STSM 2016 - Report:

Motivations for participation to citizen-science program: A meta-analysis

Authors:

Martin Jeanmougin
UMR 7204, Centre d'écologie et des sciences de la conservation (CESCO), CNRS-MNHN-UPMC-Sorbonne Université, Muséum National d'Histoire Naturelle, 43 rue Buffon CP135, 75005 Paris, France

Liat Levontin
Faculty of Industrial Engineering and Management, Technion, Israel institute of technology, Haifa, Israel

Assaf Shwartz
Faculty of Architecture and Town Planning, Technion, Israel institute of technology, Haifa, Israel

How to cite this report: Jeanmougin, Martin; Levontin, Liat; Shwartz, Assaf. 2017. Motivations for participation to citizen-science program: A meta-analysis. *STSM 2016 Report. Citizen Science COST Action CA15212*. p.19.

Purpose of the STSM

The primary aim of this STSM was to create a database for a meta-analysis on volunteers' aims and needs with respect to their engagement in citizen science (CS) projects, through a systematic quantitative review of the literature focusing on CS programs. Ultimately, results of this project will be used to design effective CS programs, extend participation and increase impact of CS.

This STSM will hence help to provide an extensive list of all motivations that have been proposed and studied to explain volunteer's participation in CS programs. Using the first result of the meta-analysis framework, it will also allow quantifying the relative importance of each category of motivation.

Description of the work carried out during the STSM

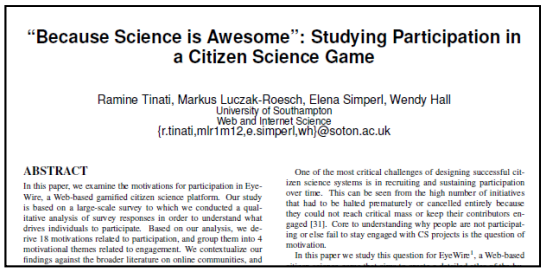
- **Literature search**

The first step of the work was to define the methodological framework to standardize the bibliographic searches. We took advantages on the recent study on publication patterns of citizen science based research developed by Follet and Strezov (2015). They followed a standardized methodology to review the literature on CS for years up to and including 2014. In their study, they classified each article into different categories to pursue their own analysis. One of these categories named "Motivation/Effects" was related to "articles exploring the motivation of participants and the effects of participation". Hence, to define a first list of articles related to CS and motivation (k_1 = 53), we used their database, filtering only article classified under this "Motivation/Effects" category.

To update k_1, we followed and slightly adapted the methodology of Follet and Strezov (2015) for years 2015 and 2016. Hence, using the same scientific literature search engines Web of Science and Scopus, we gathered all articles with "citizen science" in the topic (i.e. Title, Abstract and Keywords). The second step was to remove all duplicates. Following Follet and Strezov (2015), records with the same title and authors were considered duplicates and excluded from further analysis. Once this list was obtained, we examined the titles and abstracts of articles in the list (885 articles) and retained only those focused on the motivations of participants. Hence, we obtained a second list of publication related to CS and motivation (k_2 = 76).

In the next step we downloaded all possible publication listed in the combined list of k_1 and k_2 (k_3 = 135) to analyze, more in depth, aims and results of these articles. Each article of k_3 was read and considered relevant if it clearly referred to quantitative results (e.g. percentage, correlation) related to motivations to participate in citizen science (see Figure 1 for an example on the article of Tinati et al., 2016). The resulting list of articles considered as relevant became the final list of the subsequent analysis ($k_{relevant}$).

Potential relevant publication
for the meta-analysis



"Because Science is Awesome": Studying Participation in a Citizen Science Game

Ramine Tinati, Markus Luczak-Roesch, Elena Simperl, Wendy Hall
University of Southampton
Web and Internet Science
{r.tinati,mlr1m12,e.simperl,wh}@soton.ac.uk

ABSTRACT

In this paper, we examine the motivations for participation in EyeWire, a Web-based gamified citizen science platform. Our study is based on a large-scale survey to which we conducted a qualitative analysis of survey responses in order to understand what drives individuals to participate. Based on our analysis, we derive 18 motivations related to participation, and group them into 4 motivational themes related to engagement. We contextualize our findings against the broader literature on online communities, and

One of the most critical challenges of designing successful citizen science systems is in recruiting and sustaining participation over time. This can be seen from the high number of initiatives that had to be halted prematurely or cancelled entirely because they could not reach critical mass or keep their contributors engaged [31]. Core to understanding why people are not participating or else fail to stay engaged with CS projects is the question of motivation.

In this paper we study this question for EyeWire[1], a Web-based

Results ?

Motivations	Description of Code/Motivation	Total	Pri.	Sec.	M(count)	F(count)	M(%)	F(%)
Contribution+	Contributing to the project, not specifically related to helping science	286	165	121	174	112	37.46	26.08
Science+	Helping improve scientific knowledge. Direct mention of contributing to science	262	188	74	186	94	39.49	19.21
Fun+	For the entertainment value. No specific mention of games or competition	199	174	25	118	81	36.51	17.89
Learning+	To learn about science, or related learning purpose	95	61	34	74	21	47.97	13.23
Personal interest+	For some personal interest towards EyeWire, related to the scientific task	93	70	23	58	35	38.40	19.21
Interesting+	Due to a general interest in EyeWire	83	68	15	57	26	42.29	14.69
Procrastination+	To avoid doing another task, i.e. avoid doing school work	70	52	18	39	31	34.31	22.91
Relaxing+	As a way to relax from other tasks.	66	47	19	31	35	28.92	28.61
Gaming+	The ability to play a game, specific mention of gaming	56	41	15	32	24	35.19	22.49
Puzzle	Specific mention of enjoying the puzzle aspect of the task	39	28	11	22	17	34.74	23.33
Challenge+	Specific mention of the challenge aspect of the task	36	21	15	17	19	29.08	34.76
Community	Taking part, or feeling part of a community	27	7	20	20	7	45.61	38.42
Curious	An 'initial' interest in EyeWire, not specifically mentioning gaming or science	23	22	1	11	12	29.45	20.96
Beautiful+	Specific mention of the visually appealing aspects of EyeWire (e.g. 3D cube visuals)	13	1	12	2	11	9.47	22.64
Competition	The ability to compete with other players in the game	12	3	9	8	4	41.05	51.23
Interface	Specific mention of the 3D interface and actual design of the platform	7	5	2	3	4	26.39	30.74
Addictive+	The addictive nature of the task, often describing their flow or finding it difficult to stop	6	5	1	0	6	0.00	46.11
Citizen science	Specific mention of citizen science	5	4	1	4	1	49.26	9.61

Table 7: Motivations (Primary and Secondary) for both Genders. Motivations identified during the first iterations of coding are denoted by +. Motivations By gender (M,F) are proportional to the number of responses to the survey sample.

