UC Santa Barbara
UC Santa Barbara Previously Published Works

Title
The Accuracy of Citizen Science Data: A Quantitative Review

Permalink
https://escholarship.org/uc/item/4g60s2jn

Journal
Bulletin of the Ecological Society of America, 98(4)

ISSN
0012-9623

Authors
Aceves-Bueno, Eréndira
Adeleye, Adeyemi S
Feraud, Marina
et al.

Publication Date
2017-10-01

DOI
10.1002/bes2.1336

Copyright Information
This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, availalbe at
https://creativecommons.org/licenses/by-nc-nd/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The Accuracy of Citizen Science Data: A Quantitative Review

Erendira Aceves Bueno, eaceves@bren.ucsb.edu[1]
Adeyemi S. Adeleye, adeleye.adeyemi@epa.gov
Marina Feraud, mferaud@bren.ucsb.edu
Yuxiong Huang, yhuang@bren.ucsb.edu
Mengya Tao, mengya@ucsb.edu
Yi Yang, yyang@bren.ucsb.edu
Sarah E. Anderson, sanderson@bren.ucsb.edu

Bren School of Environmental Science & Management, University of California, Santa Barbara

Corresponding author:
Sarah E. Anderson
sanderson@bren.ucsb.edu
805-893-5886

Short running title: The Accuracy of Citizen Science Data

Keywords: citizen science; conservation; data accuracy; community-based monitoring; participatory management; monitoring

[1] Author contributions Conceived of and designed study (EAB,ASA,MF,YH,MT,YY,SEA), performed research (EAB,ASA,MF,YH,MT,YY,SEA), analyzed data (EAB,ASA,MF,YH,MT,YY,SEA), wrote paper (EAB,ASA,MF,YH,MT,YY,SEA).

Abstract

Citizen science is increasingly being used to collect data for research. However, there is often concern about the accuracy of the data. Here we use 63 peer-reviewed case studies in ecology and environmental science that compare citizen science data against reference data to statistically evaluate the accuracy of citizen-collected data. Citizen science data is not significantly different from professional data in 62% of the comparisons using p-values, shows moderate to strong correlation ($r \geq 0.5$) with professional data in 51% of the comparisons using correlations, and has at least 80% agreement with professional data in 55% of the comparisons using percent agreement. Data collected by participants who were involved for longer time periods, by participants who had training, by larger groups, and in research related to volunteers' economic and health situations are more accurate. Citizen science can provide useful data, but accuracy for a given task may be low and researchers should design tasks that increase the accuracy of data collected by citizen scientists.

1. Introduction

Citizen science involves volunteers who participate in scientific research by collecting data, monitoring sites, and even taking part in the whole process of scientific inquiry (Roy et al. 2012, Scyphers et al. 2015). In the past two decades, citizen science (also called participatory or community-based monitoring) has gained tremendous popularity (Bonney et al. 2009, Danielsen et al. 2014), due in part to the increasing realization among scientists of the benefits of engaging volunteers (Silvertown 2009, Danielsen et al. 2014, Aceves-Bueno et al. 2015, Scyphers et al. 2015). In particular, the cost-effectiveness of citizen science data offers the potential for scientists to tackle research questions with large spatial and/or temporal scales (Brossard et al. 2005, Holck 2007, Levrel et al. 2010, Szabo et al. 2010, Belt and Krausman 2012). Today, citizen science projects span a wide range of research topics concerning the preservation of marine and terrestrial environments, from invasive species monitoring (e.g., Scyphers et al., 2015) to ecological restoration and from local indicators of climate change to water quality monitoring (Silvertown 2009). They include well-known conservation examples like the Audubon Christmas Bird Count (Butcher et al. 1990) and projects of the Cornell Lab of Ornithology (Bonney et al. 2009).

