Conservation Biology

Special Section: *Moving from Citizen to Civic Science to Address Wicked Conservation Problems*

The role of automated feedback in training and retaining biological recorders for citizen science

René van der Wal,* ¶ Nirwan Sharma,† Chris Mellish,† Annie Robinson,‡ and Advaith Siddharthan†

*Aberdeen Centre for Environmental Sustainability, School of Biological Sciences, University of Aberdeen, 23 St. Machar Drive, Aberdeen, AB24 3UU, U.K.

†Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE, U.K.

‡dot.rural Digital Economy Hub, University of Aberdeen, Aberdeen, AB24 5UA, U.K.

Abstract: *The rapid rise of citizen science, with lay people forming often extensive biodiversity sensor networks, is seen as a solution to the mismatch between data demand and supply while simultaneously engaging citizens with environmental topics. However, citizen science recording schemes require careful consideration of how to motivate, train, and retain volunteers. We evaluated a novel computing science framework that allowed for the automated generation of feedback to citizen scientists using natural language generation (NLG) technology. We worked with a photo-based citizen science program in which users also volunteer species identification aided by an online key. Feedback is provided after photo (and identification) submission and is aimed to improve volunteer species identification skills and to enhance volunteer experience and retention. To assess the utility of NLG feedback, we conducted two experiments with novices to assess short-term (single session) and longer-term (5 sessions in 2 months) learning, respectively. Participants identified a specimen in a series of photos. One group received only the correct answer after each identification, and the other group received the correct answer and NLG feedback explaining reasons for misidentification and highlighting key features that facilitate correct identification. We then developed an identification training tool with NLG feedback as part of the citizen science program BeeWatch and analyzed learning by users. Finally, we implemented NLG feedback in the live program and evaluated this by randomly allocating all BeeWatch users to treatment groups that received different types of feedback upon identification submission. After 6 months separate surveys were sent out to assess whether views on the citizen science program and its feedback differed among the groups. Identification accuracy and retention of novices were higher for those who received automated feedback than for those who received only confirmation of the correct identification without explanation. The value of NLG feedback in the live program, captured through questionnaires and evaluation of the online photo-based training tool, likewise showed that the automated generation of informative feedback fostered learning and volunteer engagement and thus paves the way for productive and long-lived citizen science projects.*

Keywords: biological recording, bumblebee identification, natural language generation, training, volunteer motivation and retention

El Papel de la Retroalimentación Automatizada en el Entrenamiento y en la Retención de Registradores Biológicos para la Ciencia Ciudadana

Resumen: *El rápido crecimiento de la ciencia ciudadana, generalmente con personas laicas formando redes extensas de sensores de la biodiversidad, es visto como una solución a la disparidad entre la demanda y el suministro de datos, a la vez que compromete a los ciudadanos con temas ambientales. Sin embargo, los*

¶*email: r.vanderwal@abdn.ac.uk*

Paper submitted March 20, 2015; revised manuscript accepted July 29, 2015.

DOI: 10.1111/cobi.12705

esquemas de registro de la ciencia ciudadana requieren de consideraciones cuidadosas sobre cómo motivar, entrenar y retener a los voluntarios. Evaluamos un novedoso marco de trabajo científico y computacional que permitió la generación automatizada de retroalimentación para los ciudadanos científicos que usan tecnología de generación de lenguaje natural (GLN). Trabajamos con un programa de ciencia ciudadana basado en fotografías en el cual los usuarios también ofrecen identificación de especies con ayuda de una clave en línea. La retroalimentación es proporcionada después de presentar (e identificar) la fotografía y tiene como objetivo el mejoramiento de las habilidades de identificación de los voluntarios y el aumento en la experiencia y retención de voluntarios. Para evaluar la utilidad de la retroalimentación de GLN llevamos a cabo experimentos con novatos para así poder evaluar el aprendizaje a corto (sesión única) y a largo plazo (cinco sesiones en dos meses), respectivamente. Los participantes identificaron especímenes en una serie de fotos. Un grupo recibió solamente la respuesta correcta después de cada identificación, mientras que el otro grupo recibió la respuesta correcta además de la retroalimentación de GLN, la cual explica las razones por las que se identifica erróneamente y resalta los caracteres clave que facilitan la identificación correcta. Después desarrollamos una herramienta para el entrenamiento en la identificación con la retroalimentación de GLN como parte del programa de ciencia ciudadana BeeWatch y analizamos el aprendizaje de los usuarios. Finalmente, implementamos retroalimentación de GLN en el programa en vivo y evaluamos esto al asignar al azar a todos los usuarios de BeeWatch a grupos de tratamiento que recibieron diferentes tipos de retroalimentación al presentar la identificación. Después de seis meses, se enviaron encuestas separadas para evaluar si las opiniones sobre el programa de ciencia ciudadana y su retroalimentación variaban entre los grupos. La certeza en la identificación y la retención de novatos fueron mayores para aquellos grupos que recibieron la retroalimentación automatizada que para aquellos que sólo recibieron la confirmación de la identificación correcta sin la explicación. El valor de la retroalimentación de GLN en el programa en vivo, capturado por medio de cuestionarios y la evaluación en línea de la herramienta de entrenamiento basada en fotos, también mostró que la generación automatizada de retroalimentación informativa promueve el aprendizaje y el compromiso de los voluntarios, lo que sienta el camino para proyectos de ciencia ciudadana productivos y de larga vida.

Palabras Clave: entrenamiento, generación de lenguaje natural, identificación de abejorros, motivación y retención de voluntarios, registro biológico

Introduction

Concern about the state of the natural environment has heightened society's desire to monitor it (Mol 2008). This desire is enforced by policy instruments, such as the 1993 Convention of Biological Diversity and its successors, which act on environmental concerns and demand biological recording (Lawrence & Van Turnhout 2010), and is further enhanced by the perceived need for a strong knowledge base to inform biodiversity and ecosystem management (Adams & Sandbrook 2013). Combined, these factors have created a demand for biodiversity data that far outstrips the capacity of professional biologists to deliver (Danielsen et al. 2005). Indeed, a key driver behind the upsurge of nature-related citizen science initiatives is the desire for data at ever greater spatial and finer temporal scales (Devictor et al. 2010). Opportunely, the willingness of citizens to act as biological recorders has greatly increased because of the rise of environmentalism (Bell et al. 2008) and a societal shift towards more participatory forms of governance that actively invite citizen engagement (Conrad & Hilchey 2011). Rapid advances in distributed technologies have further stimulated biological recording by citizens (Rotman et al. 2012; Kelling et al. 2013) through numerous new citizen science initiatives (e.g., eBird [http://ebird.org/content/ebird/]; Zooniverse [https://www.zooniverse.org/]; Open Air Laboratories [http://www.opalexplorenature.org/]).