Figure 1: Example of a potential relevant publication for the meta-analysis based on the evaluation of its title and abstract. This publication was downloaded and entirely read. Results contain in "Table 7" clearly refer to quantitative results related to motivations to participate in citizen science. Hence, this publication was considered as relevant for the meta-analysis framework.

However, after consideration of this first list of articles and in order to improve our review of the literature, we also analyzed all articles that cite each articles of $k_{relevant}$. Hence, we examined the titles and abstracts of all articles that cite each articles of $k_{relevant}$ using Google Scholar search engine. This other search engine allowed us to access to grey literature (e.g. reports). Thanks to this step, we added seven records to $k_{relevant}$. Finally, we also checked all references contained in each relevant article. This final step added one record to $k_{relevant}$.

The final database of articles that deal with motivation of participants in citizen science program and that are relevant for the meta-analysis framework consists of a total of $k_{relevant} = 40$ articles. All steps describe here and all useful details to understand the data handled here can be found in the Readme document related to the three bibliographic search databases (see Appendix A in this report).

- **Data acquisition**

The next step was to extract all useful and quantitative data from each article of $k_{relevant}$. This section describes briefly which data were extracted. All mentioned types of data were extracted when they were available in the main text of the article or in supporting information. Firstly, we kept all general information (like authors, title, year of publication, etc.).

Secondly, we extracted all contextual and useful data on the methodology used in each article like the location of the study (by country at least), the type of sampling or the name of the project of citizen science. Descriptive data of respondents were also extracted, particularly the number of respondents, age/gender data, level of education and occupation data.

Finally, quantitative results on motivation of participants were extracted from each article. We classified these results into four main categories: percentage data, count data, correlation analysis and Likert scale type analysis.

All details about the data acquisition can be found in the Readme document of the database (namely "extracted_data_database", see Appendix B in this report).

- **Data categorization**

Once we had extracted all these data, we categorized the motivation statements found in each article. A wide variety of possible motivations have been proposed, but a systematic categorization of motivations is still missing. Categorization of each motivation was achieved using an adapted approach of the theoretical framework of Schwartz (2007) on individual's personal values. Motivations for participating could hence be categorized in four principal components, the four C's: Change – openness to change, Competition – self-enhancement, Conservation – in terms of tradition, and Cooperation – self-transcendence. This can be transposed for motivations for participation in CS programs: those who participate to get the satisfaction of contributing to scientific evidence or to learn are persons that are opened to change (Change), while some can participate with the willingness to potentially influence policy (Competition). Finally, participating can also be an opportunity for social experiences (Cooperation) or something coming from a certain tradition (Conservation). Also, these four C's are completed with another C: Cool – when the main reason to participate is related to the fun of the experience with respect to the concept of hedonism.

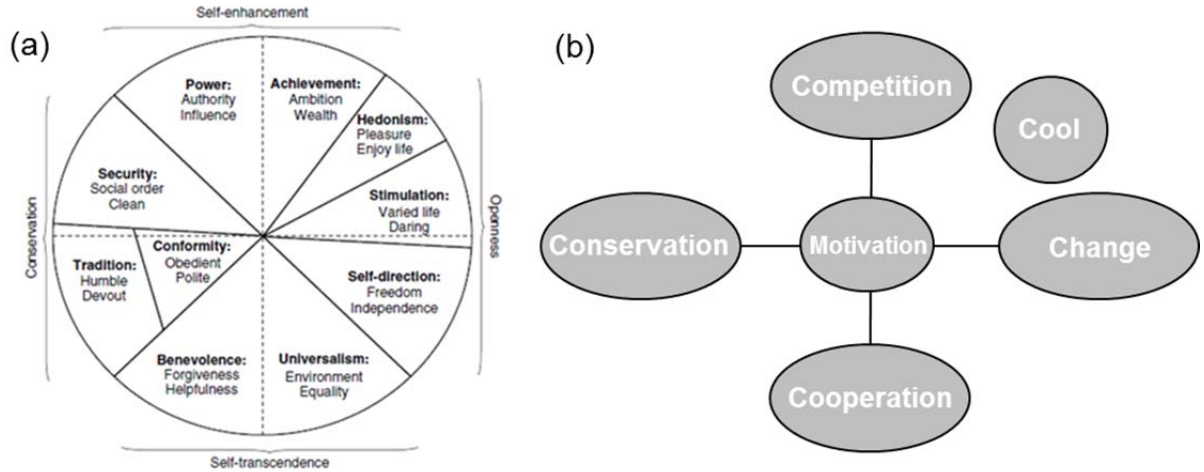


Figure 2: (a) Schwartz's (1992) circular model of personal values, (b) The four+ C's of motivation.

To these four (+ 1) C's, we add a new category named "Interaction". As numerous program of CS are related to environmental issues, some motivation statements proposed to participant in surveys were clearly related to nature connection (e.g. "I want to get outside or connect with nature"; Alender, 2016). This category of motivation was difficult to classify in the previous four (+1) C's and represented a particular motivation for participation in CS programs. Hence, it was decided to add a new category that reflects better this type of motivation.

Finally, we used these six categories of motivations (Conservation, Cooperation, Competition, Change, Cool and Interaction) to classify each motivation described in each article. Deleting the duplicated motivation statements (e.g. two articles asking the same motivation statement), we found 495 (53%) unique motivation statement on the 928 extracted from results of articles. Each of these unique motivation statements was classified under one of the six categories. A cross-validation of this classification is currently implemented by Assaf Shwartz and Liat Levontin to validate the classification made by the first author.

