Sussex Research Online

Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers

Article (Accepted Version)

Ratnieks, Francis L W, Schrell, Felix, Sheppard, Rebecca C, Brown, Emmeline, Bristow, Oliver E and Garbuzov, Mihail (2016) Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. Methods in Ecology and Evolution, 7 (10). pp. 1226-1235. ISSN 2041-2096

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/60625/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

This is the pre-peer reviewed version of the following article: **Ratnieks, F.L.W., Schrell, F., Sheppard, R. C., Brown, E., Bristow, O.E., Garbuzov, M. (2016) Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers**, which has been published in final form in *Methods in Ecology and Evolution* at http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12581/abstract, doi: 10.1111/2041-210X.12581

Abstract

- Citizen science, the involvement of volunteers in collecting of scientific data, can be a useful research tool. However, data collected by volunteers are often of lower quality than that collected by professional scientists.
- We studied the accuracy with which volunteers identified insects visiting ivy (*Hedera*) flowers in Sussex, England. In the first experiment, we examined the effects of training method, volunteer background and prior experience. Fifty-three participants were trained for the same duration using one of three different methods (pamphlet, pamphlet + slide show, pamphlet + direct training). Almost immediately following training, we tested the ability of participants to identify live insects on ivy flowers to one of 10 taxonomic categories and recorded whether their identifications were correct or incorrect, without providing feedback.
- The results showed that the type of training method had a significant effect on identification accuracy ($P = 0.008$). Participants identified 79.1% of insects correctly after using a one-page colour pamphlet, 85.6% correctly after using the pamphlet and viewing a slide show, and 94.3% correctly after using the pamphlet in combination with direct training in the field.
- As direct training cannot be delivered remotely, in the following year we conducted a second experiment, in which a different sample of 26 volunteers received the pamphlet plus slide show training repeatedly three times. Moreover, in this experiment participants received c. 2 minutes of additional training material, either videos of insects or stills taken from the videos. Testing showed that identification accuracy increased from 88.6% to 91.3% to 97.5% across the three successive tests. We also found a borderline significant interaction between the type of additional

material and the test number (*P* = 0.053), such that the video gave fewer errors than stills in the first two tests only.

- The most common errors made by volunteers were misidentifications of honey bees and social wasps with their hover fly mimics. We also tested six experts who achieved nearly perfect accuracy (99.8%), which shows what is possible in practice.
- Overall, our study shows that two or three sessions of remote training can be as good as one of direct training, even for relatively challenging taxonomic discriminations that include distinguishing models and mimics.

Introduction

Citizen science, in which volunteers collect scientific data, is an increasingly popular tool with great potential in research (Dickinson & Bonney 2012; Theobald *et al.* 2015). By providing large amounts of mostly free labour, citizen science makes certain types of research possible, such as ecological studies that collect data over large areas or time scales (reviewed in Devictor, Whittaker & Beltrame 2010; Dickinson, Zuckerberg & Bonter 2010; Conrad & Hilchey 2011; Dickinson *et al.* 2012). Furthermore, citizen science can serve additional purposes, including outreach, raising scientific literacy, changing attitudes towards nature conservation, and giving volunteers an opportunity to help science (Novacek 2008; Bonney *et al.* 2009; Cosquer, Raymond & Prevot-Julliard 2012).

However, citizen science has several challenges. In order for the data to be useful, volunteers generally need training, supervision and overall management. In addition, the data collected by volunteers is often perceived as of low quality and unreliable. For example, in 1993 an amendment was made to prohibit the US National Biological Survey from accepting the work of volunteers, following the assertion by the House of Representatives that "volunteers are incompetent and biased", which was apparently not based on any evidence (Newman, Buesching & Macdonald 2003; Silvertown et al. 2013). In many tasks, volunteers perform less well than professional scientists, yet the data collected are often of sufficiently high quality to be useful, as shown by studies on a range of organisms, including fish (Leslie, Velez & Bonar 2004), corals (Mumby *et al.* 1995), crabs (Delaney *et al.* 2008), marine (Foster-Smith & Evans 2003) and freshwater invertebrates (Fore, Paulsen & O'Laughlin 2001; Engel & Voshell 2002), amphibians (Genet & Sargent 2003), small mammals (Newman, Buesching & Macdonald 2003), birds (Greenwood 2007; Jackson, Gergel &

Martin 2015), coccinellid beetles (Gardiner *et al.* 2012), flower-visiting insects (Kremen, Ullman & Thorp 2011) and plants (Bloniarz & Ryan 1996; Brandon *et al.* 2003; Galloway, Tudor & Haegen 2006; Crall *et al.* 2011; Gollan *et al.* 2012; Fuccillo *et al.* 2015). Indeed, the large sampling effort generated by many volunteers can act as a counter-balance to unsystematic errors in data, and may even lead to the volunteer-collected datasets being more powerful than smaller datasets collected by fewer professionals (Schmeller *et al.* 2009; Gardiner *et al.* 2012), although care needs to be taken to ensure that systematic bias is either precluded in study design or accounted for in the data analysis (Dickinson, Zuckerberg & Bonter 2010; Bird *et al.* 2014; Gonsamo & D'Odorico 2014). In some cases, however, (e.g. papers examining the quality of data gathered by citizen scientists on several species monitored in the Australian project ClimateWatch, http://www.climatewatch.org.au, published in the student journal *Cygnus*, http://cygnus-biologystudentjournal.wikispaces.com), citizen science data can be unreliable and hence of little or no use in research.

One potentially important task for volunteers that is likely to involve errors is the identification of organisms to species or other taxonomic categories, such as in an ecology or conservation project. Previous studies showed that volunteers can identify large organisms, such as trees (Bloniarz & Ryan 1996), crabs (Delaney *et al.* 2008) and corals (Mumby *et al.* 1995), with reasonably high accuracy. However, few studies have directly evaluated the accuracy with which volunteers identify insects (to our knowledge, only Gardiner *et al.* 2012). Insects are an extremely diverse group of organisms that are often small and mobile, and so they are potentially a challenging group for identification. They are also the subject of numerous conservation initiatives (New 2012), including surveys involving volunteers (e.g. Big Butterfly Count, http://www.bigbutterflycount.org; BeeWalk & BeeWatch, http://bumblebeeconservation.org).

Here, we studied the accuracy with which volunteers categorized insects visiting ivy flowers into one of 10 taxonomic categories. We hypothesized that training type and intensity would affect identification accuracy. In year 1 (2013), we compared two forms of indirect training (pamphlet only, pamphlet + slide show) with direct training, in which volunteers were trained by being shown live insects on ivy flowers. However, as direct training is unsuitable for training large numbers of volunteers and/or volunteers based over a large geographic area, in year 2 (2014) we conducted a second experiment, in which we repeatedly tested a different set of volunteers trained using a method that can be delivered remotely and is suitable for training large numbers of volunteers (pamphlet + slide show). In this case, we hypothesized that accuracy would increase with successive training, as volunteers might become more familiar with the different types of insects and might learn to recognize previous mistakes with additional experience.