Publicity surrounding a citizen science initiative can reach large numbers of people. Yet, a much smaller number will volunteer their time to become genuinely involved (Worthington et al. 2012), and consideration needs to be given to how these volunteers can be kept motivated and engaged. Reasons for people to volunteer are complex and multifactorial (Clary et al. 1998; Beirne & Lambin 2013). In the context of biological recording, people volunteer for social gains (Bell et al. 2008) but also because they value nature and recording allows them to enact their relationship with nature and contribute to its protection (Lawrence & Van Turnhout 2010). The opportunity to develop and further hone skills also plays an important role (Ellis 2011). Although understudied, opportunities for learning and being able to see one's contribution are indeed deemed key components of volunteer retention in citizen science (Bonney et al. 2009; Silvertown et al. 2013). For popular species groups such as birds, very large citizen science programs may result in which numerous volunteers have high levels of skill, self-determination, and motivation (Greenwood 2007). For most other species groups larger-scale data collection by volunteers is more difficult. For example, people's interest in insects is generally low, and experts are few

(Hopkins & Freckleton 2002), despite this group of organisms representing much of Earth's species diversity. Citizen science programs addressing less popular species groups thus need to understand and manage volunteer motivation particularly well and to have measures in place that accommodate generally low levels of species identification skills among volunteers (Theobald et al. 2015).

Learning opportunities represented by citizen science initiatives are diverse and include both 'familiarity gains' relative to a species group or topic and the development of specific skills (e.g., species identification). Inviting members of the public to contribute to citizen science initiatives should oblige coordinators to provide such opportunities, and this is known to be of importance to volunteers (Kelling et al. 2013; Silvertown et al. 2013). Yet, in part because of resource limitations, most initiatives provide feedback only periodically and about the initiative as a whole, which, besides from being impersonal, does not facilitate individual learning. To address this and other constraints, the field of citizen science has become increasingly reliant on understandings from disciplines such as computing and educational sciences. For example, data mining and analytical visualization techniques (Hochachka et al. 2012; Kelling et al. 2013) have enabled effective communication of overall project achievements; these techniques can be starting points for self-directed learning by volunteers, although individual contributions are not addressed. Education sciences have brought out the critical importance of individual feedback for both learning and engagement (Butler & Winne 1995; Mansfield & Boase-Jelinek 2010; Blaschke 2014). Formative technologies are now being used to generate automated feedback in the field of education to, for instance, support the writing process of individual students and reduce the number of mistakes they make (Leacock et al. 2010; Mansfield & Boase-Jelinek 2010). Despite its recognized importance in educational sciences, very few citizen science initiatives provide automated individual feedback (Hill et al. 2012; Worthington et al. 2012; Webster et al. 2014) and none concerns feedback designed to be formative (i.e., intended to help the reader improve skills).

Here we respond to the imperative for further innovation in how citizen science initiatives engage with volunteers (Bonney et al. 2009) and advance the application of a technology from the computing sciences: natural language generation (NLG). This technology allows for the automated generation of formative feedback, which we used here to train and maintain volunteer interests in a citizen science initiative. Much of the recent focus within NLG applications research has been on data-to-text systems, which typically generate summaries of technical data for professionals such as engineers or nurses (e.g., Portet et al. 2009). Data-to-text systems have previously been used in the ecological realm to unfold textual daily journeys of satellite-tagged Red Kites (*Milvus milvus*) to engage members of the public in a reintroduction program (Ponnamperuma et al. 2013) and to contextualize submissions of conservation volunteers concerning the presence (or absence) of invasive American mink (*Neovison vison*) (Webster et al. 2014). By contrast, the NLG system we applied is based on key insights from the educational sciences to provide formative feedback aimed at improving species identification skills. We determined the utility of our NLG feedback system in both controlled settings and a live citizen science initiative, thereby explicitly testing the hypothesis that the automated provision of formative feedback to volunteers helps them improve their accuracy of species identification and enhances volunteer experience and retention.

Methods

BeeWatch as a Test Platform

We developed, with the Bumblebee Conservation Trust (BBCT) (www.bumblebeeconservation.org), an online photo submission and identification platform called BeeWatch (www.abdn.ac.uk/research/beewatch). BeeWatch allows members of the public (citizen scientists, henceforth users) to submit photos of bumblebees, along with a location and date of sighting (i.e., the basic information required for a biological record). The user is then encouraged to identify the specimen in the photograph as one of the 22 species of bumblebee present in the United Kingdom by using an online identification key. Through the interface (Fig. 1a), the user can select visual features of the bumblebee (e.g., color patterns on thorax and abdomen) to narrow down the possible species, select a species identification, and then submit the record. Subsequently, the submitted photo record is verified by a taxonomic expert at either the BBCT or Aberdeen University, and the correct identification is communicated to the user by email, along with automatically generated textual feedback aimed at helping the user improve her or his identification skills. We studied the effect of this feedback on user accuracy (of species identification) and motivation.

Generating Automatic Feedback Through NLG

We based our approach on ideas about the roles of formative feedback in learning. Sadler (1989) identified 3 expectations of a learner regarding feedback; namely, feedback ought to include the reference level to be achieved, the comparison of actual performance to this reference, and appropriate action to close the gap between the two. Another key concept from the learning literature is "parallel empathy"(Davis 1994), which demonstrates an understanding of the learner's situation. Moreover, previous studies with intelligent tutoring systems show that texts that contain comparisons are a useful device for