- **Final database of usable data for meta-analysis**

This final database was created to allow the valorization of the data and to provide first results for this STSM. However, some work is still needed to improve this database. For example, due to the high variability of studies quality and information given in each article, it is still hard to know if descriptive data extracted from each article will be actually usable for further analysis. Also, few correlations were found in the extracted data. Among these correlation data, methods also vary, from structural equation modelling (SEM) to simple spearman correlations. Finally, data based on Likert scale were also very complex to handle as some articles provide complete results (e.g. mean and standard deviation values for each question based on a Likert scale) and other studies only provide construct inferred from numerous different items found in questions based on Likert scale.

Hence, in a first part of the results, we present general results on all data extracted from relevant articles. Then we present more detailed results using the meta-analysis framework, but only for percentage data because this type of data is the most obvious to handle in a first analysis. Hence, we pooled results extracted from articles on percentage and count data, calculating these count data as percentage and created a new database (namely “percentage_results”, see Appendix C for a readme) to make this first analysis. Correlations and Likert scale data were kept separately (see “extracted_data_database”).

Description of the main results obtained

As the STSM was mainly focused on literature review and data extraction, main results are actually the different databases obtained. Even if these results are more technical than directly informative, they are nonetheless very useful because it is from these databases that future work could be carried out.

From a general point of view on articles of $k_{relevant}$ (k = 40), we found that studies on motivation of people to participate in CS programs were carried out in 18 different countries. Among these countries, there are mainly developed and Western countries (Figure 2). Also, a significant number of studies (n = 13) analyzed “online” community of people that can hence come from different cultures and countries. These results are consistent with the known world landscape of CS (Braschler 2009).

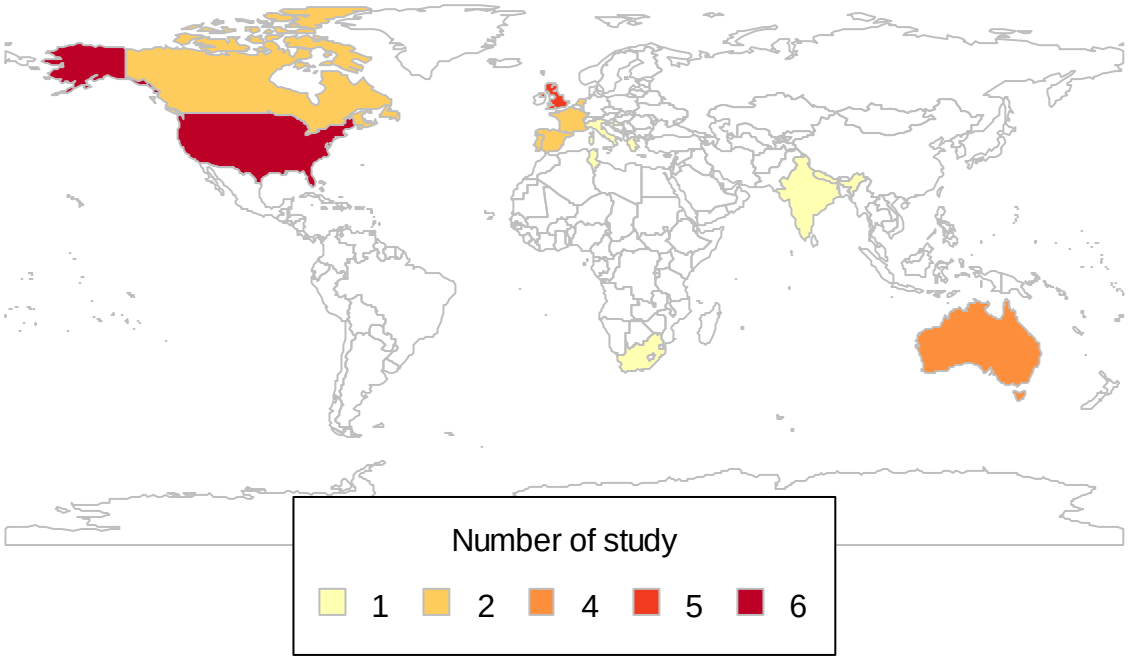


Figure 2: Location on a world map of the CS programs studied in respect with motivation of people to participate in these programs. Note that on $k_{relevant}$ (k = 40), 13 studies are worldwide.

In these countries, motivations of people were studied on 34 different CS programs that are focused on diverse subjects (e.g. astronomy, pollinators, marine ecosystem, air quality, farmland management). In a clear majority of studies, authors used surveys (and particular online surveys) to ask people's motivations. Years of publication of the relevant articles used in this work show that these articles are mostly recent (particularly compared to the general corpus of articles dealing with CS) with a particular peak in 2016 (Figure 3a). This result is expected because of a clear increase in publication of articles related to the field of CS and also because of an increase in the number of study interested in understanding people's motivations towards CS (Follett & Strezov 2015). Ratios of articles dealing with motivation on CS to those related to CS in general show that research on motivation of people to participate in CS programs is not growing faster than the general trend of publication on CS as these ratios stay more or less stable, roughly less than 10% of the general corpus of articles (Figure 3b). The same pattern is observed for publication of articles with quantitative results on motivation. This result indicates that current studies investigating motivation in CS do not rely more on quantitative results than qualitative approaches compared to former studies.