Methods

Ivy as a tool in monitoring pollinator communities

There is currently considerable concern about declines of pollinators worldwide (Potts *et al*. 2010; Goulson *et al*. 2015). Therefore, it is important to establish monitoring programs to track pollinator communities in order to inform conservation action and ultimately maintain biodiversity and ecosystem services (Lebuhn *et al*. 2013). One way to monitor pollinator communities is to look at flower visits on a common plant species that attracts a wide variety of pollinators (Frankie *et al*. 2002). In Britain and continental Europe, ivy (*Hedera* sp.) is a good candidate plant for this task, as it is a dominant source of nectar and pollen in autumn for a wide range of flower-visiting insects, including bees, wasps, flies and butterflies (Garbuzov & Ratnieks 2014). However, monitoring pollinator communities over large geographic areas and/or long time scales is a labour-intensive endeavour. LeBuhn *et al*. (2013) showed that to detect a 2-5% annual decline in species richness and abundance, c. 200-250 locations need to be sampled twice over 5 years. Citizen scientists could contribute substantially to such monitoring programs.

Insect categories

Study participants were asked to identify insects to one of the following categories, which could be achieved in practice by observation from a short distance: (i) honey bee, *Apis mellifera* (ii) social wasps, *Vespula* spp., (iii) hover flies, Diptera: Syrphidae, (iv) other flies, non-syrphid Diptera, (v) ivy bee, *Colletes hederae*, (vi) bumble bees, *Bombus* spp., (vii) butterflies, Lepidoptera: Rhopalocera, (viii) moths, Lepidoptera: Heterocera, (ix) solitary wasps, aculeate Hymenoptera excluding Anthophila (bees), Formicidae (ants) and Vespinae (social wasps). Although the participants were not required to distinguish between the different types of syrphid flies, the testers recorded the data separately for bee-mimicking (mainly *Eristalis* spp.) and wasp-mimicking species (several genera including *Episyrphus, Chrysotoxum, Helophilus, Myathropa, Syrphus, Volucella*) (Howarth, Clee & Edmunds

2000). In this way we could categorize errors made when the focal syrphid was either a honey bee mimic or a wasp mimic. We did not see any bumble bee mimicking or non-mimic hover flies.

Study participants

Participants were recruited using both personal contacts and posted advertisements on the University of Sussex campus. In Experiment 1 (2013), the 53 participants recruited were mainly students at the University of Sussex or the nearby University of Brighton (n = 49). The other four worked in the Brighton area. Their ages ranged from 20-29 (mean 21.4), with 16 males (30%) and 37 females (70%). In Experiment 2 (2014), 26 new participants were recruited, who were all university students, 16 (62%) males and 10 (38%) females aged 20-24 (mean 21.4).

Although students are not representative of the general population, or the sub-population that is likely to volunteer for a citizen science project, the use of mostly students had an advantage, as it was a relatively homogeneous group in terms of age and education. Given that the main aim of the study was to investigate the effects of different types of training on identification errors, our study sample was less likely to be influenced by confounding variables, although it also means that the results should be generalized with caution.

Experiment 1: Effects of training method, study degree and prior experience

In Experiment 1, participants were randomly assigned to one of three training methods. In method 1 (pamphlet only) participants (n = 20) were asked to spend 20 minutes studying a pamphlet on a single side of A4-sized paper, which featured colour photographs of insects in the 10 categories commonly seen on ivy flowers (Appendix S1). In method 2 (pamphlet + slide show) participants (n = 16) were asked to study the same pamphlet for 10 minutes and in the next 10 minutes were shown a slide show on a computer which included more photographs of insects on ivy flowers, provided tips on insect identification (e.g. showing specific differences between wasps or bees and their hover fly mimics) and included a ‘self-test’ at the end (Appendix S2). In method 3 (pamphlet + direct training) participants (n = 17) studied the pamphlet for 10 minutes and were then given a 10 minute direct training session in the field. Here, either FS or RCS pointed to and identified approximately 20-30 live insects on ivy flowers and gave advice on how to identify them. The duration of training was the same, 20 minutes, in each treatment group.

Additionally, the participants were asked whether the degree they were studying for, or, for the four non-students, the degree they already held, was biology related (n = 29) or not (n = 24), and whether they had any prior experience in insect identification (n = 25) or not (n = 28), as these factors were expected to possibly correlate with identification accuracy.

Experiment 2: Effects of repeated training, experience and video training material

In Experiment 2, participants were randomly assigned into two groups. In both groups, each participant studied the same pamphlet as in Experiment 1 for 10 minutes and viewed the same slide show for 10 minutes. Additionally, one group (n = 13) viewed a 1 min 53 s video (Appendix S3) showing 25 common insects on ivy flowers in clips of a few seconds with narrative comments. For the control group (n = 13), each video clip was replaced with a still image taken from the same clip (Appendix S4). Once assigned to a group, participants were trained using the same method repeatedly on three occasions at 2-4 day intervals, with the training occurring just before the test. Participants were not given any feedback by the tester on whether their identifications were correct or not. This was to mimic a citizen science project being carried out over a large area so that volunteers could not seek advice from the project organizers.

Participant testing

A few minutes after being trained, participants were tested one at a time. An expert tester (FS or RCS in 2013, EB or OEB in 2014) pointed to an insect on an ivy flower and asked the participant to identify it. The expert then noted the identification category given and the actual category, without letting the participant know whether his or her identification was correct or incorrect. Each participant was tested on an average of 49 individual insects (range: 33 – 53) in Experiment 1 and 51 (range 50-65) in Experiment 2. Ideally, each participant would have been tested on the same number of insects per category. However, this was not possible as the relative abundance of different categories varied across time, both within and among days. In addition, certain categories were generally common (syrphid and non-syrphid flies, social wasps, honey bees) and others uncommon (bumble bees, ivy bees, butterflies, moths, solitary wasps).

All training and testing took place on or near the University of Sussex campus using large patches of mature ivy growing on stone walls and in hedges in periods of weather suitable for insect activity (warm, no strong wind, no rain), between 10:00 and 16:00 BST, and between 25 September and 30 October in 2013 and between 30 September and 6 November in 2014, which is during the main period of ivy bloom in the study area.

Expert benchmark

In addition, to estimate a ‘professional’ error rate, six experts were each asked to identify c. 100 individual insects on ivy flowers in the same way as the volunteers. The experts were people who already had a good knowledge of flower-visiting insects and were tested without further training. These included two PhD students (MG and Mr. Thomas Wood), one postdoctotal researcher (Dr Margaret Couvillon), one professor (FLWR) and the two undergraduate students who were working on the project in 2013 (FS & RCS). As all experts were assumed to have equal standing in the testing procedure, the identification of each insect by an expert under test was verified by two of the other experts.