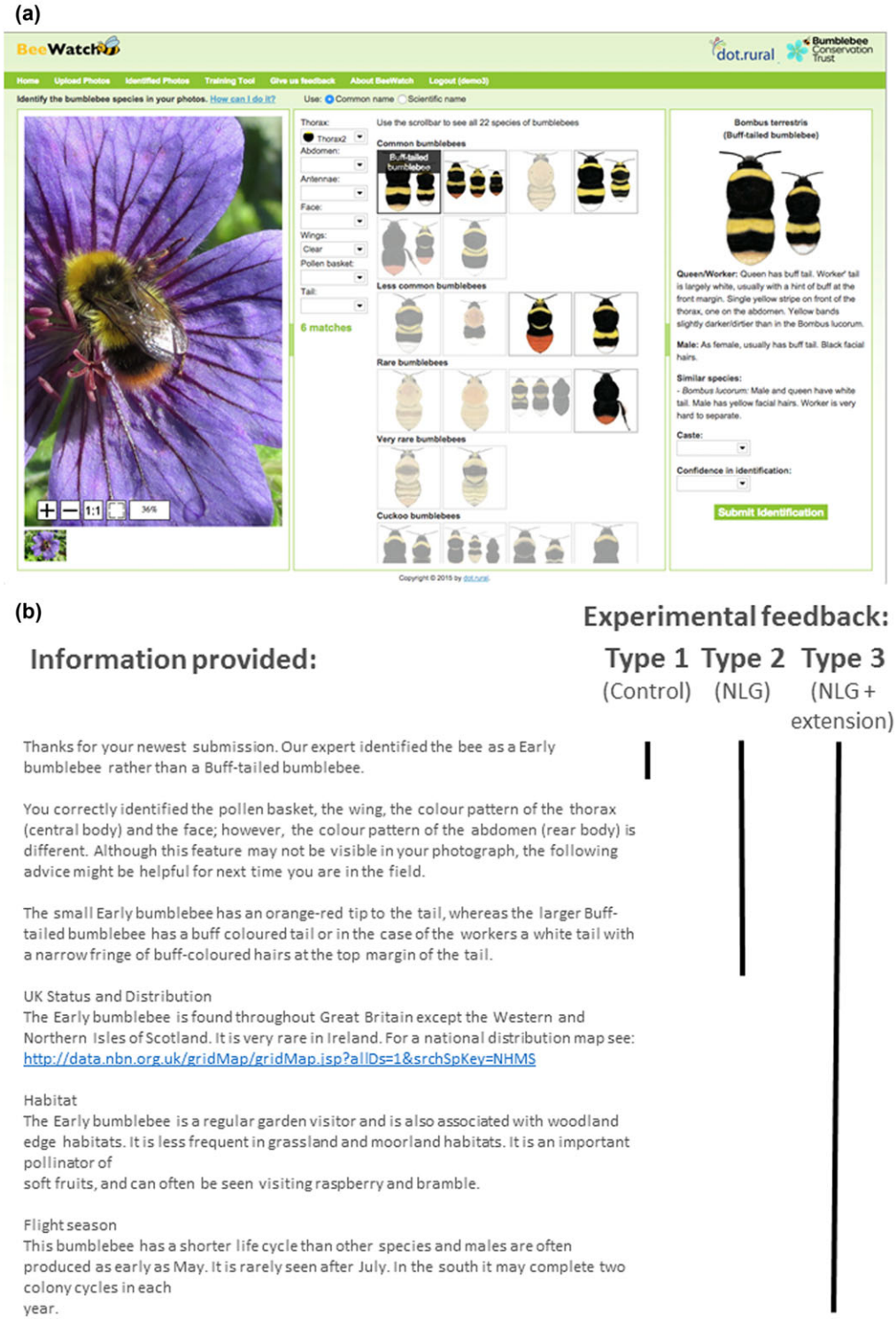

Figure 1. (a) Screenshot of the online identification guide used in the citizen science initiative BeeWatch and (b) the different types of automated feedback provided to BeeWatch participants upon submission of a bumblebee photo. In (a), the photograph on the left is the user's submission; the central panel shows the identification key and its drop-down filters which, if used, greys out species that do not comply with the selected condition(s); and the panel on the right provides details about the selected species and allows users to submit an identification. In this case, the submitted image is an early bumblebee, but the user has identified it as buff-tailed bumblebee. In (b), the automated feedback provided to users for the misidentification in (a), as an example of the three different feedback types examined (type 1, control, acknowledgement + correct ID; type 2, natural language generated (NLG) feedback on the misidentification; type 3, NLG feedback + further ecological information of the species).

augmenting the learner's existing knowledge with new knowledge (Karasimos & Isard 2004). Following these understandings, we used comparisons to augment learning and generated parallel empathy and developed formative feedback in an attempt to close the gap between the actual performance and the reference. Blake et al. (2012) provide technical details about the implementation of the NLG module and its architecture.

To determine whether NLG feedback can enhance identification accuracy and volunteer retention, we created three types of feedback (Fig. 1b). The first type acted as our experimental control (no NLG used) and consisted of an acknowledgment and the correct name of the species as identified by an expert (i.e., Sadler's reference). We labeled this type 1 feedback (Fig. 1b provides an example). The 2 further types of feedback (types 2 and 3) expanded the feedback text through the use of NLG, providing a comparison of visual characteristics of two bumblebee species: the correct one as identified by a taxonomic expert and the one solicited by the user. The resulting automated type 2 feedback had two additional paragraphs (Fig. 1b): one listed body parts in correspondence with the reference if an incorrect identification was made to generate parallel empathy, and the second honed in on specific differences between the two species to foster learning. To provide participants of BeeWatch with potentially interesting, and therefore motivating, further information a third type of feedback was supplied. Type 3 feedback contained exactly the same text as type 2 feedback plus ecological information (as fixed text that succinctly reported a species' abundance and distribution, habitat, and flight season) (Fig. 1b). Where a learner identified a species correctly, the feedback was similar but shorter for types 2 and 3, emphasizing characteristics that distinguish the species identified from other species (Supporting Information).

NLG Feedback and Short-Term Learning by Novices (Experiment 1)

To test whether the provision of NLG feedback helps volunteers improve their accuracy of species identification, 48 third-year biology students at Aberdeen University with no prior experience in bumblebee identification took part in a 45-min experimental trial in March 2012. Participants were randomly allocated to either the control group and received type 1 feedback (i.e., acknowledgement of submission + correct answer; $n = 21$ students) or to the NLG treatment group and received type 2 feedback (type 1 + feedback based on comparisons of visual features; $n = 27$ students). Each group used a separate computer class room and students had individual work stations therein. All students were set the task of identifying, on-screen, 20 photographed bumblebee specimens with the aid of a 2-page identification guide developed by the BBCT (a precursor of what became the online key illustrated in Fig. 1a). The textual feedback appeared on screen every time a species identification was entered and before a new photograph was shown. After the last identification had been submitted, each user was asked to rate how helpful they found the feedback (on a scale of 1–5, with 1 being not very helpful and 5 being very helpful).

NLG Feedback and Longer-Term Learning by Novices (Experiment 2)

To determine whether NLG feedback provision could foster learning over longer time frames and outside a classroom setting, we invited all first-, second-, and third-year biology students at Aberdeen University to take part in a 5-week-long trial commencing in January 2013. These participants were chosen because they had no prior experience with bumblebee identification but, given their field of study, could be expected to have an interest in the subject matter. Indeed, we advertised the opportunity to "gain unique identification skills and learn about the bumblebees that occur in the UK" to attract volunteers. Upon logging into the identification web interface, students were automatically assigned to either type 1 or type 2 feedback (Fig. 1b). Over 5 weeks, they were asked to identify, with the help of the online key (Fig. 1a), bumblebee specimens in batches of 10 photographs. Upon completion of a batch (to be done in a single session), a new batch was offered about a week later until the participant had identified bumblebees in 50 photographs. Feedback was provided after each identification. When a set of 10 photos had been completed the student was also provided with an average accuracy score and a request to log out and return in a week's time. Students were sent reminders if they had not attempted a new set for the week but were free to opt out of the study at any point.