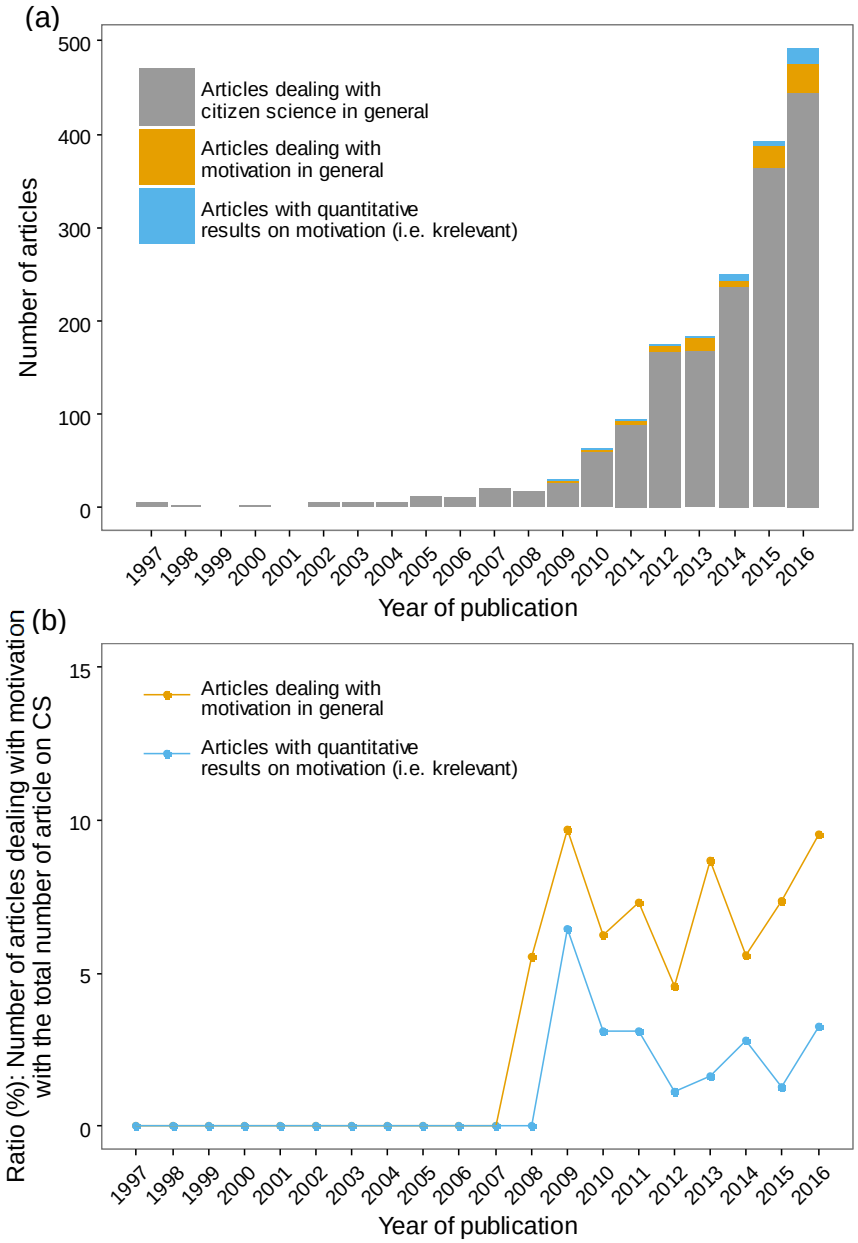


Figure 3: (a) Frequency of the year of publication of articles dealing with citizen science. Articles dealing with motivation to participate in CS programs and those dealing with quantitative results on motivation (i.e. $k_{relevant}$) are highlighted. (b) Ratio of the number of articles dealing with motivation to participate in CS programs and those dealing with quantitative results on motivation (i.e. $k_{relevant}$) to the total number of articles dealing with CS. Note that the y-axis is largely truncated.

Descriptive data of respondents are not well reported in studies on motivation. Data on age of respondents are presented in less than half of all relevant articles. The aggregation of these data is complicated because of the diversity of the way authors chose to present them. Only 10 articles give a mean and a standard deviation of age of respondents. Also, data on gender are presented in half of all relevant studies. We found data on level of education in only 16 (40%) articles and these data were usually incomplete. Also, we found data on occupation of respondents in only 11 (28%) articles and these data were incomplete too. Due to this bad quality of descriptive data of respondents, it seems unfortunately difficult to use them for further analysis.

On the data extracted from results of articles of $k_{relevant}$, survey samples strongly vary among the studies (from 14 to 10992). These samples mainly correspond to the number of respondents to the study survey but in some case, authors used other methods and thus the sample may correspond e.g. to a number of forum posts. For a particular study, the sample size was hence dramatically higher than for the others (Figure 4a) so to better reflect the real distribution of sample sizes, this particular study was remove from the distribution (Figure 4b). This new distribution of sample size shows that a majority of study used quite small sample sizes compared to the entire distribution (Figure 3b) even if the number of studies with less than 30 respondents stays reasonable (9).

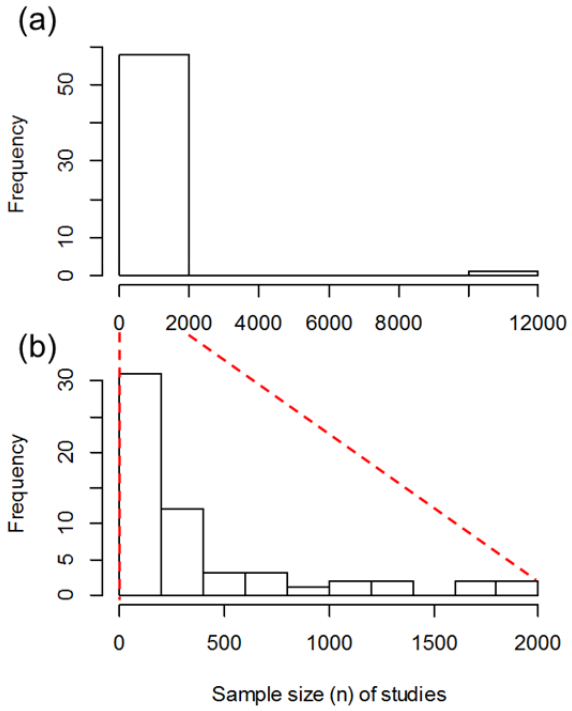


Figure 4: (a) Distribution of all sample sizes of relevant studies on motivation of people to participate in CS programs. (b) Distribution of sample sizes excluding the outlier point (i.e. between 0 and 2000). Frequency of these sample sizes exceed $k_{relevant}$ because a same article can have multiple CS program studied for example.