Statistical analyses

To analyse the effects of training method on the proportion of insects correctly identified in each category (Experiment 1), we fitted a series of generalized linear models (GLM) using three factors as explanatory variables: (i) training treatment, (ii) study degree (biology related or non-related) and (iii) prior experience in identifying insects (yes/no), and their two-way interactions. The three-way interaction was not included, as it was not part of the original hypothesis and may have resulted in model overfitting. Similarly, a series of generalized linear mixed models (GLMMs) were used to analyse the effects of (i) repeated training, (ii) type of additional material (narrated video or stills) and their two-way interaction, with ‘participant’ included as a random factor (Experiment 2). The response variable, the proportion of insects identified correctly by each participant, was arcsine square root transformed prior to the analyses. Models were fitted for only those insect groups where the sample size (number of participants who encountered at least one insect of that category) was equal to 3 × the number of predictors (main effects and interactions) or greater (Forstmeier & Schielzeth 2011), i.e. ≥ 18 (Experiment 1) or ≥ 9 (Experiment 2). As a result, data on bumble bees, ivy bees (Experiment 2 only), butterflies, moths and solitary wasps were not analysed due to insufficient sample size, but are presented descriptively. Models were fitted in R v.3.1.1 (R Core Team 2014) using the generalized least squares method with maximum likelihood (package *nlme*, functions *gls* and *lme*, (Pinheiro *et al.* 2013)), which allows for heteroscedasticity. The α-level was not adjusted for multiple models since our aim was to provide an exploratory analysis across a wide range of insect taxa (Moran 2003). When factors were significant, factor levels were compared pairwise using Tukey's HSD post-hoc test (function *glht*, package *multcomp*, (Hothorn, Bretz & Westfall 2008)). All values reported are means per participant ± standard error.

Results

Experiment 1: effects of training method, study degree and prior experience on identification error rate

Type of training method had a significant effect on the proportion of all insects identified without error ($P = 0.008$, Table 1), which was lowest in the 'pamphlet only' treatment (79.1±4.5%, a), followed by the 'pamphlet + slide show' treatment (85.6±1.4%, ab) and the 'pamphlet + field training' treatment (94.3±1.3%, b) (lowercase letters indicate pairwise comparisons based on Tukey's HSD test, where treatments not sharing a letter are significantly different at $\alpha=0.05$) (Fig. 1). Type of training method also had a significant effect on the proportion of bee-mimicking hover flies identified without error ($P < 0.001$, Table 1), which followed the same pattern, increasing from 63.3±6.3% (a) to 83.7±1.6% (ab) to 97.3±1.5% (b) through the same three training methods, respectively. Identification accuracies of the other insect categories analysed were not significantly affected by training method, although there were trends for errors to decrease in most insect categories as training went from method 1 to 2 to 3 (Fig. 2). For example, identification accuracy of honey bees was comparable in methods 1 and 2 (76.6±7.0% and 75.9±4.3%, respectively), but noticeably, although non-significantly ($P = 0.308$), higher in method 3 (88.8±3.8%) (Fig. 2).

Participant study degree, prior experience in insect identification, and the three two-way interactions among the three factors explored did not have significant effects on identification accuracy in any insect category or in all insects combined (Table 1). However, there were trends in expected directions: identification accuracy tended to be higher in participants with a biology related degree (89.2±1.9%) *vs*. non-biology related degree

(82.1±3.9%); and in participants with prior experience (90.9±1.8%) *vs*. without prior experience (81.6±3.5%).

Experiment 2: effects of repeated training and additional video training material

Participants became significantly more accurate at identifying all insects combined (P < 0.001) from test 1 (87.8±1.9%, a) to test 2 (93.3±1.7%, b) to test 3 (97.5±0.5%, c) (Fig. 1). A similar trend was also significant in social wasps alone (P = 0.007), where accuracy increased from test 1 (88.6±3.6%, a) to 2 (91.3±4.2%, a) to 3 (99.7±0.2%, b), and marginally non-significant for bee-mimicking hover flies (P = 0.072) and non-syrphid flies (P = 0.064, Fig. 3, Table 2). In all other insect groups analysed, the accuracy of identification did not improve significantly with repeated training.

The type of additional material (video or stills) did not have a significant main effect on identification accuracy in any insect group. However, it did have borderline significant interactions with repeated training in non-syrphid flies and all insects combined (Fig. 4, Table 2). The interaction meant that the greater identification accuracy of the video group *vs.* the stills group was largest at test 1 (5.9%), less at test 2 (4.5%) and negative, but very close to zero, at test 3 (-0.3%) (Fig. 4).

Expert benchmark

Of the 629 insects identified by the experts, only one was identified incorrectly. This was a relatively rare wasp-mimicking non-syrphid fly, *Conops* sp. (Conopidae), which was mistaken for a solitary wasp. All other insects, 99.8%, were identified correctly, showing that a very low error rate is easily achievable by experts (Fig. 5).

What mistakes did participants make and how did these change with training?

Experiment 1: three training methods

As expected, the highest errors made by participants who received pamphlet only training were in the confusion of honey bees or social wasps with their hover fly mimics. A large proportion of bee-mimicking hover flies (21.6%) were misidentified as honey bees in the pamphlet only training method, which reduced to 13.4% in the pamphlet + slide show training method and to only 2.7% in the pamphlet + direct training method (Table 3). Similarly, misidentifications of wasp-mimicking hover flies as social wasps reduced from 17.1% in the pamphlet only training method to 2.8-4.4% in the two other more intensive training methods (Table 3), although this trend was not statistically significant (Table 1). Similarly, in the pamphlet only and pamphlet + slide show training methods, honey bees were most frequently misidentified as hover flies (7.9-11.2%) and ivy bees (8.5-16.2%); these misidentification rates reduced to 3.5% and 7.2%, respectively, in the pamphlet + direct training method (Table 3), although this trend was also not statistically significant (Table 1).

Experiment 2: one training method repeated three times

Similar to Experiment 1, a large proportion of bee-mimicking hover flies (21.3%) were misidentified as honey bees in test 1, this misidentification rate reduced to 9.3-9.7% in tests 2 and 3 (Table 4), although in contrast to Experiment 1, this trend was only marginally statistically non-significant (Table 2). The proportions of wasp-mimicking hover flies misidentified as social + solitary wasps reduced slightly with repeated training (10.6%, 6.8%, 6.0% in tests 1, 2 and 3, respectively, Table 4), but this was not statistically significant (Table 2). The proportion of non-syrphid flies misidentified as hover flies fell from 2.5% in test 1 to 0.5-0.9% in tests 2 and 3 (Table 4), which was marginally non-significant (Table 2). The only insect group affected significantly by repeated training was social wasps (Table 2), in which the total proportion of misidentifications was similar in tests 1 and 2 (11.4% and 8.7%, respectively) and was spread in comparable proportions among several insect groups (Table 4). All of these errors declined to zero or nearly zero in test 3, making the total proportion of misidentifications only 0.4% (Table 4).

Discussion

Effects of training and experience

Our results clearly show that training method has a significant effect on the accuracy of insect identification. As expected, Experiment 1 showed that pamphlet + direct training was the best method, and that pamphlet alone was the worst method. However, direct training cannot be provided remotely and is labour-intensive to use with large numbers of volunteers. Thus, the highest identification accuracy which can immediately be achieved with remote training materials is that of the pamphlet + slide show method, approximately 86%. This is an encouraging result, given that identification accuracy of at least 80% is considered "acceptable" in most ecological studies relying on citizen science (Cohn 2008).