NLG Feedback Training Tool and Its Use by Citizen Scientists

Following experiment 2, a training tool was developed for BeeWatch that we used to study the utility of NLG feedback on actual citizen science volunteers. We used the same set and order of 50 photographs as in experiment 2 but provided all users with type 2 feedback because the tool was meant to be a learning resource for BeeWatch users. No constraints were put on users. They could attempt bumblebee identification on as many photographs as they wished in a session and could restart where they left off. However, after having gone through all photographs the user could no longer access the training tool, and there was no possibility to revisit and revise earlier attempts. We assessed the effect of type 2 (NLG) feedback provision on the rate of learning of training tool users by comparing results from the citizen scientists with results from the control group of experiment 2 (receiving type 1 feedback).

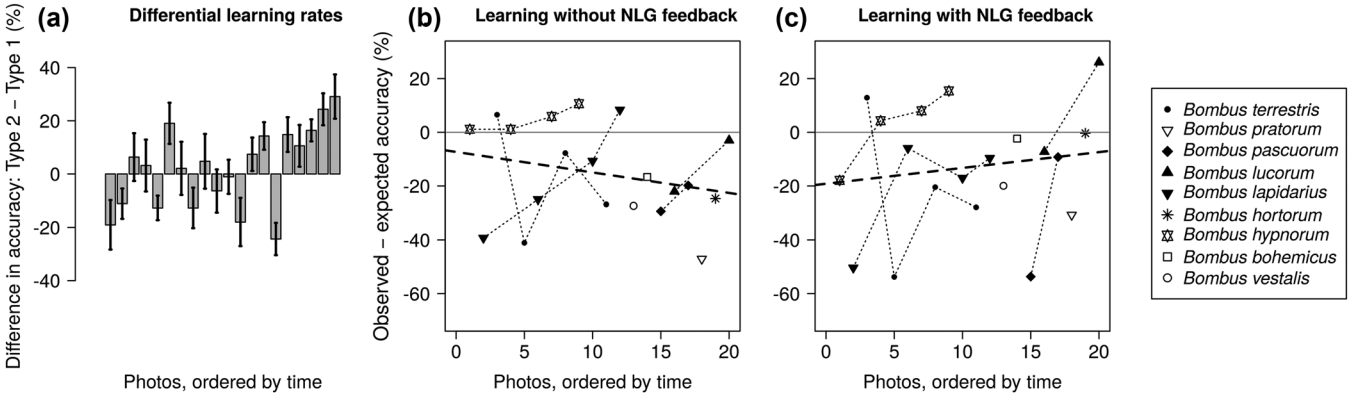

Figure 2. Effect of natural language generated (NLG) feedback on short-term learning in novice bumblebee identifiers over time: (a) difference in accuracy of identifying specimens of bumblebees in a photograph (mean and SE) between participants who received NLG feedback (type 2, correct answer + feedback based on comparisons of visual features) and those who received only the correct answer (type 1 feedback) over the course of 20 photographs in one sitting and the (b, c) difference between observed accuracy and expected accuracy (calculated using the average accuracy of BeeWatch citizen scientists in identifying that species) of participant identifications of bumblebees for each photograph for those who received (b) type 1 and (c) type 2 feedback (dotted lines connect photos of the same species; dashed lines, least-square model fits). Without NLG feedback (type 1), no overall improvement in accuracy occurred, while with NLG feedback (type 2), average accuracy of the group increased and approached the performance of BeeWatch users. Learning over time at the species level occurred in both treatment groups.

Appraising the Value of NLG Feedback in BeeWatch (Experiment 3)

To assess how citizen science recorders within BeeWatch appraised the NLG feedback, all existing and new BeeWatch users in 2012 were randomly allocated to one of three treatment groups with different types of feedback upon identification submission (types 1, 2, or 3 [Fig. 1b]). Unlike for the earlier experiments, we were able to include type 3 feedback, which was type 2 (NLG) feedback plus contextual ecological information about the recorded species (which is only informative in the context of an actual photo submission by a user). Three separate surveys were sent out at the end of 2012 to assess whether views on BeeWatch and its feedback differed among the three groups of users.

Statistical Approach

All statistical analyses were run in R version 3.1.1 (R Core Team 2014). The same generalized linear mixed model, with logit error distribution, was fitted to the identification accuracy data (a binomial variable) for each of the three experiments. These included the fixed effects expected accuracy (a continuous variable that accounted for differences in identification difficulty among species); feedback type (type 1 or 2); time (order of presentation of photos) or set (order of presentation of sets of photos) (both continuous variables); and the interaction between feedback type and time or set. Participant was included as random effect, and parameter estimates were computed using maximum likelihood estimation (Laplace approximation). Species-specific expected-accuracy estimates were calculated from crowd-sourcing data within BeeWatch (Siddharthan et al. 2015), where participants identified bumblebees in photographs submitted by other BeeWatch users.

Results

NLG Feedback and Short-Term Learning (Experiment 1)

Providing third-year students with instantaneous NLG feedback on species identifications improved their ability to identify bumblebee species. The difference in accuracy between the two treatment groups was almost 20% by the end of the trial in favor of those receiving NLG feedback alongside the correct species identification (NLG feedback × time: $z = 2.58$, $p < 0.01$; Supporting information) (Fig. 2a). Expected accuracy, included in the model to correct for the fact that some bumblebee species are harder to identify than others, was a highly significant term ($z = 8.42$, $p < 0.0001$). When making this correction, it emerged that only the group of students who received NLG feedback gradually approached the level of performance achieved by BeeWatch participants (see diverting dashed lines in Figs. 2b-c). Out of the 20 photos several were of the same species whereas others occurred in the set only once. Improved accuracy over time was observed for all species that occurred multiple times, with the exception of the easily confused