On the type of data extracted, there are less data presented as correlation than the other types (Figure 5). Data derived from Likert Scale questions are the most reported by authors. However, count data and percentage are considered equivalent data and if count data are calculated as percentage, this latter type of data is clearly the most represented in the analysis of motivation of people to participate in CS programs (27 times found in $k_{relevant}$). Usually, articles contain multiple analysis and types of data.

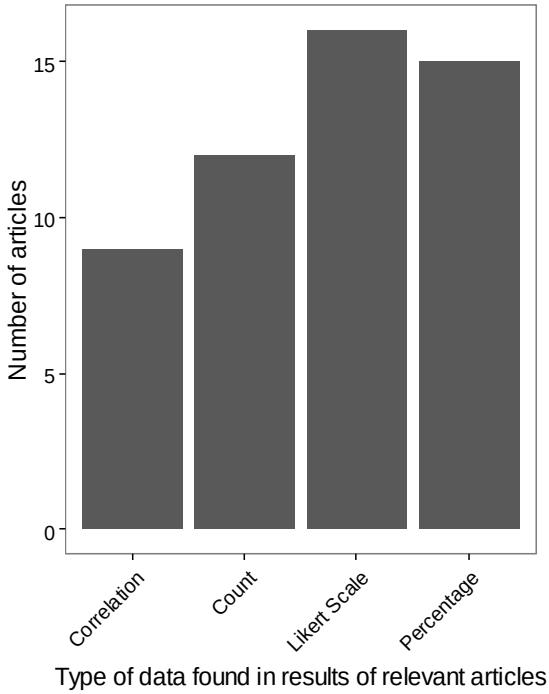


Figure 5: Number of articles that contain the different types of data.

Focusing now on percentage results (“percentage_results” data), the database contains 539 records combining each motivations statement of authors found in the relevant articles categorized in the six different categories of motivation (i.e. the four (+1) C’s and the “interaction” category) and values of percentage that represent the percentage of respondents that chose the motivation statement considered. However, 125 motivation statements were not categorized because not relevant (e.g. when authors proposed a response like “Others” to respondents, this motivation statements cannot be categorized in one of our six categories) or because the motivation statement was not clear (e.g. lacking description). These records were deleted along with 12 records where raw numbers of percentage contained in a figure of the considered article could not be extracted, resulting in a final database of 402 records.

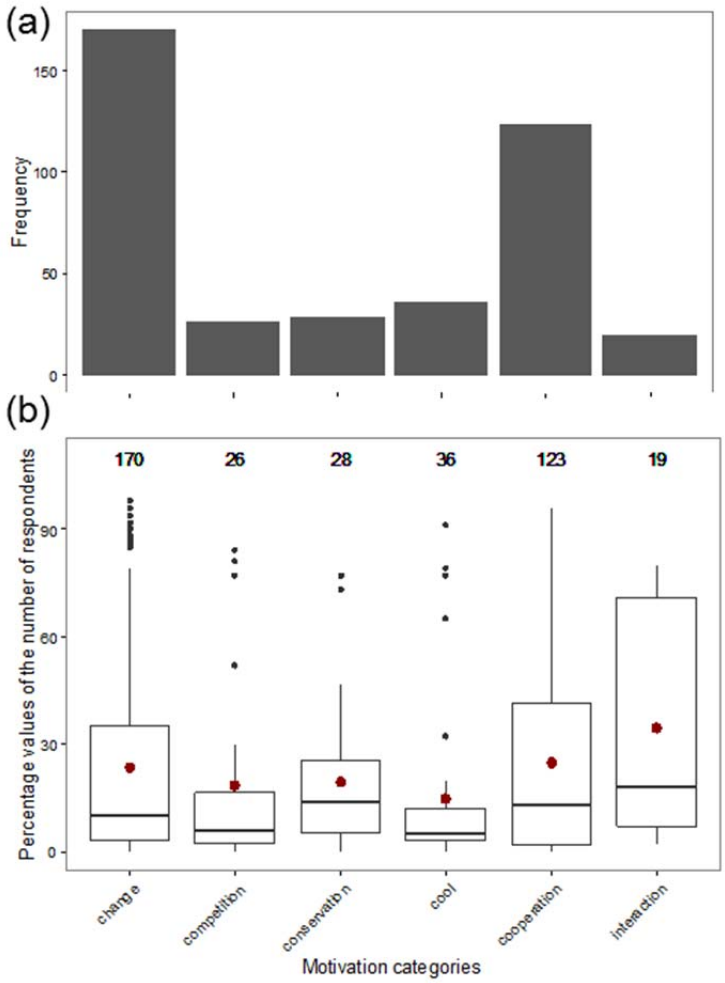


Figure 6: (a) Frequency of each category of motivation asked to respondents and found in relevant articles when presenting their results as percentage data. (b) Distribution of percentage of respondents of each study that chose a motivation statement between the different categories of motivation. E.g. in two particular studies (or surveys), 90% and 25% respectively of respondents chose the response categorized *a posteriori* as "competition". Red points indicate the mean.

Results on percentage data show that usually, motivations statements asked to people in study of $k_{relevant}$ are mainly categorized as "change" or "cooperation" (respectively 42.1% and 30.4%, see Figure 6a). The other motivation statements are roughly distributed in the four other categories (around 7% each). This first results show that authors of studies on motivation put the emphasis on questions related to "change" or "cooperation" motivation categories but neglect the others categories of motivation. This potential bias in questions asked to people should be addressed and corrected in future work on people motivation to participate in CS programs.

The distribution of percentage of respondents that chose a motivation statement between the different categories of motivation (Figure 6b) shows that there is no clear pattern between the different categories of motivation. We found no particular effect of motivation categories [one-way ANOVA, Tukey honestly significant difference (HSD) for unequal sample size; $P > 0.05$ in all cases]. This result means that, on average, people respond to all motivations category with the same frequency and that there is no motivation more likely to be selected by respondents. Hence, focusing on a particular motivation category to attract people to participate in a particular CS program will be ineffective on

average. The best option seems to develop potential tools into a CS program that answer to a majority of the six motivation categories. This should be taken into account in the development of future CS program. However, these first results are preliminary and future work is still needed to consolidate these recommendations, particularly on the definition on the different category of motivations that can bias the results.