Despite the growth in the number of citizen science projects, scientists remain concerned about the accuracy of citizen science data (Danielsen et al. 2005, Crall et al. 2011, Gardiner et al. 2012, Law et al. 2017). Some studies evaluating data quality have found volunteer data to be more variable than professionally collected data (Harvey et al. 2002, Uychiaoco et al. 2005, Belt and Krausman 2012, Moyer-Horner et al. 2012) and others that volunteers' performance is comparable to that of professionals or scientists (Hoyer et al. 2001, Canfield Jr et al. 2002, Oldekop et al. 2011, Hoyer et al. 2012). For example, Danielsen et al. (2005) concluded that the 16 comparative cases studies they reviewed only provided cautious support for volunteers' ability to detect changes in populations, habitats, or patterns of resource use. In a more recent review, (Dickinson et al. 2010) found that the potential of citizen scientists to produce datasets with error and bias is poorly understood.

The evidence of problems with citizen science data accuracy (e.g., Hochachka et al. 2012; Vermeiren et al. 2016) indicates a need for a more systematic analysis of the accuracy of citizen science data derived from individual studies of accuracy. To our knowledge, despite useful qualitative reviews (e.g., Lewandowski and Specht 2015), there are to date no reviews that combine the case studies to quantitatively evaluate the data quality of citizen science. In this paper, we conduct a quantitative review of citizen science data in the areas of ecology and environmental science. We focus on the universe of peer-reviewed studies in which researchers compare citizen science data to reference data either as part of validation mechanisms in a citizen science project or by designing experiments to test if volunteers can collect sufficiently accurate data. We code the authors' qualitative assessments of data accuracy and we code the quantitative assessments of data accuracy. This enables us to evaluate both whether the authors believe the data to be accurate enough to achieve the goals of the program and the degree of accuracy reflected in the quantitative comparisons. We then use a linear regression model to assess correlates of accuracy. With citizen science playing an increasingly important role in expanding our scientific knowledge and enhancing the management of the environment, we conclude with recommendations for assessing data quality and for designing citizen science tasks that are more likely to produce accurate data.

2. Methods

This study uses the case survey method to compile the set of studies published before 2014 that directly compare citizen science data with reference data. The goal of this method is to supplement qualitative, in-depth case studies with a quantitative analysis. As with all large-n studies, this prioritizes generalizability over detailed analysis of each case. It supplements existing published case studies and qualitative reviews (e.g., Freitag and Whiteman 2016, Kosmala, et al. 2016).

2.1 Compilation of comparative case studies

We used a 'snowball' approach to identify studies published before 2014 that compare citizen science data with some sort of reference data. Beginning with the 16 studies reviewed in Danielsen et al. (2005), we performed a cited reference search on Google Scholar (http://scholar.google.com/) for papers that cited these 16 studies. Next, we identified every paper cited in this group of papers that compared citizen science data to reference data and again performed a cited reference search on this new group of papers. We repeated this process iteratively until we encountered no new case studies, giving confidence that we had identified the universe of papers in ecology and environmental science that compare citizen science data to reference data. This process yielded a preliminary list of 72 articles. We eliminated 9 studies either because they presented their statistical results in figures (e.g., (Rock and Lauten 1996, Osborn et al. 2005, Thelen and Thiet 2008), did not directly compare citizen science data against professionally collected data (e.g., (Mellanby 1974), or conducted only qualitative comparisons (e.g., (Mueller et al. 2010). Bibliographic information on each of the 63 studies used in this study is provided in Appendix A.

2.2 Extraction of statistical information

For each of the 63 papers, we identified each comparison between citizen scientists and professionals that was made. This yielded 1,363 comparisons, which spanned a wide range of measurements from identification and counts of specific species (Lovell et al. 2009) to calculation of total nitrogen concentration in water (Loperfido et al. 2010). We extracted quantitative statistical results for each comparison. For example, in a study on invasive species (Crall et al. 2011), volunteers' estimates of cover across species were compared to professionals' estimates using a Student t-test, so we recorded the t-statistic, p-value, and degrees of freedom when provided. In that same paper, citizen scientists' correct identification of species was compared to professionals' using percent agreement and a chi-squared test, so each of those values (% agreement, chi-squared value, and p-value) was recorded. That paper also included breakdowns of easy and difficult species identification, as well as the presence or absence of species, resulting in five observations that compare the data from volunteers to that of professionals. To assure data quality, the accuracy of the data extracted from each paper was checked by a second coder after inclusion into the database.