More encouragingly, Experiment 2 showed that repeated indirect training without external feedback resulted in accuracies of 93.3% in the second test and 97.5% in the third, which was comparable to direct training in Experiment 1, 94.3%. Similar 'learner' effects are not new in citizen science surveys. For example, participants of the Breeding Bird Surveys were shown to underestimate bird counts in their first year of participation, compared to subsequent years (Kendall, Peterjohn & Sauer 1996; Jiguet 2009). Our study, however, shows that the learning curve can operate over only a few days and is the first such result in the identification accuracy of insects. In contrast, Mumby *et al.* (1995) found that participant accuracy in identifying species of coral did not increase with repeated surveys. However, the volunteers underwent an 8-day training programme only in the beginning and were not given repeated training. Similarly, Kremen, Ullman and Thorp (2011) provided a 2-day training course to volunteers and found that their accuracy in identifying flower-visiting insects did not increase between the first and the second sample rounds. These differences suggest that both repeated training and experience, rather than repeated experience alone, is an important part of the learning process, probably due to the opportunity to realize and reflect on one's previous decisions, both correct and incorrect. Our results also support the idea that repeated experience might reinforce learning through reducing memory-related biases (some memories can change over time with certain details becoming sharper or fuzzier, Jordan *et al*. 2012 and references therein). Furthermore, training delivered in stages could result in better learning outcomes, as material taught over a longer period of time is learned better than the same material taught more intensively over a shorter period (spacing effect, Challis 1993; Jordan *et al*. 2012).

Our study tested the ability of volunteers to identify insects almost immediately after training. However, in real citizen science projects, volunteers could experience variable delays between training and data collection. Understanding the effects of this delay on identification accuracy would be a worthwhile topic for further research.

We also found that the additional training material in the form of a video was better than that in the form of still images only in the first two tests. Participant background and pre-experiment experience had only marginally non-significant effects on identification accuracy, suggesting that their influence, if any, is relatively unimportant compared to the effects of training materials and repeated training.

Specific error types

Our results also showed that there were specific identification difficulties with insects on ivy flowers. This is important as it shows where training is most needed. The most common errors in our study were in the misidentification of honey bees or social wasps with their respective hover fly mimics. In addition, honey bees were also often confused with ivy bees and *vice versa*. Similar challenges are likely to be faced when identifying flower-visiting insects on other flowers. Honey bees and hover flies are common insects and, although ivy bees are only found in the autumn on ivy flowers, many other species of bee that are not honey bees or bumble bees do occur.

In contrast, all bumble bees, butterflies and moths were identified correctly, probably because they are well known and have characteristic appearance. Even though few individuals were seen, this result appears to be robust. However, these insect groups, some of which may be of conservation concern, can potentially present identification challenges. For example, bumble bee mimicking hoverfly species do occur in the study area but they are not nearly as commonly seen as honey bee and wasp-mimicking hoverflies, and are rarely seen in the autumn (FLWR pers. obs.). This may be because bumble bees themselves are not common in the autumn. Two species of bumble bee-mimicking moths, *Hemaris* (Sphingidae), occur in southern England, but are rare and fly in spring. Bee flies, *Bombylius* (Bombyliidae), somewhat resemble bumble bees, or bees in general, and are commonly seen in spring, as are several large and stocky species of wild bee (e.g. *Andrena fulva*, *Anthophora plumipes*, *Eucera longicornis*) that can also be confused with bumble bees by novices. In other locations, there may be other sources of confusion. For example, large carpenter bees, *Xylocopa*, are almost absent from Britain but are common in many areas of North America and Europe and are very similar in shape and size to a queen bumble bee.

The high identification accuracies achieved by participants in this study were undoubtedly influenced by the fact that the identification of most insects, other than the honey bee and the ivy bee, was not at the species level. To identify all insects to species would be a very challenging task and, in some genera, would not be possible without capture and microscopic or genetic analysis. This would likely be something that could not be carried out by volunteers, and so would be a necessary limitation of citizen science projects.

Concluding recommendations

Our results allow some recommendations for future citizen science surveys. (i) The use of remote training materials can give high identification accuracy and may be acceptable if provision of direct training is not feasible. (ii) Repeated training without feedback can improve identification accuracy and cause it to reach levels comparable to direct training. (iii) Data from early surveys could well be of lower quality and may need to be discarded. (iv) A preliminary study such as this is desirable in that it would allow an assessment of the 'learner' effect and identification of specific difficulties, which will help in the development of targeted training methods and materials.

Acknowledgements

M.G. was funded by The Body Shop Foundation and Rowse Honey Ltd. We thank Dave Goulson and Norman Carreck for providing feedback on an earlier draft of this manuscript; and Thomas Wood and Margaret Couvillon for allowing us to measure their expert identification accuracy.

Data accessibility

Data are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.rd6v7 (Ratnieks *et al*. 2016).

Supporting Information

Training pamphlet (Appendix S1), training slide show (Appendix S2), and additional training material in the form of narrated video (Appendix S3) and narrated stills (Appendix S4) 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.

References

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M. & Stuart-Smith, J.F. (2014) Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation,* **173,** 144-154.

Bloniarz, D.V. & Ryan, H.D.P. (1996) The use of volunteer initiatives in conducting urban forest resource inventories. *Journal of Arboriculture,* **22,** 75-82.

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

Brandon, A., Spyreas, G., Molano-Flores, B., Carroll, C. & Ellis, J. (2003) Can volunteers provide reliable data for forest vegetation surveys? *Natural Areas Journal,* **23,** 254-262.

Challis, B.H. (1993) Spacing effects on cued-memory tests depend on level processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **19,** 389-396.

Cohn, J.P. (2008) Citizen science: can volunteers do real research? *BioScience,* **58,** 192-197.

Conrad, C.C. & Hilchey, K.G. (2011) A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental Monitoring and Assessment,* **176,** 273-291.

Cosquer, A., Raymond, R. & Prevot-Julliard, A.-C. (2012) Observations of everyday biodiversity: a new perspective for conservation? *Ecology and Society,* **17**.

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J. & Waller, D.M. (2011) Assessing citizen science data quality: an invasive species case study. *Conservation Letters,* **4,** 433-442.

Delaney, D.G., Sperling, C.D., Adams, C.S. & Leung, B. (2008) Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions,* **10,** 117-128.

Devictor, V., Whittaker, R.J. & Beltrame, C. (2010) Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions,* **16,** 354-362.

Dickinson, J.L. & Bonney, R. (2012) *Citizen science: public participation in environmental research*. Cornell University Press, Ithaca.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012) The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment,* **10,** 291-297.

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N. (2010) Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics,* **41,** 149-172.

Engel, S.R. & Voshell, J.R. (2002) Volunteer biological monitoring: can it accurately assess the ecological condition of streams? *American Entomologist,* **48,** 164-177.

Fore, L.S., Paulsen, K. & O'Laughlin, K. (2001) Assessing the performance of volunteers in monitoring streams. *Freshwater Biology,* **46,** 109-123.

Forstmeier, W. & Schielzeth, H. (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology,* **65,** 47-55.

Foster-Smith, J. & Evans, S.M. (2003) The value of marine ecological data collected by volunteers. *Biological Conservation,* **113,** 199-213.