Future collaboration with the host institution

The database created during this STSM will be a good basis for future collaboration. Lot of work is still needed to valorize the data. The host institution is clearly implicated in the Working Group 4 (WG4) of the COST Action CA15212 and future collaboration will involve Liat Levontin and Assaf Shwartz for the valorization of the created database. In particular, Deliverable 1 of the WG4 will take clear advantages from this work. Also, a new post-doctoral position on the diversity of the participatory environmental observatories just opened in the home institution and could benefit from the work of this STSM. This post-doctoral will be a good opportunity to pursue collaboration between the host institution and the potential future post-doctorate recruited on the project mentioned.

Foreseen publications/articles resulting from the STSM

After the STSM, numerous potential deliverables could be derived from this database. One of the first foreseen publications will be based on the analysis of the distribution of motivations. From a medium term perspective, the Deliverable 1 of the WG4 will be partly based on the work of this STSM. Finally, this database could be used to write and published a review focus on participants' motivations. This review will help to raise awareness of the citizen science community by showing how scientific evidences on motivation of participants are still scarce and should merit better attention in order to design effective citizen science programs.

References

Alender, B. (2016). Understanding volunteer motivations to participate in citizen science projects: a deeper look at water quality monitoring [WWW Document]. *JCOM - J. Sci. Commun.* URL https://jcom.sissa.it/archive/15/03/JCOM_1503_2016_A04

Braschler, B. (2009). Successfully Implementing a Citizen-Scientist Approach to Insect Monitoring in a Resource-poor Country. *BioScience*, 59, 103–104.

Follett, R. & Strezov, V. (2015). An Analysis of Citizen Science Based Research: Usage and Publication Patterns. *PLOS ONE*, 10, e0143687.

Schwartz, S.H. (2007). Basic human values: theory, methods, and application. *Risorsa Uomo*.

Tinati, R., Luczak-Roesch, M., Simperl, E. & Hall, W. (2016). Because science is awesome: studying participation in a citizen science game. ACM Press, pp. 45–54.

Appendix A

Readme for the three bibliography databases

This readme aims to explain each database constructed in the work flow and also aims to clarify the exact meaning of each sheet and column of each database.

- **Database from Follet and Strezov (2015), (Follet_Strezov_2015_database_final.xlsx)**

Sheet "raw_database":

This sheet contains the raw database obtained from Follet and Strezov (2015).

Column "ID" defines a unique ID for each record (i.e. article).

Column "X" is a former ID that is not used in further analysis. This ID is preserved because it relies on a former ID defined by Follet and Strezov that can be useful to make potential link with their original database.

Columns "Author", "Title", "Source", "Abstract" and "Year" contains primary information on each article.

Columns "Project", "Methodology", "Validation", "Motivation.Effect", "General", "Action", "Conservation", "Investigation", "Virtual" and "Education" referr to categories defined by Follet and Strezov for the purpose of their analysis. A filter is applied on the "Motivation.Effect" column to remove all records that not meet the motivation criteria defined by Follet and Strezov. This filter can be released to see all records (k = 888).

Sheet "relevance_motivation":

This sheet contains only records categorized by Follet and Strezov in their "Motivation.Effect" category. For each record, information is given on their relevance regarding a meta-analysis framework, particularly the presence or not of data.

Columns "ID", "Author", "Title", "Source", "Abstract" and "Year" are duplicated from the previous sheet.

Column "Downloaded" indicates if the PDF of the article was download (1) or not (0). If not, the next column "Remarks" contains the reason, with potential other remarks for other records.

Column "NotRelevant" indicates if an article is considered not relevant (1) for the meta-analysis framework (e.g. no result or only qualitative results). A blank cell indicates that information on this article can be found in another column (i.e. "NotSure" or "RelevanceMetaAnalysis").

Column "NotSure" indicates if an article was hard to classify as relevant (1) and hence needs more expertise. A blank cell indicates that information on this article can be found in another column (i.e. "NotRelevant" or "RelevanceMetaAnalysis"). Also, the next column "NotSure:why" explain why an article was classified as "NotSure".

Column "RelevanceMetaAnalysis" indicates if an article is considered as relevant for the meta-analysis framework (1) or not (0). Some articles may be classified as relevant and also as "NotSure".

This is because the column “RelevanceMetaAnalysis” contains the final choice made on the relevance of the article (e.g. an article firstly classified as “NotSure” has been finally classified as relevant).

Column “Data” indicates where quantitative data can be find in the PDF of the article (pagination is based on the downloaded PDF, see in the PDF folder).

Column “DataType:percentage” indicates if there is data as percentage (1) or not (0). Next columns “DataType:numberof”, “DataType:frequency”, “DataType:correlation” and “DataType:other” indicate information labelled, respectively data as a count (e.g. respondents), as frequency, as correlation or other types of data. A same article can have different types of data.

Column “CitationGoogleScholar” indicates how many times each relevant article was cited based on Google Scholar database (in March 2017). The next column “CitationGoogleScholar:2015-2016” indicates the same information but only for publication in 2015 and 2016.

- **Database from literature review carried out during the STSM (STSM_biblio_database_final.xlsx)**

Sheet “raw_database”:

This sheet contains the raw database of literature related to citizen science searched on Web of Science (wos) and Scopus for years 2015 and 2016.

Column “ID_raw” defines a unique ID for the raw database. Hence, it is possible to know efficiently which articles were deleted in the next steps.

Columns “authors”, “title”, “year”, “journals”, “abstract”, “keywords” and “doi” are obvious columns that contain the information labelled.

Column “type” indicates the type of publication, i.e. if it is an article, a review, a book chapter, etc. based on category defined by WoS and Scopus.

Column “biblio” indicates for each article from which search engine it was found (i.e. wos or scopus).

Sheet “-duplicated-correction”:

This sheet has the same information than “raw_database” but all duplicates were removed. When a duplicate was found between the two search engines, the record from the Web of Science database was preferentially preserved to ensure the best consistency of records. Also, all records labelled as “correction” (i.e. published correction of an article) were removed.

Sheet “-year2014-2017”:

This sheet has the same information than “-duplicated-correction” but all records belonging to year 2014 or 2017 were removed.