Each comparison of different tasks or different subsets of the tasks is used as an observation here. Where more than one statistical test was used to compare the same set of observations, each was included in the summary of the data presented here. As a result, some comparisons appear more than once among the 1,363 comparisons. Specifically, 182 observations were counted twice and five observations were counted three times to capture all statistical methods that researchers reported in 63 studies. These duplications were eliminated in the analysis that compares citizen science data to professionals. The order of selection for the p-value where multiple tests were used was Student t-test, Wilcoxon signed rank, ANOVA, then Mann-Whitney. In the few examples where no p-value was available and a correlation r-value was available, the correlation r-value was used. We define minimally acceptable levels of accuracy as not being significantly different ($p<0.05$) according to statistical tests, having a correlation greater than 0.5, or having at least 80% agreement. These are relatively low standards for accuracy. We return to what defines an acceptable level of accuracy in our recommendations for comparing citizen science and professional data (section 5.1).

2.3 Authors' Qualitative Evaluations of Citizen Science Data

In addition to collecting the statistical comparisons between citizen science and reference data, we qualitatively code the authors' evaluations of the quality of the citizen science data. For each paper, a coder read the abstract and qualitatively coded whether the authors used words like *accurate*, *reliable*, *comparable*, *statistically similar*, or *valuable* to describe the citizen science data or whether they used words like *no significant correlations*, *overestimated*, or *contradictions.* This results in a binary coding of the authors' assessment of the data as either positive or negative. A second coder confirmed the binary coding of the authors' assessments of the data.

2.4 Covariates of Accuracy

In addition to coding the statistical comparisons between citizen science data and reference data, we coded the attributes of the task and citizen scientists that might affect accuracy. To characterize the task, we coded the discipline as geology, atmospheric science, biology of animals, or botany and the location of the research as marine, freshwater, terrestrial, or the atmosphere. We also coded whether the author noted any particular difficulty with the task, as difficulty affects accuracy (Kosmala et al. 2016). To understand the attributes of the citizen scientists, we coded the length of participation of the citizen scientists into 6 categories ranging from 0-1 month to more than 10 years, whether they participated only once or repeatedly, and the number of citizen scientists participating. We coded whether the paper mentioned that the citizen scientists received training prior to the task and whether the citizen scientists had an economic or health stake in the scientific/research question. Details of the coding are in Appendix B. A linear regression model was fit to assess whether various attributes of the citizen science project affect the percent agreement between citizen science data and reference data.

3. Results

3.1 Characteristics of the Data

Figure 1 provides a summary of the characteristics of the papers. Most of the studies focused on terrestrial systems (47.7%), followed by freshwater systems (29.2%), marine systems (21.5%) and atmospheric studies (1.5%). The majority (69.0%) of the studies were relatively short, with lengths of participation of less than 1 month; a smaller fraction had longer monitoring periods, varying from 2-6 months (34.2%) to 7-12 months (8.5%) to 1-5 years (2.8%). The number of citizen scientists participating in studies tended to be small, with 20.55% of studies using fewer than 10 people. Very few studies (2) used more than 1000 people (2.7%). Other studies engaged 11-50 people (19.2%), 51-100 people (13.7 %), 101-500 people (16.4%), or 501-1000 people (6.9%). Figure 2 shows that more than 60% of the statistical comparisons we analyzed were from animal studies, followed by botany studies and geology-related studies which comprised slightly over 20% and 18%, respectively. Only 0.6% of the comparisons generated by citizen science studies focused on the atmosphere.

Citizen science data and professional data were compared using more than 10 different statistical methods (Figure 2). The comparisons most commonly used percent agreement (42.0%), Mann-Whitney (13.7%), or Student's t-tests (14.2%). The least-used comparison methods were correlations such as linear regression, Spearman's Rank correlation, and Pearson's correlation. Table 1 shows the number of studies and the number of comparisons using each of the statistical methods. Each test measures accuracy in a slightly different way.

3.2 Statistical Comparisons of Citizen Science and Reference Data

While authors tend to be optimistic about the use of citizen science data in their qualitative discussions, we find only 51 to 62% of the comparisons between citizen science data and reference data show accuracy levels that meet our minimum thresholds for accuracy in scientific research. We present results from each of the main data comparison methods (percent agreement, statistics using p-values, correlations, and authors' qualitative evaluations of accuracy) separately in this section and present results from regression analysis in the following section.

