enables communication and sharing of resources (cognitive or technological) among the members of a group (even providing chat boxes) can positively affect attention and engagement. By expanding the number of possible feedbacks on a task increases the possibilities for individuals to observe the effects on a product or a course of action, thus improving the quality of the work [38, 93]. 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, 94]). 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], it follows, then, in citizen science processes as well as in the broader area of social representation, we can distinguish between (1) the level of the generalised consensus in the scientific community, which make 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 ideas 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 [39, 70]. This understanding of social learning is most closely related to the ideas of situated learning [49, 88], distributed cognition [66, 67], 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 [49]. Citizen science projects such as iNaturalist [61] and iSpot [77] rely on creating communities of
nature enthusiasts uploading photographs of plant and animal species as well as identifying photos shared by other members of the community. More specifically, they rely on multiple annotations for producing reliable data, highlight expertise of the members using user ratings (a method to highlight individual uniqueness) and 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 [77]. 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 [50, 56, 82, 84], 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 [56]. Zooniverse users utilize the 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 [49, 59]. However, these forums are not directly linked to the independent classification models and thus might limit opportunities for social agency [39] 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, which are important for Citizen Science practice [13, 35, 89]. Hence, the objective here is to look more closely at how identification tasks derive their meaning of ‘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 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 is novel and deploys an AI to mediate the task.

In addition to the design 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 in 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 and how they affect performance, learning and engagement within online citizen science communities.

2.1 Dataset

We used data from the citizen science platform BeeWatch (www.abdn.ac.uk/research/beewatch), which provides participants with the option to submit images of bumblebees online as well as to independently identify images submitted by fellow BeeWatch participants through crowdsourcing [84].

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 the identification difficulty for different species [76], 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, thus 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 the 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 [76] - 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 it 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. 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 these same participants could review their identifications in the light of new information generated from the independent peer
annotations from the first step. We focused on four different types of on-line collaboration interfaces described below, drawn from both literature and practice, as summarised in Sections 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 for the experimental procedure. Each interface implements one of four types of feedback, which impact on 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[50, 57, 60]. 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 [92]. 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 [95].

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 differentiated to prevent comparison of their own expertise with others. The identities of participants were anonymized and two ratings were shown as two 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 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 sound/mobile signal etc).

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...
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).

of a group [9, 43, 46, 77, 85]. 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, 73].

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, 38, 93, 94]. 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 [61, 77]. 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, angle of the specimen, etc.) 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 focussing on the features that may help their peers in 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 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, 39, 97].

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.
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.

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 for that species. Then the interface displays comparison NLG texts for all other species submitted by other users.

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 the potential differences in visual features between the independent identifications by participants. This machine-generated text for 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, 84]. In short, when a citizen 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 citizen 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 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 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, 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 focussed on feedback in social learning, this fourth interface focussed 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 [39].

2.4 Procedure

From previous experience with studies around user performance with BeeWatch data, we sought at least 50 participants for the study and 15 images each for the four interfaces.

For this purpose, 72 photographs from BeeWatch 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 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).

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.

Fig. 5. Popups for collecting reasoning information when 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 even after reviewing the popup also 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 where the user can either select that their existing identification was correct or provide the reasons using free text.

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. All the interfaces had almost equal distribution of responses between them (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 [69]. New accuracy was fitted using R’s base glm function with old accuracy, interface type, consensus type, interfaces used and the interaction of interface type and consensus type where interfaces used was the number
of interfaces each participant used during the experiment and the Anova function from the car package was then used to test for significance. To test interface effect by consensus type hypothesis testing of count data (change in accuracy by interface) was done using the fisher.test() for CC, CNC and NC consensus types 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 species.

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 user) 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 for these cases. Figure 6 shows the percentage of identification 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.

The results in Fig. 7 show that the level of consensus has changed from before to after collaboration for all the 72 images across all four interface types. Consensus was defined as proportion of the group (between 0 and 1) that selected the most-selected species, and was different from accuracy, where 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 low, and where consensus was already high, reviewing frequently reduced the level of consensus.

### 3.2 Effect of interface and consensus type

![Fig. 8. Bar Plots showing mean percentage 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). 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.](image)

Significant variation in new accuracy was explained from the effect of old accuracy ($\chi^2 = 158.4$, df = 1, p-value < 0.001), Interface type ($\chi^2 = 7.67$, df = 3, p-value = 0.053) and Consensus type ($\chi^2 = 27.26$, df = 2, p-value < 0.001), while the number of interfaces used 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. 8), indicating that the Social interface design led to the largest average increase in accuracy.

When taking Consensus type into consideration, it was found that when the initial consensus towards the correct species (Fig. 8A), the increase in accuracy differed significantly by interface (p-value < 0.05), suggesting the usefulness of the User Ratings interface for these images that enabled participants to review their identifications. Where consensus was towards the incorrect species however, no significant difference of interface type was found (Fig. 8B), though Social and
NLG performed numerically better. Finally, for images where there was no consensus towards a particular species (Fig. 8C), differences across interface type were again significant (p-value < 0.01). The User ratings only improved accuracy for the Consensus correct images and had no improvements in accuracy for other image types highlighting the potential negative effect it may have in such cases. Overall, the Social type achieved greatest accuracy, followed by the NLG interface, contrasting sharply with the clear lack of the effect of User Ratings and Distribution interfaces.