Sheet “relevance_motivation”:

This sheet contains the analysis of each article regarding their relevance for the meta-analysis framework (i.e. focused on motivations of participants).

Column “ID” defines a unique ID for each record. This is the final ID for this database. Hence, this ID is used in all subsequent analyses.

Columns “authors”, “title”, “year”, “journals”, “abstract”, “keywords”, “type”, “doi” and “biblio” are equivalent to the previous sheet.

Column “relevanceTitle” indicates if an article was considered potentially relevant for the meta-analysis based on its title (1) or not (0). If the cell is blank, it means than more information is needed to assess the relevance of the article, particularly by reading the abstract.

Column “relevanceTitleAbstract” indicates if an article was considered potentially relevant for the meta-analysis based on its title and abstract (1) or not (0). If the cell is blank, it means than more information is needed to assess the relevance of the article, particularly by reading the entire text of the article (see the next column).

Column “Downloaded” indicates if the PDF of the article was download (1) or not (0). All articles found potentially relevant in the previous column were downloaded.

Column “Remarks” contains potential remarks on each article.

Column “NotRelevant” indicates if an article is considered not relevant (1) for the meta-analysis framework (e.g. no result or only qualitative results). A blank cell indicates that information on this article is found in another column (i.e. “NotSure” or “RelevanceMetaAnalysis”).

Column “NotSure” indicates if an article was hard to classify as relevant (1) and hence needs more expertise. A blank cell indicates that information on this article is found in another column (i.e. “NotRelevant” or “RelevanceMetaAnalysis”). Also, the next column “NotSure:why” explain why an article was classified as “NotSure”.

Column “RelevanceMetaAnalysis” indicates if an article is considered as relevant for the meta-analysis framework (1) or not (0). Some articles may be classified as relevant and also as “NotSure”. This is because the column “RelevanceMetaAnalysis” contains the final choice made on the relevance of an article (e.g. an article firstly classified as “NotSure” has been finally classified as relevant).

Column “Data” indicates where quantitative data can be find in the PDF of the article (pagination is based on the downloaded PDF, see in the PDF folder).

Column “DataType:percentage” indicates if there is data as percentage (1) or not (0). Next columns “DataType:numberof”, “DataType:frequency”, “DataType:correlation” and “DataType:other” indicate information labelled, respectively data as a count (e.g. respondents), as frequency, as correlation or other types of data. A same article can have different types of data.

Column “CitationGoogleScholar” indicates how many times each relevant article was cited based on Google Scholar database (in March 2017). The next column “CitationGoogleScholar:2015-2016” indicates the same information but only for publication in 2015 and 2016.

- **Database of articles from citation of previous relevant paper (citation_biblio_database_final.xlsx)**

Sheet "relevance_motivation":

This sheet contains the relevant articles for the meta-analysis that were found in articles that cited the relevant article of the two previous databases.

Column "ID" defines a unique ID for these records (in the continuity of the ID of "STSM_biblio_database_final.xlsx").

Column "citationFromID" indicates the ID of an article cited by a new relevant reference found.

Column "referenceFromID" indicates the ID of an article from which a new relevant reference was found.

Columns "authors", "title", "year", "journals", "abstract", "keywords" and "doi" are obvious columns that contain the information labelled.

Column "type" indicates the type of publication, i.e. if it is an article, a review, a book chapter, etc. based on category defined by WoS and Scopus.

Column "biblio" indicates for each article how it was found (i.e. using the Google Scholar "cited by" function for citation or in reference of a relevant article).

Column "Downloaded" indicates if an article was download (1) or not (0).

Column "Remarks" contains potential remarks on each article.

Column "RelevanceMetaAnalysis" indicates if an article is considered as relevant for the meta-analysis framework (1) or not (0).

Column "Data" indicates where quantitative data can be find in the PDF of the article (pagination is based on the downloaded PDF, see in the PDF folder).

Column "DataType:percentage" indicates if there is data as percentage (1) or not (0). Next columns "DataType:numberof", "DataType:frequency", "DataType:correlation" and "DataType:other" indicate information labelled, respectively data as a count (e.g. respondents), as frequency, as correlation or other types of data. A same article can have different types of data.

Column "CitationGoogleScholar" indicates how many times each relevant article was cited based on Google Scholar database (in March 2017). The next column "CitationGoogleScholar:2015-2016" indicates the same information but only for publication in 2015 and 2016.

- **Reference cited:**

Follett, R. & Strezov, V. (2015). An Analysis of Citizen Science Based Research: Usage and Publication Patterns. PLOS ONE, 10, e0143687.

Appendix B

Readme of the database containing extracted quantitative data

This readme aims to explain the database that was built from the extraction of data from each relevant article. It also helps to clarify the exact meaning of each sheet and column of this database.

Sheet "general_info":

This sheet contains general information on each article like Authors, Title, Year, etc. The column "ID" defines a unique ID for each record (i.e. article). This ID is based on previous bibliographical analysis (see Readme_bibliography_database).

Sheet "data_methods":

This sheet contains all general information on methodology used in each article like the location of the study, type of sampling, CS program name, etc. The column "descriptiveData" indicates whether there is descriptive data (1) or not (0) in the article. These data are stored in the next sheet. The column "numberOfProgram" indicates the number of different CS program studied in the considered article. However, some articles do not focus on particular CS programs but are more focused on general public motivation to participate in CS. In this case, "general public" is written in the column.

Sheet "data_descriptive":

This sheet contains descriptive data of respondents of the survey like the "n" (columns "n:survey" and "n:motivation" indicate the total number of respondent in the survey and the total number of respondent to questions on motivation in the survey, respectively). Age data (if applicable) are stored in the columns "age:typeOfData" and "age:data". Gender data (if applicable) are stored in the columns "sexe:typeOfData" (e.g. count or percentage), "sexe:male" (raw number), "sexe:female" (raw number) and "sexe:noresponse" (raw number).

If there is any information in columns "levelOfeducation" and "occupation", it means that there is data on level of education of respondents and on their occupation. These data are stored in the two next sheets and the type of these data is indicated here (e.g., if "percentage" is indicated in the column "occupation" for a particular article, it means that there is data on occupation in the sheet "occupation" of the database for this particular article and that the raw numbers found in the sheet "occupation" are percentage).