Percent Agreement: Is there agreement between the data collected by citizen scientists and professionals?

The most common means of comparing citizen science data to data collected by professionals was percent agreement (525 out of 1363; Table 1); yet this method does not allow for hypothesis testing. As shown in Figure 3, 55.2% of comparisons had a percent agreement equal to or greater than 80%. There was at least 50% agreement in about 86.1% of the comparisons. Percent agreement of 10% or less was reported less than 2% of the time. We note that percent agreement fails to account for agreement by chance (Lombard et al. 2002), so these figures likely overstate the degree of accuracy of citizen scientists.

Statistics using p-values: Are the data collected by citizen scientists and professionals different?

A total of 528 comparisons used various statistical tests that resulted in p-values to test the hypothesis that citizen scientist and professional data are different. Considering a $p\text{-value} \leq 0.05$ as significant, differences between citizen science and professional data were significant in 203 observations (38.4%) and not significant in 325 observations (61.6%), as shown in Figure 4. Each comparison of citizen scientists to professionals was given the same weight, regardless of the sample size or the degree of replication. Alternately, Fisher's method aggregates the results and suggests that there are significant differences between citizen science and professional data when all studies are considered together (results in Appendix C).

Correlations: Are there significant correlations between the data collected by citizen scientists and professionals?

The correlation between citizen scientist and professional data was reported in 81 pairings. Overall, 72% of correlations were significantly greater than zero, but a quarter of the positive correlations were quite weak. We considered values of $r \geq 0.5$ to show moderate to strong correlation between citizen scientist and scientist data. There were 41 observations (50.6%) with $r \geq 0.5$, of which 36 (87.8%) were significant ($p \leq 0.05$), 2 (4.9%) were not significant, and 3

(7.3%) were not reported. A total of 35 observations (43.2%) showed weak positive correlation between citizen scientist and scientist data ($0 \leq r < 0.5$). Of these observations, 12 (34.3%) were significant, 17 (48.6%) were not significant, and 6 (17.1%) had no reported p-values. A total of 5 observations (6.2%) indicated a negative correlation between citizen scientist and scientist data, and in all of these cases the correlations were not significant (Figure 5).

3.3 Authors' Qualitative Evaluations of Citizen Science Data

This analysis shows that, depending on the comparison method, between 51% and 62% of the comparisons resulted in accurate citizen science data. In the 63 papers analyzed, 73% of the abstracts described the contributions of citizen science positively, using words like *accurate*, *reliable*, *comparable*, *statistically similar*, or *valuable*. Only 8 of the papers (13%) assessed citizen scientists' performance negatively, using words like *no significant correlations*, *overestimated*, or *contradictions* in their abstracts. There are two likely reasons for these differences. First, many papers have multiple comparisons between citizen science and reference data, which may allow the authors to conclude that citizen science data is sufficiently accurate for certain tasks. In other words, the authors of the studies frequently saw the usable data within the noise. Second, there is no agreed-upon definition of terms like "reliable". For some scholars, 70% agreement is reliable, yet for others 70% agreement would not be sufficient for the scientific questions they seek to answer. This highlights the crucial role that research design and researcher judgement plays in deciding whether data are accurate enough for a given use.

4. Covariates of Accuracy

The main covariates of citizen scientists' accuracy are location, participation length, monitoring frequency, group size, training, and volunteer type, with about 20% of the total data variance explained by the model (Table 2). Research conducted in marine and terrestrial locations tends to have over 40% higher percent agreement than in freshwater locations. A longer participation length and holding a training session have a positive effect on the percent agreement, both with around 20% increases. This suggests that the studies to quantitatively compare citizen science data to professional data currently available may underestimate the accuracy of projects with longer participation. Surprisingly, citizen scientists who participate repeatedly in the monitoring program perform about 13% worse than those participate only once. If the citizen scientists have an economic or health stake in the outcome, percent agreement is, on average, 68% higher than the general volunteer type.