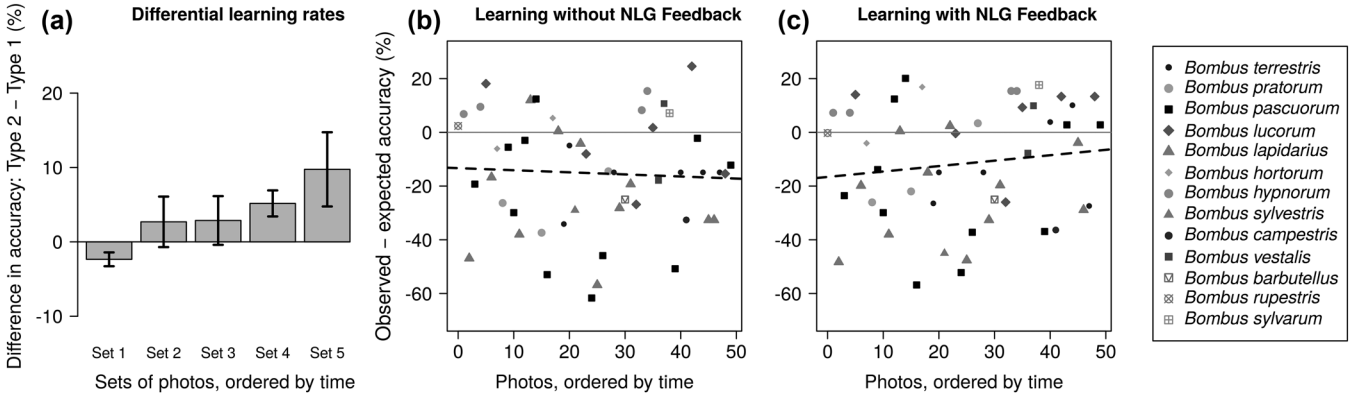

Figure 3. Effect of natural language generated (NLG) feedback on longer-term learning in novice bumblebee identifiers over time: (a) difference in accuracy of identifying specimens of bumblebees in a photograph (mean and SE) between participants who received NLG feedback (type 2) and those who were provided with only the correct answer (type 1 feedback) over the course of 5 sets of 10 photographs each within 60 days and (b, c) difference between the observed accuracy of participants for each photograph and the expected accuracy, calculated using the average accuracy of BeeWatch citizen scientists in identifying that species, for those who received either (b) type 1 or (c) type 2 feedback (dashed lines, least-square model fits). Without NLG feedback, no improvement was observed, whereas with NLG feedback accuracy of the group improved and approached the performance of BeeWatch users.

white-tailed bumblebee (*B. leucorum*). However, a single instance of NLG feedback, which detailed why a specimen was of a certain species, appeared to have a more immediate learning effect than only being told the right species (Figs. 2b-c). Participants who received the NLG feedback found the feedback on average more helpful than those in the control group (helpfulness scores of 3.85 and 3.09, respectively; $t = 2.78$; $p < 0.01$). Collectively, the experiment demonstrated that exposure to NLG feedback was deemed valuable and allowed novices to rapidly acquire specialist identification skills.

NLG Feedback and Longer-Term Learning (Experiment 2)

The 5-week trial confirmed that whether a photographed species was identified correctly or not strongly depended on what species it concerned, as was clear from the highly significant effect of expected accuracy ($z = 10.51$, $p < 0.001$; Supporting information). Despite the preponderance of this species effect, those who received NLG feedback gradually became better at bumblebee identification than those in the control group (NLG feedback $\times$ time: $z = 1.79$, $p = 0.07$) (Fig. 3a); the difference in average accuracy was 12% by the end of the trial. Indeed, the group that received NLG reached a level of accuracy close to that of the average BeeWatch user, and this was not witnessed for the control group (Figs. 3b-c).

Although NLG feedback was explicitly designed to allow users to get better at bumblebee identification, its provision also influenced participant retention. During the 5-week trial, 72% of participants in the control group stopped prematurely, whereas this was 58% for the group receiving NLG feedback (Fig. 4a). There was no evidence for selective drop out based on achieved accuracy (i.e., the best or worst prevailing). Although few did not finish the set of 10 photos they started on, more participants failed to complete a set in the control group (6 out of 36) than in the NLG feedback group (2 out of 38). Both findings indicate that the NLG feedback provided increased engagement with the task, which in turn led to greater participant retention. Recipients of NLG feedback submitted 1156 identifications out of a possible 1900 (61%), whereas those in the control group submitted 992 out of a possible 1800 (55%) ($z = 3.53$, $p < 0.001$).

Unexpectedly, type of feedback provided also influenced the punctuality of participants in terms of when sets were completed. Those in the NLG group completed most of their sets in the intended periods (grey areas in Figs. 4b-f) throughout the trial, suggesting that weekly reminders were largely acted upon. An increasing number of those in the control group, however, completed their sets late and needed an increasing number of reminders.

NLG Feedback Training Tool and its Use by Citizen Scientists

The BeeWatch training tool, designed to determine whether NLG feedback would also enable learning in an uncontrolled setting where users may differ greatly in skill level, was used by 338 out of 1091 active BeeWatch volunteers (31%) from 15 May to 13 October 2014. Although the training tool was well used the number of identifications attempted by volunteers and their achieved accuracies varied greatly. Regarding the latter, some individuals identified the specimen correctly for 9 or more of the first 10 photographs. Few of those users