Frankie, G.W., Vinson, S.B., Thorp, R.W., Rizzardi, M.A., Tomkins, M. & Newstrom-Lloyd, L.E. (2002) Monitoring: an essential tool in bee ecology and conservation. *Pollinating Bees: The Conservation Link Between Agriculture and Nature* (eds. P.G. Kevan & V.L. Imperatriz-Fonseca), pp. 187-198. Ministry of Environment, Brasília.

Fuccillo, K.K., Crimmins, T.M., de Rivera, C.E. & Elder, T.S. (2015) Assessing accuracy in citizen science-based plant phenology monitoring. *International Journal of Biometeorology,* **59,** 917-926

Galloway, A.W.E., Tudor, M.T. & Haegen, W.M.V. (2006) The reliability of citizen science: a case study of Oregon white oak stand surveys. *Wildlife Society Bulletin,* **34,** 1425-1429.

Garbuzov, M. & Ratnieks, F.L.W. (2014) Ivy: an underappreciated key resource to flower-visiting insects in autumn. *Insect Conservation and Diversity,* **7,** 91-102.

Gardiner, M.M., Allee, L.L., Brown, P.M., Losey, J.E., Roy, H.E. & Smyth, R.R. (2012) Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. *Frontiers in Ecology and the Environment,* **10,** 471-476.

Genet, K.S. & Sargent, L.G. (2003) Evaluation of methods and data quality from a volunteer-based amphibian call survey. *Wildlife Society Bulletin,* **31,** 703-714.

Gollan, J., de Bruyn, L., Reid, N. & Wilkie, L. (2012) Can volunteers collect data that are comparable to professional scientists? A study of variables used in monitoring the outcomes of ecosystem rehabilitation. *Environmental Management,* **50,** 969-978.

Gonsamo, A. & D’Odorico, P. (2014) Citizen science: best practices to remove observer bias in trend analysis. *International Journal of Biometeorology,* **58,** 2159-2163.

Goulson, D., Nicholls, E., Botías, C. & Rotheray, E.L. (2015) Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science,* **347,** 1255957.

Greenwood, J.J. (2007) Citizens, science and bird conservation. *Journal of Ornithology,* **148,** 77-124.

Hothorn, T., Bretz, F. & Westfall, P. (2008) Simultaneous inference in general parametric models. *Biometrical Journal,* **50,** 346-363.

Howarth, B., Clee, C. & Edmunds, M. (2000) The mimicry between British Syrphidae (Diptera) and aculeate Hymenoptera. *British Journal of Entomology and Natural History,* **13,** 1-39.

Jackson, M.M., Gergel, S.E. & Martin, K. (2015) Citizen science and field survey observations provide comparable results for mapping Vancouver Island White-tailed Ptarmigan (*Lagopus leucura saxatilis*) distributions. *Biological Conservation,* **181,** 162-172.

Jiguet, F. (2009) Method learning caused a first-time observer effect in a newly started breeding bird survey. *Bird Study,* **56,** 253-258.

Jordan, R.C., Ehrenfeld, J.G., Gray, S.A., Brooks, W.R., Howe, D.V., Hmelo-Silver, C.E. (2012) Cognitive considerations in the development of citizen science projects. *Citizen Science: Public Participation in Environmental Research.* (eds. J.L. Dickinson & R. Bonney), pp. 167-178. Cornell University Press, Ithaca.

Kendall, W.L., Peterjohn, B.G. & Sauer, J.R. (1996) First-time observer effects in the North American breeding bird survey. *The Auk,* **113,** 823-829.

Kremen, C., Ullman, K.S. & Thorp, R.W. (2011) Evaluating the quality of citizen-scientist data on pollinator communities. *Conservation Biology,* **25,** 607-617.

Lebuhn, G., Droege, S., Connor, E.F., Gemmill-Herren, B., Potts, S.G., Minckley, R.L., Griswold, T., Jean, R., Kula, E., Roubik, D.W., Cane, J., Wright, K.W., Frankie, G. & Parker, F. (2013) Detecting insect pollinator declines on regional and global scales. *Conservation Biology,* **27,** 113-120.

Leslie, L.L., Velez, C.E. & Bonar, S.A. (2004) Utilizing volunteers on fisheries projects: benefits, challenges, and management techniques. *Fisheries,* **29,** 10-14.

Metcalfe, D.J. (2005) *Hedera helix* L. *Journal of Ecology,* **93,** 632-648.

Moran, M.D. (2003) Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos,* **100,** 403-405.

Mumby, P.J., Harborne, A.R., Raines, P.S. & Ridley, J.M. (1995) A critical assessment of data derived from Coral Cay Conservation volunteers. *Bulletin of Marine Science,* **56,** 737-751.

New, T.R. (2012) Insect conservation: past, present and prospects. Springer, Dordrecht.

Newman, C., Buesching, C.D. & Macdonald, D.W. (2003) Validating mammal monitoring methods and assessing the performance of volunteers in wildlife conservation—"*Sed quis custodiet ipsos custodies*?". *Biological Conservation,* **113,** 189-197.

Novacek, M.J. (2008) Engaging the public in biodiversity issues. *Proceedings of the National Academy of Sciences,* **105,** 11571-11578.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2013) nlme: linear and nonlinear mixed effects models. R package version 3.1-118.

Potts, S.G., Biesmeijer, J.C., Kremen, C., Neumann P., Schweiger, O. & Kunin, W.E. (2010) Global pollinator declines: trends, impacts and drivers. *Trends in Ecology & Evolution,* **25,** 345-353.

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org [accessed 1 November 2014]

Ratnieks, F.L.W., Schrell, F., Sheppard, R.C., Brown, E., Bristow, O.E., Garbuzov, M. (2016) Data from: Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. *Dryad Digital Repository,* doi: 10.5061/dryad.rd6v7.

Schmeller, D.S., Henry, P.Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Deri, E. & Budrys, E. (2009) Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology,* **23,** 307-316.

Silvertown, J., Buesching, C.D., Jacobson, S., Rebello, T. & Birtles, A. (2013) Citizen science and nature conservation. *Key Topics in Conservation Biology 2* (eds D.W. Macdonald & K.J. Willis), pp. 127-142. John Wiley & Sons, Chichester.

Theobald, E., Ettinger, A., Burgess, H., DeBey, L., Schmidt, N., Froehlich, H., Wagner, C., HilleRisLambers, J., Tewksbury, J. & Harsch, M. (2015) Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation,* **181,** 236-244.

Tables

Table 1. Experiment 1: F- and P-values of the effects of participant training method, study degree (biology related *vs.* non-related), prior experience and their two-way interactions on per-participant identification error rate of insects foraging on ivy flowers in a series of Generalized Linear Models (GLMs). P-values highlighted in bold are significant at α = 0.05.