The column "program" is used to make a difference between CS programs within the same article (i.e. some articles deal with more than one CS program, see data_methods sheet, column "numberOfProgram").

Sheet "level_of_education":

This sheet contains data on level of education of respondents. Column "n:levelOfEducation" indicates the number of respondents to questions related to their level of education. This "n" may be different from the "n" of the study or the "n" of motivation responses. Column "description:levelOfEducation" indicates the level of education as written by the authors of the corresponding article. Column "data:levelOfEducation" contains raw data. The type of data, like "percentage" or "count", is found in the data_descriptive sheet. The final column provides potential remarks.

Sheet "occupation":

This sheet contains data on occupation of respondents. Column "n:occupation" indicates the number of respondents to questions related to their occupation. This "n" may be different from the "n" of the study or the "n" of motivation responses. Column "description:occupation" indicates the occupation as written by the authors of the corresponding article. Column "data:occupation" contains raw data. The type of data, like "percentage" or "count", is found in the data_descriptive sheet. The final column provides potential remarks.

Sheet "data_type":

This sheet contains information on the type of motivation-related data found in the result section of each article. There are four different types of data:

- percentage: results on motivation are presented in the considered article as percentage (e.g. "6% of respondents are motivated to participate in this CS program because they want to learn more about science")
- count: results on motivation are presented in the considered article as count data (e.g. "70 respondents are motivated to participate in this CS program because they want to learn more about science"). If the "n" of the survey is provided, these results can be easily calculated as percentage. However, in this database, data are kept as raw as possible.
- correlation: results on motivation are presented in the considered article as a correlation with another studied variable (e.g. a study presenting the relationship between motivation of respondents and their age using a Spearman's rank correlation coefficient).
- Likert Scale data: results on motivation are based on Likert scale questions on the survey and results are presented for example as a mean of scale values chose by respondents.

Each corresponding column indicates whether this type of data is contained (1) or not (0) in the considered article. All raw data are then stored in the corresponding next sheets (e.g. sheet raw_percentage for percentage data).

Sheet "raw_percentage":

This sheet contains all raw data of results found as percentage values. A new ID (column "ID_percentage") is created because these data are then used in further analysis. Column "motivation" indicates the motivation as expressed and written by authors of articles in their results. Column "percentage" indicates the percentage values. Column "n:percentage" indicates the number of respondents. Column "category" refers to the proposed classification of each motivation. Finally, column "remarks" provides some potential information on each article (e.g. results for different CS program). If "from figure" is written in this column, it means that data are inferred from figure.

Sheet "raw_count":

This sheet contains all raw data of results found as count data. A new ID (column "ID_count") is created because these data are then used in further analysis. Column "motivation" indicates the motivation as expressed and written by authors of articles in their results. Column "count" indicates the count data. Column "percentage" is the calculated percentage values from the count. Column "n:count" indicates the number of respondents. Column "category" refers to the proposed classification of each motivation. Finally, column "remarks" provides some potential information on each article (e.g. results for different CS program or between genders). If "from figure" is written in this column, it means that data are inferred from figure.

Sheet "raw_correlation":

This sheet contains all raw data of results found as correlations. A new ID (column "ID_ correlation") is created because these data are then used in further analysis. Column "n:correlation" indicates the number of data. Column "dependentVariable:motivation" indicates the motivation as expressed and written by authors of articles and used as the dependent variable of the analyzed relationship. Column "IndependentVariable" indicates the variable used as the explicative variable of the relationship as expressed and written by authors of articles. Column "correlation" gives the correlation values between the previous mentioned dependent and independent variables. Column "pvalue" indicates, if applicable, the significance of the correlation values. Column "correlationType" indicates the type of analysis (e.g. Spearman's rank correlation coefficient, correlation of Structural Equation Modelling (SEM), regression analysis, etc.). Column "category:DependantVariable" refers to the proposed classification of each motivation and column "category:IndependantVariable" is a proposed categorization of the independent variable used in the correlation analysis, e.g. age or activity measures of people doing CS program. Finally, column "remarks" and "remarks2" provide some potential information on each article (e.g. results for different CS program or between genders).

Sheet "raw_likertScale":

This sheet contains all raw data of results based on Likert Scale survey questions. A new ID (column "ID_ likert") is created because these data are then used in further analysis. Column "motivation" indicates the motivation as expressed and written by authors of articles in their results. Column "likertScaleData" gives values calculated from Likert Scale results as written by authors of article. Usually, these values are the mean of the value on the Likert scale chosen by respondents (if applicable, standard deviations of mean are given in the column "likertScaleSD"). However, these data can also be of other types (e.g. a construct inferred from different items). See next columns for more information. Column "likertScaleRange" indicates the range of the scale used in the survey question (e.g. from 1 (very unlikely) to 5 (very likely) will be coded "1-5"). Column "construct" indicates with a "1" if the "motivation" that is considered in the respective column is a construct inferred from different items by authors of the article or indicates to which construct the "motivation" is related. If the cell is blank, it means than there is no construct in the considered article. Column "cronbach'sAlpha" indicates (if applicable) the cronbach's Alpha value of related construct. Column "n:likertScale" indicates the number of data. Column "factor:eigenvalues" indicates (if applicable) the eigen values of the factor inferred from different motivation items. Column "category" refers to the proposed classification of each motivation. Finally, column "remarks" and "remarks2" provide some potential information on each article (e.g. results for different CS program or between genders).

Appendix C

Readme of the database containing analyzed percentage data (“percentage_results”)

This readme aims to present the database used in the analysis of percentage data (i.e. count and percentage data pooled together). This database was created from the “extracted_data_database” in order to be managed in the R environment and to conduct some statistical analysis.

The database is composed of one sheet named “percentage_results”.

The first column is a new ID named “ID_data” to facilitate management of the dataset on the R environment. The two next columns are previous ID. The column “percentage” indicates percentage values of the extracted results. The column “n:percentage” indicates the sample size of each study. The column “category” indicates to which category of motivation the percentage value is referred. The column “remarks” contains potential useful information to avoid bias and to manage the dataset for further analysis on the R environment.