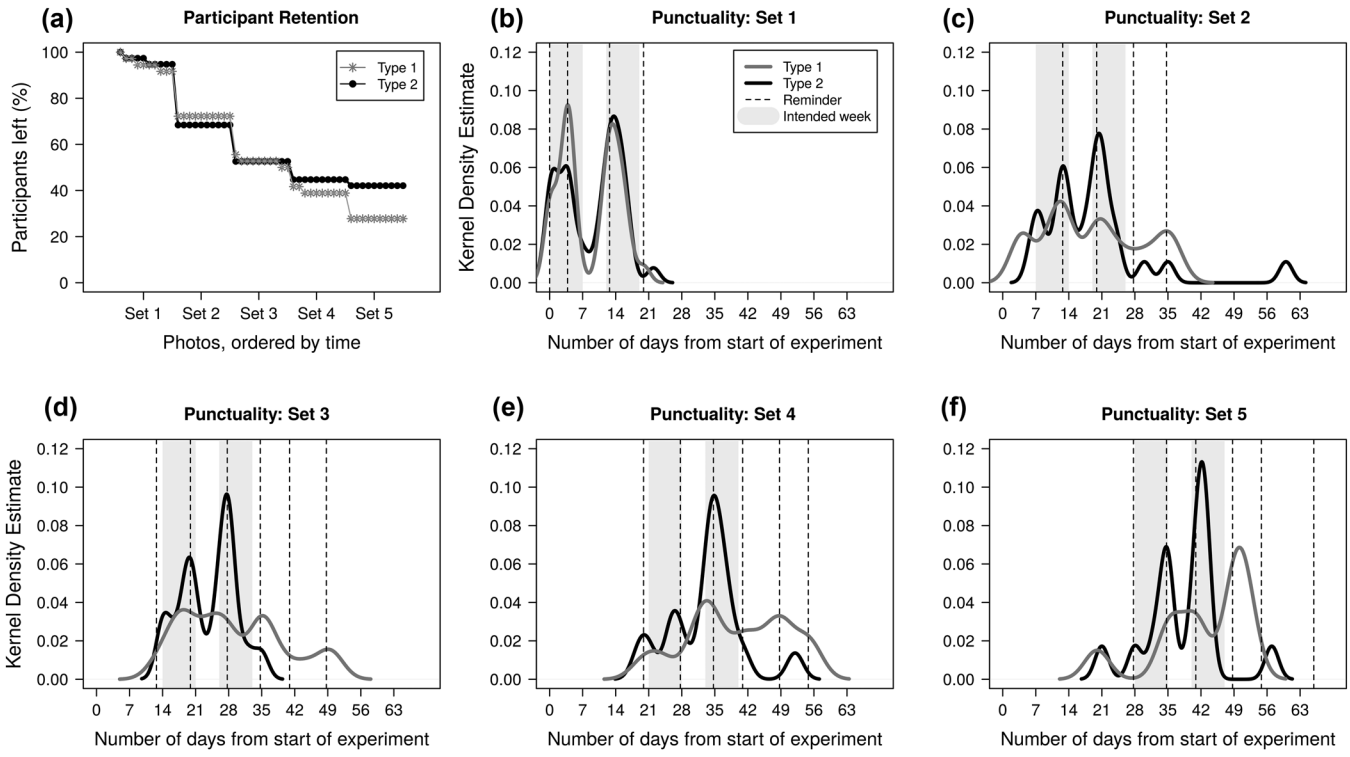

Figure 4. The effect of formative feedback on participant retention and punctuality in experiment 2: (a) percentage of participants who submitted species identifications over the entire experimental period (60 days, during which they were invited to work on 5 sets of 10 photographs of bumblebees each) who received either type 1 feedback (correct answer) or natural language generated (NLG) type 2 feedback (correct answer + text based on comparisons of visual features) and (b–f) kernel density estimates of when participants attempted sets of 10 photographs (grey, weeks when participants were expected to attempt to identify a set of photographs; dashed lines, when reminders were sent out).

went beyond the first 10 photographs, suggesting that these were highly skilled individuals who used the new tool out of curiosity rather than for learning. When we excluded those performing above expectation (because they would benefit little from further online training and thus were not the target audience for which the tool was developed), users gradually improved, reaching an accuracy level comparable to the average BeeWatch user (Fig. 5c), as was observed for the NLG group in experiment 2. This result suggests NLG feedback was also effective in an uncontrolled setting, where users practiced at their own pace. Indeed, when comparing training-tool users with those in the control group of experiment 2 and taking into account the differential difficulty among species (expected accuracy; $z = 16.04$, $p < 0.001$; Supporting Information), those receiving type 2 (NLG) feedback (when using the training tool) gradually became better bumblebee identifiers than those who received type 1 feedback in experiment 2 (Fig. 5b) (NLG feedback $\times$ time: $z = 3.42, p < 0.001$). There was again no evidence for selective drop-out of those achieving low accuracy; hence, we view the observed increase in accuracy as evidence of genuine learning.

The Value of NLG Feedback in BeeWatch (Experiment 3)

Although the survey, disseminated across all users to assess the utility of NLG feedback in the actual citizen science program, revealed that levels of appreciation of BeeWatch were high in general, users who received NLG were most satisfied (very satisfied: 44%, 53%, and 55% for feedback types 1–3, respectively) and more likely to recommend it to others (very likely: 50%, 56%, and 63% for types 1–3, respectively) (Table 1). The greatest difference was in the appraisal of the feedback received. More respondents in the two NLG feedback groups found the feedback useful (64% for both types 2 and 3) than those who received type 1 feedback (47%). Half of the type 1 group wanted to receive more feedback (54%), compared with 23% and 25% of type 2 and type 3 respondents, respectively. Approximately one-third of all respondents thought their bumblebee identification skills had definitely improved as a result of using BeeWatch.

Qualitative survey results further illustrated that NLG feedback was appreciated. Common denominators in the responses of users to the question whether bumblebee identification skills improved as a result of using BeeWatch was the notion of "learning" and that the tool

Table 1. Summary of responses to questions in a survey sent out to participants in the citizen science project BeeWatch who received automatically generated feedback[a] upon submission of bumblebee photos and their identification of the specimens therein.

		Percentage response[b]		
Question	*Descriptor*	*type 1*	*type 2*	*type 3*
Were you satisfied with the overall performance of BeeWatch?	very satisfied	44	53	55
	satisfied	44	37	36
	neither satisfied nor dissatisfied	11	7	9
	dissatisfied	0	3	0
How likely are you to recommend BeeWatch to others?	very likely	50	56	63
	likely	44	35	34
	not very likely	6	7	3
	not at all	0	1	0
How useful did you find the feedback given?	very	47	64	64
	okay	47	30	27
	not very	6	6	8
Would you prefer more or less feedback or was it about the right level?	more	54	23	25
	about right	46	77	73
	less	0	0	2
Have your bumblebee identification skills improved as a result of using BeeWatch?	yes, definitely	35	35	29
	yes, a little	56	56	63
	No, I still can't ID any bumblebee.	4	7	5
	No, I already had good ID skills.	6	1	3

[a] *Volunteers received 1 of 3 types of feedback: 1, acknowledgement + correct species ID; 2, NLG feedback on species ID; 3, NLG feedback on species ID + ecological information.*
[b] *Breakdown, per question, of the percentage of respondents selecting one of the descriptors. Overall survey response rate was 16% (n = 258 volunteers).*

allowed "novices" to be trained in a task that was perceived as rather "difficult." Although one of the type 1 users indicated that BeeWatch "aids are very useful to beginners like me," several users receiving NLG type 2 feedback were more explicit about the value of the tool: "having expert confirmation [. . .] has been excellent [. . .]. It no longer feels like I am taking educated, probably wrong, guesses." One participant even exclaimed, "I enjoyed it so much I wrote a newspaper article about it," wherein she said, "[I have] taken a keen interest in sneaking up as close as I can and photographing them." One of the advantages of NLG feedback appeared to be that it gave users a sense of the number of different species of bumblebee and what to look for to differentiate among them. Participants said, for example, "I didn't realise there were so many different types until I went to the beewatch web site." and ". . . a great way to learn about the different species. It is also fun because now I am always on the look-out for new ones I have not seen before." Some type 3 feedback users put this in even starker terms: "It promotes understanding by not just providing an answer to 'what is this?' [. . .] in other words educationally sound." and "BeeWatch identified it and our pleasure in watching was intensified." These responses indicated that NLG feedback fostered enjoyable and effective learning of a new skill.

When asking users how BeeWatch could be improved, differences between all three types of feedback became clear. The majority of comments made by type 1 users concerned technical issues (e.g., "Make the body part selection more flexible." "Response time was a bit slow."). Comments of type 2 users were distinctly more appreciative, for example, "no need to improve - quite satisfied" and ". . . individual feedback [. . .] was a pleasant surprise." Type 3 users were most complimentary, however; almost half their comments indicated approval (e.g., "Seems pretty good to me already." "I think it is brilliant." "BeeWatch is doing a terrific job."). The desire for feedback beyond the correct species name also came to the fore; type 1 users asked for type 2 feedback (e.g., "Where the ID was wrong some comments on why might help."), and type 2 users asked for type 3 feedback and beyond (e.g., ". . . [provide] more information on how long they live, what happens to them in winter, and what they feed on." "Some general context of the individual results: e.g. how many other sitings of the gypsy bumblebee I saw have been recorded this year, ideally within the area.").