	Training	Degree	Experience	Training × degree	Training × experience	Degree × experience
Honey bees	$F_{2,31} = 1.22$ $P = 0.308$	$F_{1,31} = 0.61$ $P = 0.440$	$F_{1,31} = 1.25$ $P = 0.272$	$F_{2,31} = 2.15$ $P = 0.133$	$F_{2,31} = 0.28$ $P = 0.756$	$F_{1,31} = 0.33$ $P = 0.568$
Social wasps	$F_{2,43} = 1.06$ $P = 0.359$	$F_{1,43} = 2.29$ $P = 0.138$	$F_{1,43} = 0.84$ $P = 0.369$	$F_{2,43} = 0.08$ $P = 0.925$	$F_{2,43} = 2.50$ $P = 0.094$	$F_{1,43} = 0.47$ $P = 0.499$
Bee-mimicking hover flies	$F_{2,43} = 11.78$ **$P < 0.001$**	$F_{1,43} = 0.91$ $P = 0.345$	$F_{1,43} = 1.40$ $P = 0.243$	$F_{2,43} = 0.01$ $P = 0.989$	$F_{2,43} = 0.87$ $P = 0.428$	$F_{1,43} = 0.27$ $P = 0.606$
Wasp-mimicking hover flies	$F_{2,40} = 1.98$ $P = 0.151$	$F_{1,40} = 0.05$ $P = 0.831$	$F_{1,40} = 0.10$ $P = 0.758$	$F_{2,40} = 0.77$ $P = 0.469$	$F_{2,40} = 0.45$ $P = 0.640$	$F_{1,40} = 0.01$ $P = 0.911$
Non-syrphid flies	$F_{2,43} = 2.20$ $P = 0.123$	$F_{1,43} = 0.32$ $P = 0.577$	$F_{1,43} = 0.17$ $P = 0.686$	$F_{2,43} = 0.79$ $P = 0.460$	$F_{2,43} = 0.37$ $P = 0.693$	$F_{1,43} = 0.17$ $P = 0.681$
Ivy bees	$F_{2,13} = 1.24$ $P = 0.321$	$F_{1,13} = 0.08$ $P = 0.785$	$F_{1,13} = 0.06$ $P = 0.806$	No data in some factor level combinations		
All insects	$F_{2,43} = 5.47$ **$P = 0.008$**	$F_{1,43} = 3.23$ $P = 0.079$	$F_{1,43} = 3.08$ $P = 0.086$	$F_{2,43} = 0.13$ $P = 0.875$	$F_{2,43} = 0.54$ $P = 0.587$	$F_{1,43} = 0.34$ $P = 0.563$

Table 2. Experiment 2: F- and P-values of the effects of repeated training, the type of additional material (video or stills, both with narration) and their two-way interaction on per-participant identification error rates of insects foraging on ivy flowers in a series of Generalized Linear Mixed Models (GLMMs). P-values highlighted in bold are significant at α = 0.05.

	Honey bees	Social wasps	Bee-mimicking hover flies	Wasp-mimicking hover flies	Non-syrphid flies	All insects
Test number (1st, 2nd or 3rd)	$F_{2,40}$ = 2.14 *P* = 0.131	$F_{2,48}$ = 5.58 ***P* = 0.007**	$F_{2,42}$ = 2.81 *P* = 0.072	$F_{2,43}$ = 1.09 *P* = 0.345	$F_{2,48}$ = 2.92 *P* = 0.064	$F_{2,48}$ = 44.02 ***P* < 0.001**
Additional material (video or stills)	$F_{1,24}$ = 1.01 *P* = 0.325	$F_{1,24}$ = 3.43 *P* = 0.076	$F_{1,24}$ < 0.01 *P* = 0.973	$F_{1,24}$ = 0.42 *P* = 0.522	$F_{1,24}$ = 0.20 *P* = 0.661	$F_{1,24}$ = 1.21 *P* = 0.282
Test number × additional material	$F_{2,40}$ = 0.35 *P* = 0.708	$F_{2,48}$ = 2.12 *P* = 0.132	$F_{2,42}$ = 0.51 *P* = 0.604	$F_{2,43}$ = 1.00 *P* = 0.377	$F_{2,48}$ = 3.21 ***P* = 0.049**	$F_{2,48}$ = 3.14 ***P* = 0.053**

Table 3. Experiment 1: breakdown of identification errors made by participants in the three training treatments. Arrows in (b) and (c) represent percentage point increases (↑) or decreases (↓) in misidentification rate as compared to the training treatment of preceding intensity, i.e. (a) and (b), respectively. Only insect groups coloured in red were affected significantly by training method, (see Table 1). Values presented are means per participant (n = 20 (a), 16 (b), 17 (c). Not all participants encountered every type of insect).

(a) Pamphlet only

		Actual insect					
		Honey bee	Social wasp	Bee-mimicking hover fly	Wasp-mimicking hover fly	Other non-syrphid fly	Ivy bee
Participant identification call	Honey bee		2.8%	21.6%	3.5%	0.0%	5.0%
	Social wasp	0.0%		1.1%	17.1%	1.3%	5.0%
	Hover fly	11.2%	3.8%			3.7%	37.5%
	Other non-syrphid fly	1.6%	0.3%	1.8%	1.9%		0.0%
	Ivy bee	8.5%	4.4%	4.3%	4.1%	1.3%	
	Other bee	1.0%	0.3%	4.7%	0.5%	1.5%	0.0%
	Solitary wasp	0.5%	0.5%	2.8%	2.0%	0.0%	0.0%
	Bumble bee	0.5%	0.0%	0.5%	0.0%	0.0%	0.0%

(b) Pamphlet + slide show

		Actual insect					
		Honey bee	Social wasp	Bee-mimicking hover fly	Wasp-mimicking hover fly	Other non-syrphid fly	Ivy bee
Participant identification call	Honey bee		4.6%↑	13.4%↓	0.0%↓	0.0%–	16.7%↑
	Social wasp	0.0%–		0.0%↓	2.8%↓	0.0%↓	0.0%↓
	Hover fly	7.9%↓	2.3%↓			1.1%↓	0.0%↓
	Other non-syrphid fly	0.0%↓	0.0%↓	0.0%↓	0.0%↓		0.0%–
	Ivy bee	16.2%↑	9.8%↑	2.6%↓	3.6%↓	0.0%↓	
	Other bee	0.0%↓	0.0%↓	0.0%↓	0.0%↓	0.0%↓	0.0%–
	Solitary wasp	0.0%↓	0.0%↓	0.4%↓	2.2%↑	0.0%–	16.7%↑
	Bumble bee	0.0%↓	0.0%–	0.0%↓	0.6%↑	0.0%–	0.0%–

(c) Pamphlet + direct training

		Actual insect					
		Honey bee	Social wasp	Bee-mimicking hover fly	Wasp-mimicking hover fly	Other non-syrphid fly	Ivy bee
Participant identification call	Honey bee		0.7%↓	2.7%↓	0.0%–	0.0%–	1.0%↓
	Social wasp	0.5%↑		0.0%–	4.4%↑	0.0%–	4.2%↑
	Hover fly	3.5%↓	2.9%↑			5.9%↑	0.0%–
	Other non-syrphid fly	0.0%–	0.0%–	0.0%–	4.9%↑		0.0%–
	Ivy bee	7.2%↓	1.7%↓	0.0%↓	4.2%↑	0.0%–	
	Other bee	0.0%–	0.0%–	0.0%–	0.0%–	0.0%–	0.0%–
	Solitary wasp	0.0%–	0.0%–	0.0%↓	0.6%↓	0.0%–	0.0%↓
	Bumble bee	0.0%–	0.0%–	0.0%–	0.0%↓	0.0%–	0.0%–

Table 4. Experiment 2: breakdown of identification errors made by participants in the three tests. Arrows in (b) and (c) represent percentage point increases (↑) or decreases (↓) in misidentification rate as compared to the preceding test. Insect groups coloured in red were affected significantly by training method, in light red marginally non-significantly, in white non-significantly, and in grey not analysed (see Table 2). Values presented are means per participant (n = 26, but not all participants encountered every type of insect).