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 smoky 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 (Fig. 8). While this interfaces gives information about the existing disagreement among the group, it offers limited opportunities to engage in a dialogic interaction with the group/interface [6, 83]. In line with the theorisation of feedback via persuasion [58, 90], 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 thus also offering limited impetus to the 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 highly persuasive for participants to review their identifications which led them to carefully 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 which were 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 (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 type of comments that were 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 in the identification as requested by the interface (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 (bleached by the sun) red mason bee is correct.” {6269}), while in other they did not agree with the reasoning information (“To identify this, I’d need to get a better view of the abdomen and the bottom of the thorax.” {532}); “The photo does not show it so Just looked at the main features available on the photo/photos I.e strips, colour and antennae - 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 if they had overlooked any features, which might have led to different identifications. The comments left also suggested higher engagement with the overall process of identification. Using this interface, participants were also 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 in 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 and user learning.

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 focussed 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 highlighted that the participant may have used 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.

This highlights that participants may have used the reasoning information from the NLG texts to change their identifications. The NLG texts were designed to provide a comparison of the differences in features of species and the texts may be argued to simulate a dialogic interaction through divergence of feature analysis in order to identify the correct species. This may have helped participants to focus attention on the features which are relevant to compare their choice from the group and change their identification, if needed, thereby leading to convergence of thinking between the group. The overall increase in accuracy associated with using the NLG interface was comparable to that of the User Ratings interface, however neither matched the level of increase in accuracy associated with the Social interface. The difference in Social and NLG may be because the NLG interface can only provide information on differences across species without considering specific contextual aspects (e.g. angle of the specimen, image quality, features visible on the photographs) whilst the Social interface can.

4 **DISCUSSION**

In this article, we wanted to investigate how participants involved in citizen science and online species identification projects respond to different forms of feedback in online collaboration, designed to 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, derived from collaboration literature and citizen science practice, studied the role of majority vote, user expertise, communication and sharing of resources through social interactions and automatically generated texts proxying expert knowledge.

Highlighting “level of agreement” in the form of visualisations is widely utilised as a technique in online communities and is suggested as a persuasive method in literature [57] for supporting collaboration; however, we found it had very limited affect in improving 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 [38], the participants still didn’t find this information persuasive enough to reconsider their choice as it didn’t communicate any species
identification knowledge. Validating prior literature \[9, 43, 46, 85\], 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 \[90\] discussed in Section 1.2 may have been elicited from this interface as when the existing consensus was correct accuracy and consensus increased, however, when the consensus was incorrect or there was no consensus among the crowd, participants were biased towards the incorrect answer \[58\]. Nevertheless, as users utilised the identification key while reviewing their identifications, they also had to look at reasoning behind reviewing their identifications which may have limited social conforming effects and bias. This may not be the case in online communities where such keys or learning resources don’t exist or are not easily 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 also 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 \[38, 96\] leading to comments with detailed reasoning and preventing “general commenting” behaviour as is the case with 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 \[56, 77, 82\]. Results from the analysis of the Social interface also indicated that this kind of task-focussed 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 \[27, 50\].

The effect of AI in mediating the collaboration task using the NLG interface revealed that it was as effective as the user ratings interface in improving user performance. More importantly the effect was consistent across all consensus types showing that this technology can be a useful intervention for supporting social collaboration in citizen science. Though the increases in accuracy of the NLG interface were not as high as those resulting from the Social interface, they were quite consistent across all consensus types. 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. Social and NLG interfaces especially help with creating consensus in situations where there is either no consensus or the consensus is towards the wrong answer, which make them suitable methods for building consensus in web-based citizen science. The distinct capabilities of the interfaces also highlight their potential to be used in combination for 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 \[77\]. 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 collaboration 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 \[9, 26, 38, 48, 52\]. The study also provides some novel insights such as the limited and in some cases negative effects of the Distribution and User Ratings interfaces.
and understandings into the role 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. However, such independent validation may also lead to missed opportunities for collaboration and peer-learning, and 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 [56]. 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 in the novel area of citizen science and policy-making in International development contexts which demand greater sensitivity to historical, linguistic and contextual dimensions of specific environmental problems [33].

Machine learning and AI algorithms are being explored increasingly for automated species identification and for some species groups the algorithms are very efficient [22, 54]. 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. 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 [75]. Finally, researching these technologies and the results from our study have wider implications in multiple disciplines, including taxonomic research into developing novel species identification technologies vital for ecological and conservation activities [86] 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 learning 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 [74]. Importantly, our study corroborates the value of the socio-material frame to make sense of ‘learning through feedback’ [7, 65], 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 consciousness [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 prone to biases. 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

We thank the many thousands of citizen scientists who submitted data to BeeWatch and the bumblebee experts who helped us to verify these. We are particularly grateful for the contributions of the participants in this study and to the referees for their insightful comments. This work was supported by funds from grant no. XXXX.

REFERENCES


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.