Discussion

Citizen science invites the public to participate in both scientific thinking and data collection (Bonney et al. 2009). The model of citizen science currently having the greatest influence on the environmental realm involves monitoring biodiversity at broad geographic scales. Citizen science data are notoriously noisy, partially because

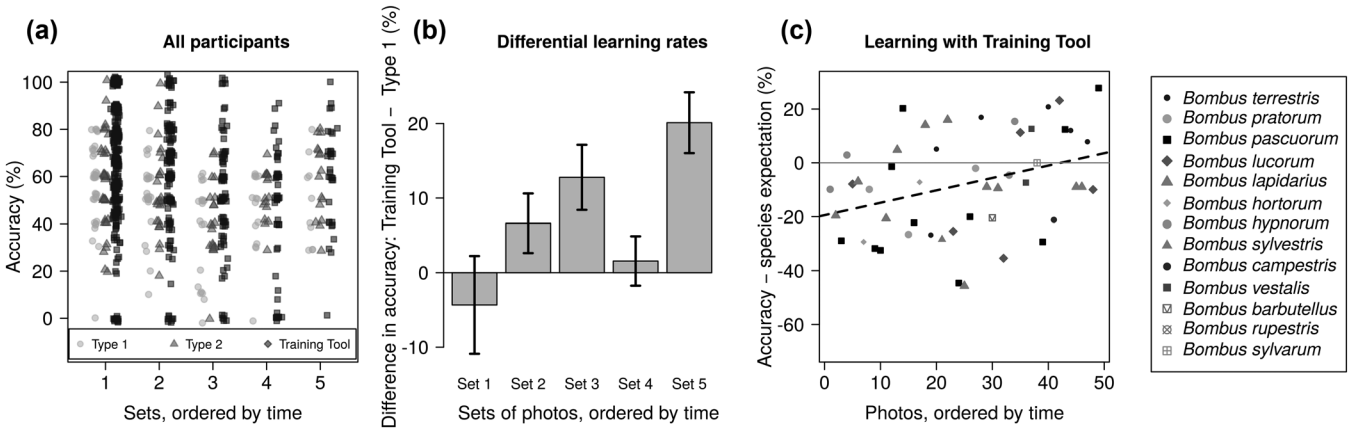

Figure 5. Formative feedback in a citizen science program: (a) bumblebee identification accuracy of training-tool users (who received natural language generated [NLG] feedback) relative to accuracy of participants who were novices and received type 1 (no NLG) and type 2 (NLG) feedback for each set of 10 photos (experiment 2); (b) difference in accuracy of training-tool users and those who received no NLG feedback (type 1) in experiment 2; and (c) observed accuracy of identification of each photograph relative to expected accuracy, calculated using the average accuracy of BeeWatch citizen scientists in identifying that species (dashed line, least-square fit of the accuracy of the training tool users).

of highly variable skill levels among the contributing volunteers (Siddharthan et al. 2015; Theobald et al. 2015). We addressed the issue of data quality by trying to increase skill levels of volunteers through training. The importance of providing educational material to volunteers—for both training and motivational purposes—is widely recognized and practiced using both consultative (e.g., internet-based pictorial guides and videos) and interpersonal (e.g., training workshops) approaches. Indeed, to view learning as a key factor behind volunteer engagement has been proposed as being central to the success of citizen science initiatives (Rotman et al. 2012). With initiatives growing in size and becoming more geographically dispersed, the delivery of personal training becomes more difficult and internet or paper-based approaches more feasible. Ensuring that such non-personal forms of training indeed increase volunteer identification skills as well as their motivation is deemed critical (Silvertown et al. 2013). Our study is a rare example of automated feedback delivery upon data submission in a citizen science initiative and the first to demonstrate its influence on both the rate of learning and motivation of volunteers.

Qualitative data from BeeWatch users revealed that bumblebee identification was difficult for novices despite the fact that drawings of the correct species were always presented among the candidate labels in the species key, a factor that has been identified as important for nonexperts to achieve successful image annotation (He et al. 2013). Given their struggles, it is perhaps unsurprising that being provided with any relevant feedback on their submissions was already highly valued. Also, most web-based citizen science initiatives provide only acknowledgement of a submission; yet, our lowest level of feedback provision already included the correct identity of the specimen. Although NLG feedback on individual submissions was highly valued, a relatively low number (one-third) of respondents to the survey thought their identification skill had definitely improved as a result of using BeeWatch. We suspect that this may be due to the large differences in identification difficulty among species (as evidenced by the large species effect in our linear models; see also Siddharthan et al. 2015). This effect may not have been transparent to users, who were not informed explicitly through the feedback as to whether the species identified was easy or difficult to identify.

The use of the training tool allowed novices to learn through receiving formative feedback on as many images as they wanted and at their own pace. The principle of such a training tool, with extensive feedback on identification specific to misidentification by a user, could play an important role in the formation of skilled recorders and may contribute to halting the decline of species experts for less popular taxa (Hopkins & Freckleton 2002). NLG feedback allowed for the construction of such a training tool with little effort, and we found that it enhanced the rate of learning and motivated volunteers to hone their identification skills—two key drivers in biological recording (Ellis 2011).

The automated feedback provision system we developed removed a major bottleneck for the BBCT and allowed them to scale up from a public engagement exercise to one of the largest providers of U.K. bumblebee records (van der Wal et al. 2015). Although this feedback system was developed for bumblebees, it

is a proof-of-concept and is equally applicable to other species groups. The point is that although social factors are important in volunteering (Bell et al. 2008), biological recording is often a solitary activity; this makes the provision of individual feedback particularly important. We complemented tailored NLG feedback only with ecological information (type 3 feedback). Combining formative feedback with more political notions, such as causes of pollinator decline or expansion of introduced species, would suit the nature conservation agenda well and could create deeper awareness of environmental problems. Although our approach sits in the science education strategy of teaching knowledge and skills, rather than environmental education, which also stresses the incorporation of values and changing behaviors (Wals et al. 2014), our findings indicate that what was designed as learning tool also motivated and drew people into a new world. Some of the unsolicited feedback by email from BeeWatch users (e.g., "Thank you for letting me know what type of bees are in my garden, I am really enjoying watching them and have bought some bee loving plants today so hopefully I will attract some more in future.") support the idea that NLG feedback has the potential to influence environmental behavior.

Acknowledgments

The authors thank H. H. Nguyen for his early development work on the BeeWatch interface; E. O'Mahony, I. Pearce, and R. Comont for identifying numerous photographed bumblebees; B. Darvill, D. Ewing, and G. Perkins for enabling our partnership with the Bumblebee Conservation Trust; and S. Blake for his investments in developing the NLG feedback. The study was part of the Digital Conservation project of dot.rural, the University of Aberdeen's Digital Economy Research Hub, funded by RCUK (grant reference EP/G066051/1).

Supporting Information

An example of type 2 feedback provided where a learner identified a species correctly (Appendix S1) and a summary of the statistical analyses of the 3 experiments testing the utility of NLG feedback (Appendix S2) are available online. The authors are solely responsible for the content and functionality of these materials. Queries (other than absence of the material) should be directed to the corresponding author.