(a) Test 1

		Actual insect					
		Honey bee	Social wasp	Bee-mimicking hover fly	Wasp-mimicking hover fly	Other non-syrphid fly	Ivy bee
Participant identification call	Honey bee		3.8%	21.3%	0.5%	0.0%	30.0%
	Social wasp	3.9%		0.7%	6.0%	0.0%	0.0%
	Hover fly	8.9%	0.5%			2.5%	0.0%
	Other non-syrphid fly	2.9%	0.0%	0.3%	11.3%		0.0%
	Ivy bee	2.6%	0.6%	0.0%	1.6%	0.0%	
	Other bee	1.6%	0.0%	0.3%	0.0%	0.2%	0.0%
	Solitary wasp	1.2%	6.6%	0.3%	4.6%	0.0%	0.0%

(b) Test 2

		Actual insect					
		Honey bee	Social wasp	Bee-mimicking hover fly	Wasp-mimicking hover fly	Other non-syrphid fly	Ivy bee
Participant identification call	Honey bee		0.0%$^{\downarrow}$	9.7%$^{\downarrow}$	2.0%$^{\uparrow}$	0.0%$^{-}$	0.0%$^{\downarrow}$
	Social wasp	0.6%$^{\downarrow}$		0.7%$^{-}$	6.1%$^{\uparrow}$	0.0%$^{-}$	33.3%$^{\uparrow}$
	Hover fly	8.2%$^{\downarrow}$	1.4%$^{\uparrow}$			0.5%$^{\downarrow}$	0.0%$^{-}$
	Other non-syrphid fly	0.0%$^{\downarrow}$	0.0%$^{-}$	1.0%$^{\uparrow}$	0.7%$^{\downarrow}$		0.0%$^{-}$
	Ivy bee	9.7%$^{\uparrow}$	3.5%$^{\uparrow}$	1.0%$^{\uparrow}$	4.0%$^{\uparrow}$	0.0%$^{-}$	
	Other bee	0.0%$^{\downarrow}$	0.0%$^{-}$	0.0%$^{\downarrow}$	0.0%$^{-}$	0.4%$^{\uparrow}$	0.0%$^{-}$
	Solitary wasp	0.0%$^{\downarrow}$	3.8%$^{\downarrow}$	0.0%$^{\downarrow}$	0.7%$^{\downarrow}$	0.0%$^{-}$	0.0%$^{-}$

(c) Test 3

		Actual insect					
		Honey bee	Social wasp	Bee-mimicking hover fly	Wasp-mimicking hover fly	Other non-syrphid fly	Ivy bee
Participant identification call	Honey bee		0.2%$^{\uparrow}$	9.3%$^{\downarrow}$	0.0%$^{\downarrow}$	0.0%$^{-}$	No data
	Social wasp	0.0%$^{\downarrow}$		0.0%$^{\downarrow}$	6.0%$^{\downarrow}$	0.3%$^{\uparrow}$	
	Hover fly	3.1%$^{\downarrow}$	0.2%$^{\downarrow}$			0.9%$^{\uparrow}$	
	Other non-syrphid fly	0.0%$^{-}$	0.0%$^{-}$	2.2%$^{\uparrow}$	8.5%$^{\uparrow}$		
	Ivy bee	6.5%$^{\downarrow}$	0.0%$^{\downarrow}$	0.0%$^{\downarrow}$	0.0%$^{\downarrow}$	0.0%$^{-}$	
	Other bee	0.0%$^{-}$	0.0%$^{-}$	0.0%$^{-}$	0.0%$^{-}$	0.3%$^{\downarrow}$	
	Solitary wasp	0.0%$^{-}$	0.0%$^{\downarrow}$	0.0%$^{-}$	0.0%$^{\downarrow}$	0.0%$^{-}$	

Figures

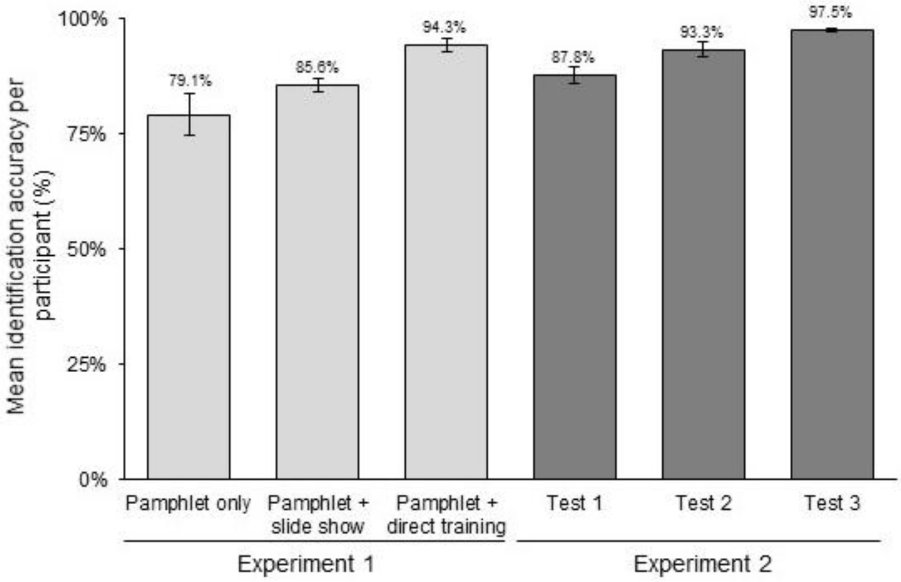

Figure 1. Identification accuracy of all insects combined in Experiment 1, where three groups of participants were trained using a different method and tested immediately after training; and Experiment 2, where all participants were trained and tested three times, using the same training method (pamphlet + slide show + additional material). Bars show means ± standard error.