5. Discussion and Conclusions

5.1 Recommendations to increase transparency and make determination of accuracy more comparable across studies

- Most importantly, we recommend that authors be explicit about their criterion for determining whether the data are "good enough", as assessment criteria appeared to vary considerably. Ideally, this threshold should be determined prior to data collection to more quickly identify problematic tasks during collection and to avoid post-hoc rationalization of the accuracy of collected data. For example, if the goal is to identify catastrophic changes in mussel coverage in the intertidal zone, sufficient accuracy might be that citizen scientists can detect changes of at least one or two standard deviations in existing data. In other research, sufficient accuracy might require detecting much smaller changes.

This lack of explicit criteria for accuracy is particularly acute when correlations are used. For example, one paper reported a Spearman's rank correlation of 0.55 with $p<0.001$. While this allows for a significance test (an advantage over percent agreement), it is unclear whether 0.55 should be considered a high enough correlation. These definitions of accuracy are specific to the research question for which the data will be used and should be specified before data collection commences or analysis proceeds.

- Since percent agreement fails to account for agreement by chance (Lombard et al. 2002), we recommend augmenting it with Fleiss's K coefficient, a more conservative index (Landis and Koch 1977) that is less likely to overstate agreement. While percent agreement is appealing for ease of interpretation, Fleiss's K coefficient has been employed extensively in studies requiring intercoder reliability and both can be reported to balance ease of interpretation with conservative estimates of accuracy.

5.2 Limitations

The case survey method of analysis has well-known shortcomings. First, the case survey method relies on published case studies, which may not adequately cover all areas. In this case, many well-known citizen science projects are long-term and use many citizen scientists. Studies evaluating data quality, however, typically analyze data over a short period of time with fewer participants (Wiggins et al. 2011). The available comparisons of citizen science and reference data may not be fully representative of citizen science projects, which leaves open the possibility that the longer term and larger projects have better data quality. Thus, the conclusions here should be taken to apply mainly to shorter projects. It is clear that studies comparing citizen science data to reference data should continue, as there is more to learn about the correlates of data quality and how to design citizen science projects that produce quality data. Second, the analysis hinges on the quality of the data in the studies. There are reasons to believe that the studies used here likely represent relatively good quality data. They were primarily designed explicitly to test the quality of citizen science data, which likely indicates that the researchers put more thought into how to obtain quality data. Most of the studies here (75.3%) provided training, which improves data quality. Nonetheless, this study must rely on published comparisons and data quality issues are not unique to citizen science. The papers examined here most often compare citizen science data to professional data, a common means of assessing data quality that often makes the assumption that the professional data is fully accurate (Kosmala et al. 2016). Yet data collected by professionals can also have quality issues (Dickinson et al. 2010, Crall et al. 2011, Lewandowski and Specht 2015). We are therefore cognizant that the conclusions drawn here necessarily come from a subset of the citizen science activities that are undertaken, compared with professional data, and published so care must be taken in generalizing to other citizen science projects.

5.3 Conclusions

Despite these limitations of the case survey methodology, it offers the best way to draw quantitative conclusions across the published case studies, since most citizen science studies are not designed with reference data for comparison. As a result, researchers can only qualitatively assess the accuracy of the data. Such qualitative assessments can be valuable, as when a researcher notices citizen scientists struggling to identify uncommon species. But they may be overly optimistic. Although the abstracts of papers comparing citizen science data to professional data indicated that the citizen science data quality was good in 73% of the abstracts, the results of our quantitative assessment cast more doubt on the accuracy of the data. For those studies reporting p-values we found that citizen science was not significantly different from professional data in 62% of the cases. We also found a moderate to strong correlation in 51% of the comparisons reporting correlation, and 55% of the comparisons reporting percent agreement had at least 80% agreement with professional data. Depending on the needs of the researchers, such levels of accuracy may not be sufficient. Monitoring in marine or terrestrial environments, longer participation length, prior training program, larger group size, and conducting research related to volunteers' economic and health situations are good ways to increase the accuracy of the data. This analysis of more than 1,300 comparisons between citizen science and professional data offers some actionable recommendations for researchers using or considering the use of citizen science.