Literature Cited

Adams WM, Sandbrook C. 2013. Conservation, evidence and policy. Oryx **47:**329–335.

Beirne C, Lambin X. 2013. Understanding the determinants of volunteer retention through capture-recapture analysis: answering social science questions using a wildlife ecology toolkit. Conservation Letters **6:**391–401.

Bell S, et al. 2008. What counts? Volunteers and their organisations in the recording and monitoring of biodiversity. Biological Conservation **17:**3443–3454.

Blake S, Siddharthan A, Nguyen H, Sharma N, Robinson A, Mellish C, Van der Wal R. 2012. Natural language generation for nature conservation: automating feedback to help volunteers identify bumblebee species. Pages 311–324 in Proceedings of the 24th International Conference on Computational Linguistics International Committee on Computational Linguistics. Available from http://anthology.aclweb.org/C/C12/C12-1020.pdf (accessed November 2015).

Blaschke LM. 2014. Heutagogy and lifelong learning: a review of heutagogical practice and self-determined learning. International Review in Open and Distant Learning **13:**56–71.

Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg KV, Shirk J. 2009. Citizen Science: a developing tool for expanding science knowledge and scientific literacy. BioScience **59**: 977–984.

Butler DL, Winne PH. 1995. Feedback and self-regulated learning: a theoretical synthesis. Review of Educational Research **3:**245–281.

Clary EG, Snyder M, Ridge R, Copeland J, Stukas AA, Haugen J, Miene P. 1998. Understanding and assessing the motivation of volunteers: a functional approach. Journal of Personality and Social Psychology **74:**1516–1530.

Conrad CC, Hilchey KG. 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. Environmental Monitoring and Assessment **176:**273–291.

Danielsen F, Burgess ND, Balmford A. 2005. Monitoring matters: examining the potential of locally-based approaches. Biodiversity and Conservation **14:**2507–2542.

Davis MH. 1994. Empathy: a social psychological approach. Westview Press, Boulder.

Devictor V, Whittaker RJ, Beltrame C. 2010. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. Diversity and Distributions **16:**354–362.

Ellis R. 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:**769–790.

Greenwood JJD. 2007. Citizens, science and bird conservation. Journal of Ornithology **148**(Suppl 1):S77–S124.

He J, Van Ossenbruggen J, De Vries AP. 2013. Do you need experts in the crowd? A case study in image annotation for marine biology. Pages 57–60 in Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire, Paris. Available from http://dl.acm.org/citation.cfm?id = 2491792 (accessed November 2015).

Hill A, et al. 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. ZooKeys **209:**219–233.

Hochachka WM, Fink D, Hutchinson RA, Sheldon D, Wong W-K, Kelling S. 2012. Data-intensive science applied to broad-scale citizen science. Trends in Ecology & Evolution **27:**130–136.

Hopkins GW, Freckleton RP. 2002. Declines in the numbers of amateur and professional taxonomists: implications for conservation. Animal Conservation **5:**245–249.

Karasimos A, Isard A. 2004. Multi-lingual evaluation of a natural language generation system. Pages 829–832 in Proceedings of the Fourth International Conference on Language Resources and Evaluation. European Language Resources Association, Paris. Available from http://www.lrec-conf.org/proceedings/lrec2004/pdf/134.pdf (accessed November 2015).

Kelling S, Lagoze C, Wong W-K, Yu J, Damoulas T, Gerbracht J, Fink D, Gomes C. 2013. eBird: a human/computer learning network to

improve biodiversity conservation and research. AI Magazine Spring 2013**:**10–20.

Lawrence A, Van Turnhout E. 2010. Personal meaning in the public sphere: the standardisation and rationalisation of biodiversity data in the UK and the Netherlands. Journal of Rural Studies **26:**353–360.

Leacock C, Chodorow M, Gamon M, Tetreault J. 2010. Automated grammatical error detection for language learners. Synthesis Lectures on Human Language Technologies **3:**1–134.

Mansfield P, Boase-Jelinek D. 2010. Optimising automated feedback systems to motivate students. eCULTURE **3:**95–109.

Mol APJ. 2008. Environmental reform in the information age: the contours of informational governance. Cambridge University Press, New York.

Ponnamperuma K, Siddharthan A, Zeng C, Mellish C, Van der Wal R. 2013. Tag2Blog: Narrative generation from satellite tag data. Pages 169–174 in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics, Stroudsburg, Pennsylvania. Available from http://www.aclweb.org/anthology/P13-4029 (accessed November 2015).

Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, Sykes C. 2009. Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence **173:**789–816.

Rotman D, Preece J, Hammock J, Procita K, Hansen D, Parr C, Lewis D, Jacobs D. 2012. Dynamic changes in motivation in collaborative citizen-science projects. Pages 217–226 in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW). Association for Computing Machinery, New York. Available from http://dl.acm.org/citation.cfm?id = 2145238 (accessed November 2015).

Sadler DR. 1989. Formative assessment and the design of instructional systems. Instructional Science **18:**119–144.

Siddharthan A, Lambin C, Robinson A, Sharma NN, O'Mahony E, Mellish C, Van der Wal R. 2015. Crowdsourcing without a crowd: using Bayesian models to minimise crowd size for reliable online species identification. ACM Transactions on Intelligent Systems and Technology **5**. DOI:http://dx.doi.org/10.1145/2776896.

Silvertown J, Buesching CD, Jacobson SK, Rebelo T. 2013. Citizen science and nature conservation. Pages 124–142 in Macdonald DW, Willis KJ, editors. Key topics in conservation biology 2. Wiley-Blackwell, Malden, Massachusetts.

Theobald EJ, et al. 2015. Global change and local solutions: tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation **181:**236–244.

van der Wal R, Anderson H, Robinson A, Sharma N, Mellish C, Roberts S, Darvill B, Siddharthan A. 2015. Mapping species distributions: a comparison of skilled naturalist and lay citizen science recording. Ambio **44** (Suppl. 4):S584–S600.

Wals AEJ, Brody M, Dillon J, Stevenson RB. 2014. Convergence between science and environmental education. Science **344:**583–584.

Webster G, Sripada S, Mellish C, Melero Y, Arts K, Lambin X, Van der Wal R. 2014. Determining content for unknown users: lessons from the MinkApp case study. Pages 113–117 in Proceedings of the 8th International Natural Language Generation Conference (INLG). Association for Computational Linguistics, Stroudsburg, Pennsylvania. Available from http://www.aclweb.org/anthology/W14-4417 (accessed November 2015).

Worthington JP, Silvertown J, Cook L, Cameron R, Dodd M, Greenwood RM, McConway K, Skelton P. 2012. Evolution MegaLab: a case study in citizen science methods. Methods in Ecology and Evolution **3:**303–309.