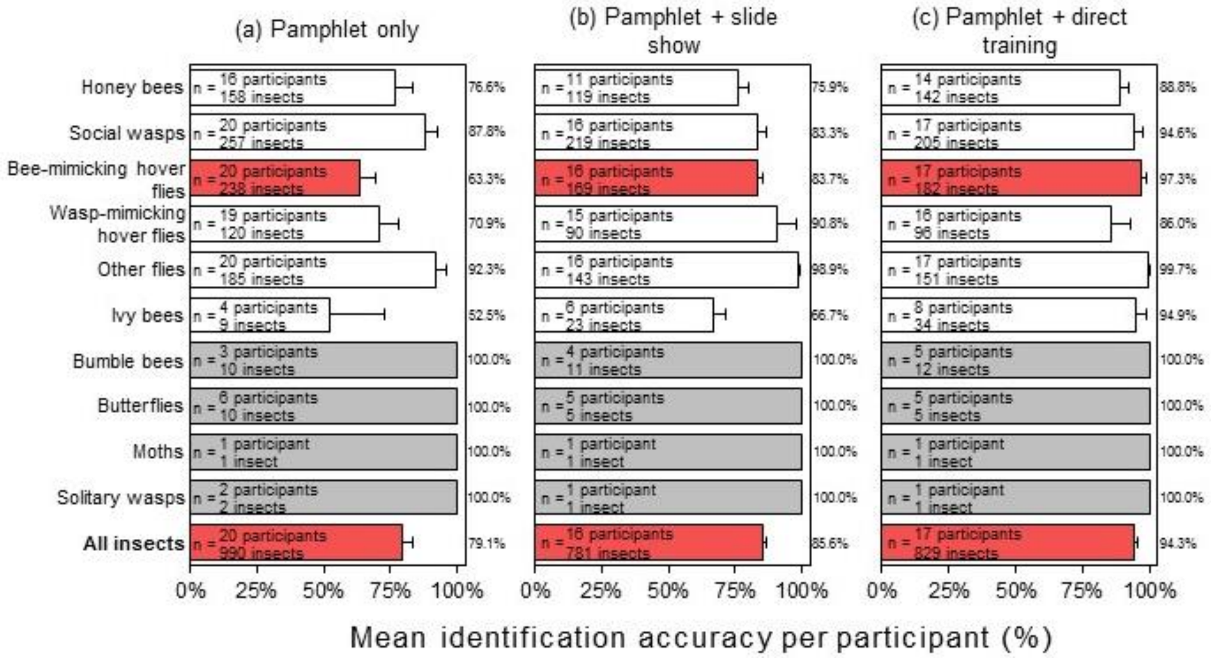

Figure 2. Experiment 1: identification accuracy of insects foraging on ivy flowers by 53 participants in three training method treatments: (a) pamphlet only, (b) pamphlet + slide show and (c) pamphlet + direct training in the field. Insect groups coloured in grey were not compared statistically among training methods due to low sample size. Insect groups coloured in red were significantly different among training method treatments (*P* < 0.05, see Table 1). Bars show mean identification accuracy (%) per participant + standard error.

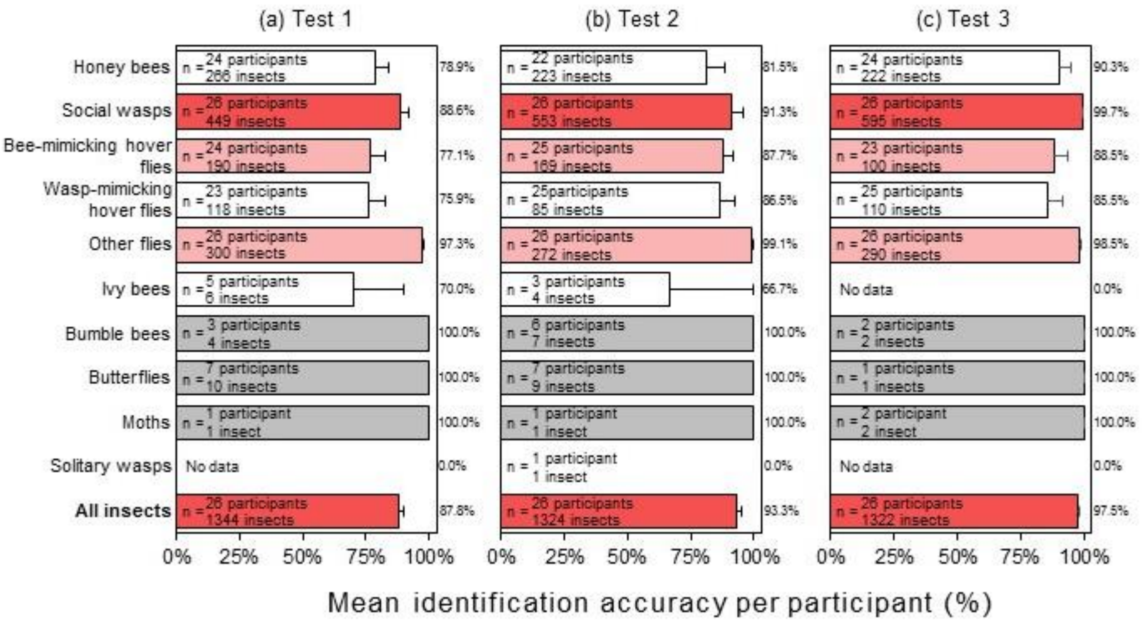

Figure 3. Experiment 2: identification accuracy of insects foraging on ivy flowers by 26 participants across the three consecutive tests. Insect groups coloured in grey were not compared statistically due to low sample size. Insect groups coloured in red were significantly different among tests ($P \leq 0.05$) and in light red marginally non-significantly different ($0.05 < P \leq 0.10$, see Table 2). Bars show mean identification accuracy (%) per participant + standard error.

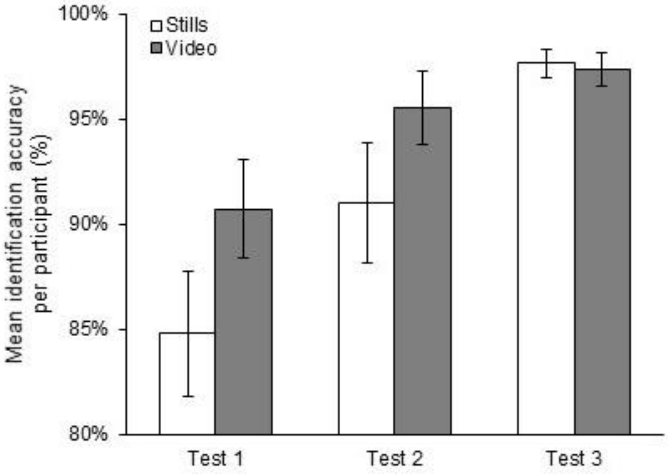

Figure 4. Interactive effects of the type of additional training material (stills or video) and the repeated training and testing on the identification accuracy of all insect groups combined.

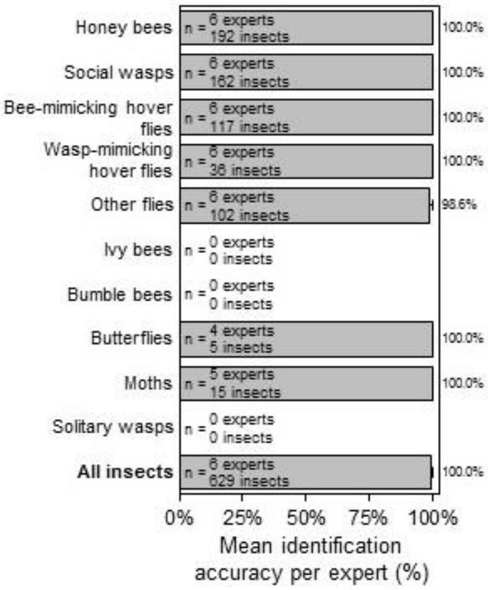

Figure 5. Insect identification accuracy data from 6 experts, which was not part of the experimental data set or the data analysis and is presented for contextualization and general comparison.