First, the low overall accuracy of the data suggests that *researchers should consider collecting reference data* so as to easily identify suspect citizen science data. If collection of reference data is impractical, researchers should closely supervise citizen scientists to enable qualitative accuracy checks or employ other quality assurance methods. Jacobs (2016) analyzes existing methods for automated and semi-automated quality assurance and existing citizen science projects are constantly innovating to improve data quality (Jacobs). For example, the eBird project establishes a maximum number of birds that may be entered for every species in each month for a given region and then follows up with the original observers if these values are exceeded (Wood et al. 2011) and has continued to improve its data quality procedures.

Second, *researchers should design citizen science tasks with the skill of the citizens in mind and* *employ strategies to improve data quality.* Our regression results suggest that researchers should strive to employ citizen science on projects where citizens participate for longer time periods and should provide training sessions. Training, in particular, has been shown elsewhere to enhance accuracy and credibility (Freitag et al. 2016, Kosmala et al. 2016). A novel finding from this research is that scientists should consider seeking out volunteers with an economic or health stake in the research outcomes, as these volunteers produce data of better quality. For example, researchers might recruit citizens for a mussel study from among recreational harvesters, rather than the general population. Kosmala, et al. (2016) offer other strategies, such as iterative project design, employment of statistical methods for error correction, and good data curation, for improving data quality.

This somewhat pessimistic assessment of citizen science accuracy should not discourage researchers from using citizen science for conservation science, as it has other advantages such as cost-effectiveness and stakeholder engagement (Aceves-Bueno et al. 2015, Newman et al. 2017). Nonetheless, it does call into question the accuracy of the data and suggest that researchers put safeguards like the recommendations above into place when employing volunteers in monitoring and data collection.

Acknowledgments

Isaac Perlman and Trevor Zink participated in early stages of the project. We thank them for their assistance. We would like to thank Michael Bostock for the d3.js script that we used to produce Sankey diagrams in this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aceves-Bueno, E., A. S. Adeleye, D. Bradley, W. Tyler Brandt, P. Callery, M. Feraud, K. L. Garner, R. Gentry, Y. Huang, I. McCullough, I. Pearlman, S. A. Sutherland, W. Wilkinson, Y. Yang, T. Zink, S. E. Anderson, and C. Tague. 2015. Citizen science as an approach for overcoming insufficient monitoring and inadequate stakeholder buy-in in adaptive management: Criteria and Evidence. Ecosystems **18**:493-506.

Belt, J. J., and P. R. Krausman. 2012. Evaluating population estimates of mountain goats based on citizen science. Wildlife Society Bulletin **36**:264-276.

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

Brossard, D., B. Lewenstein, and R. Bonney. 2005. Scientific knowledge and attitude change: The impact of a citizen science project. International Journal of Science Education **27**:1099-1121.

Butcher, G. S., M. R. Fuller, L. S. Mcallister, and P. H. Geissler. 1990. An evaluation of the christmas bird count for monitoring population trends of selected species. Wildlife Society Bulletin **18**:129-134.

Canfield Jr, D. E., C. D. Brown, R. W. Bachmann, and M. V. Hoyer. 2002. Volunteer lake monitoring: testing the reliability of data collected by the Florida LAKEWATCH program. Lake and Reservoir Management **18**:1-9.

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

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

Danielsen, F., K. Pirhofer-Walzl, T. P. Adrian, D. R. Kapijimpanga, N. D. Burgess, P. M. Jensen, R. Bonney, M. Funder, A. Landa, N. Levermann, and J. Madsen. 2014. Linking public participation in scientific research to the indicators and needs of international environmental agreements. Conservation Letters **7**:12-24.

Dickinson, J. L., B. Zuckerberg, and D. N. Bonter. 2010. Citizen science as an ecological research tool: challenges and benefits. Annual review of ecology, evolution and systematics **41**:149-172.

Freitag, A., R. Meyer, and L. Whiteman. 2016. Strategies employed by citizen science programs to increase the credibility of their data. Citizen Science: Theory and Practice **1**.

Gardiner, M. M., L. L. Allee, P. M. Brown, J. E. Losey, H. E. Roy, and R. R. Smyth. 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.

Harvey, E., D. Fletcher, and M. Shortis. 2002. Estimation of reef fish length by divers and by stereo-video: A first comparison of the accuracy and precision in the field on living fish under operational conditions. Fisheries Research **57**:255-265.

Holck, M. H. 2007. Participatory forest monitoring: An assessment of the accuracy of simple cost–effective methods. Biodiversity and Conservation **17**:2023-2036.

Hoyer, M. V., N. Wellendorf, R. Frydenborg, D. Bartlett, and D. E. Canfield Jr. 2012. A comparison between professionally (Florida Department of Environmental Protection) and volunteer (Florida LAKEWATCH) collected trophic state chemistry data in Florida. Lake and Reservoir Management **28**:277-281.

Hoyer, M. V., J. Winn, and D. E. Canfield Jr. 2001. Citizen monitoring of aquatic bird populations using a Florida lake. Lake and Reservoir Management **17**:82-89.

Jacobs, C. Data quality in crowdsourcing for biodiversity research: issues and examples. European Handbook of Crowdsourced Geographic Information:75.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons. 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment **14**:551-560.

Landis, J. R., and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics **33**:159-174.

Law, E., K. Z. Gajos, A. Wiggins, M. L. Gray, and A. Williams. 2017. Crowdsourcing as a Tool for Research: Implications of Uncertainty.

Levrel, H., B. Fontaine, P.-Y. Henry, F. Jiguet, R. Julliard, C. Kerbiriou, and D. Couvet. 2010. Balancing state and volunteer investment in biodiversity monitoring for the implementation of CBD indicators: A French example. Ecological Economics **69**:1580-1586.

Lewandowski, E., and H. Specht. 2015. Influence of volunteer and project characteristics on data quality of biological surveys. Conservation Biology **29**:713-723.

Lombard, M., J. Snyder‐Duch, and C. C. Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. Human communication research **28**:587-604.

Loperfido, J., P. Beyer, C. L. Just, and J. L. Schnoor. 2010. Uses and biases of volunteer water quality data. Environmental Science & Technology **44**:7193-7199.

Lovell, S., M. Hamer, R. Slotow, and D. Herbert. 2009. An assessment of the use of volunteers for terrestrial invertebrate biodiversity surveys. Biodiversity and Conservation **18**:3295-3307.

Mellanby, K. 1974. A water pollution survey, mainly by British school children. Environmental Pollution (1970) **6**:161-173.

Moyer‐Horner, L., M. M. Smith, and J. Belt. 2012. Citizen science and observer variability during American pika surveys. The Journal of Wildlife Management **76**:1472-1479.

Mueller, J. G., I. H. B. Assanou, I. Dan Guimbo, and A. M. Almedom. 2010. Evaluating rapid participatory rural appraisal as an assessment of ethnoecological knowledge and local biodiversity patterns. Conservation Biology **24**:140-150.

Newman, G., M. Chandler, M. Clyde, B. McGreavy, M. Haklay, H. Ballard, S. Gray, R. Scarpino, R. Hauptfeld, and D. Mellor. 2017. Leveraging the power of place in citizen science for effective conservation decision making. Biological Conservation **208**:55-64.

Oldekop, J. A., A. J. Bebbington, F. Berdel, N. K. Truelove, T. Wiersberg, and R. F. Preziosi. 2011. Testing the accuracy of non-experts in biodiversity monitoring exercises using fern species richness in the Ecuadorian Amazon. Biodiversity and Conservation **20**:2615-2626.

Osborn, D. A., J. S. Pearse, and C. A. Roe. 2005. Monitoring rocky intertidal shorelines: a role for the public in resource management. Pages 624-636 *in* California and the World Ocean '02, conf. proc. American Society of Civil Engineers, Reston, VA.

Rock, B. N., and G. N. Lauten. 1996. K-12th grade students as active contributors to research investigations. Journal of Science Education and Technology **5**:255-266.

Roy, H., M. Pocock, C. Preston, D. Roy, J. Savage, J. Tweddle, and L. Robinson. 2012. Understanding citizen science and environmental monitoring: Final report on behalf of UK environmental observation framework.

Scyphers, S. B., S. P. Powers, J. L. Akins, J. M. Drymon, C. W. Martin, Z. H. Schobernd, P. J. Schofield, R. L. Shipp, and T. S. Switzer. 2015. The role of citizens in detecting and responding to a rapid marine invasion. Conservation Letters **8**:242-250.

Silvertown, J. 2009. A new dawn for citizen science. Trends in Ecology & Evolution **24**:467-471.

Szabo, J. K., P. A. Vesk, P. W. J. Baxter, and H. P. Possingham. 2010. Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. Ecological Applications **20**:2157-2169.

Thelen, B. A., and R. K. Thiet. 2008. Cultivating connection: Incorporating meaningful citizen science into Cape Cod national seashore's estuarine research and monitoring programs. Park Science **25**:74-80.

Uychiaoco, A. J., H. O. Arceo, S. J. Green, T. Margarita, P. A. Gaite, and P. M. Aliño. 2005. Monitoring and evaluation of reef protected areas by local fishers in the Philippines: tightening the adaptive management cycle. Biodiversity & Conservation **14**:2775-2794.

Wiggins, A., G. Newman, R. D. Stevenson, and K. Crowston. 2011. Mechanisms for data quality and validation in citizen science. Pages 14-19 *in* e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on. IEEE.

Wood, C., B. Sullivan, M. Iliff, D. Fink, and S. Kelling. 2011. eBird: Engaging birders in science and conservation. PLoS Biol **9**:e1001220.

Table 1. Methods[a] applied by the studies reviewed to test the accuracy of citizen science data.

Methods	**# of studies**	**# of comparisons**
Percentage Agreement	27	525
T-test	15	183
Spearman's Rank Correlation	9	69
Wilcoxon Signed Rank Test	8	61
Pearson's Correlation	8	52
ANOVA	6	21
Linear Regression	5	18
Mann-Whitney Test	4	185
Chi-Square	4	25
ANOSIM	2	7
Kendall's Coefficient of Rank	2	12

[a] Only methods used by 2 or more papers are presented. This table includes comparisons where multiple methods were used. Later analyses eliminate these duplicates.

Table 2. The fitted model coefficients and the corresponding significant levels and standard errors.

Coefficient	Estimate	Standard Error	P-value
Intercept	74.87	10.29	1.51E-12
Location – marine	54.49	8.04	3.78E-11
Location – terrestrial	44.90	6.81	1.20E-10
Participation length – 7 months to 1 year	18.80	9.87	0.057
Monitoring frequency – repeated	-12.92	3.45	0.0002
Group size – medium	0.61	8.22	0.94
Group size – small	-8.38	8.20	0.31
Training – yes	22.14	5.05	1.44E-05
Volunteer type - volunteer	-67.84	7.21	< 2E-16
Specialized knowledge - yes	10.40	4.56	0.023
Adjusted R-Squared	0.20		

Figures

Figure 1. The characteristics of the 63 papers used to compare citizen science and professional data (from left to right): study location, length of participation, citizen scientist group size, and training. NA means data could not be inferred and NR means not reported.

Figure 2. The statistical comparisons of data employed by the papers reviewed in this study. The papers reviewed were grouped into distinct disciplines (first column). This figure shows the type of statistical analysis performed in each study (second column) and the type of result reported (third column). The grey bars represent the proportion of analyses that performed each type of statistical analysis and reported each type of result.

Figure 3. Percent agreement between citizen science data and reference data. The bars represent the amount of analyses (y axis) that reported each level of percent agreement (x axis). The percentage of papers reporting each level of agreement is shown on top of each bar.

Figure 4. Number of comparisons where the data collected by citizen scientists and professionals are significantly different (grey) or not significantly different (pattern). For p-values > 0.05 where the exact p-value was not reported, we randomly and uniformly generated values between 0.051 and 1. A total of 137 comparisons were treated in this way.

Figure 5. Correlation r values for data collected by citizen scientists and professionals, and their associated p-values. Significant correlations are shown in grey, non-significant correlations are shown in pattern, and correlations with no reported p-values are shown in blank. The numbers

within columns represent the number of observations.