**Keeping it simple: An experiment in online classification design**

Alexandra Swanson, Helen Spiers, Lucy Fortson, Ted Cheeseman, Chris Lintott

**Abstract**

Volunteers completed the Yes/No workflow in half of the time it took to complete the Survey workflow, with no cost to accuracy or volunteer engagement, as measured by classification rates or participation in the discussion forum.

**Introduction**

The Zooniverse is an online citizen science platform that engages volunteers by asking for help in the interpretation of large-scale datasets. At [www.zooniverse.org](http://www.zooniverse.org), researchers from across disciplines ask volunteers to apply their pattern-recognition skills to classify what they see in images or videos, such as the species of wildlife captured in camera traps or the morphology of galaxies in telescope imagery (Cite ALL THE THINGS). Since its inception with the launch of [www.galaxyzoo.org](http://www.galaxyzoo.org) in 2007, the Zooniverse has grown to include over 1.6 million registered volunteers who participate across more than 100 different research projects; the need for citizen science in data interpretation shows no signs of abating. Since the launch in 2015 of a project builder platform (www.zooniverse.org/lab), which allows researchers to launch projects without customized software development, the number of research projects within the Zooniverse has more than doubled (Figure 1). The accelerating production of big data and concomitant growth of citizen science projects threaten to outpace the growth of the volunteer community; citizen science projects need to become more efficient and more engaging to keep pace with demand for data analysis.

One obvious solution is to combine human and machine classification; Wright et al. (2017) recently demonstrated that the combination of human and machine classifiers for a deployed Zooniverse project outperformed either one alone. Another solution would be to find new ways to engage volunteers, for example through a mobile app. However, the diversity of citizen science projects makes it hard to find a general solution. For example, machine learning solutions developed for one complex classification may not easily be transferred to other projects, as seen with Norouzzadeh et al.’s (2017) convolutional neural networks that successfully identify Serengeti wildlife but have yet to be applied to other species. Similarly, some projects cannot be reduced to interfaces simple enough for mobile development, such as ecology projects (e.g. [www.snapshotserengeti.org](http://www.snapshotserengeti.org), [www.cameracatalogue.org](http://www.cameracatalogue.org)) that ask volunteers to identify which of over forty different species is present in an image.

Here we introduce “cascade filtering” as an alternative approach to more traditional image classification projects. We introduce this term to describe the conversion of a complex task into a series of simple binary (i.e. yes or no) questions presented separately to volunteers. By reducing all projects to a similar set of problems, cascade filtering allows projects to deploy common approaches to better leverage mobile technology and machine learning methodologies. A traditional species identification task could, for example, be broken down into a series of questions that asks first about the presence of any animal, then about the presence of a specific animal or group of animals, and so on, with each binary question successively filtering the dataset down to the images the research team needs (see Figure 2).

Reducing a complex task into binary questions enables the integration of simpler machine learning for automated image identification at every step of the filtering process (see e.g. Allwein et al 2000). It also enables better portability to mobile devices: the classification task is simpler, thus requiring less screen “real-estate,” and can also be reduced to gross motor movements such as “swiping” left or right. Thus, although images would ultimately require more classifications to produce the same amount of final information, the ease of classification and ability to integrate simple machine learning routines could provide benefits to efficiency that far outweigh the increase in classification counts.

However, altering project design can have wide-ranging and unanticipated impacts on the nature of volunteer participation. For example, a study of volunteer behavior on *Snapshot Serengeti* (www.snapshotserengeti.org), which asks volunteers to identify African animals as one of >50 species (Swanson et al. 2015), revealed that volunteers spent more time classifying when they received high proportions of “blank” images, which took less time and effort to classify (Bowyer et al. 2016). Recent work by Spiers and Lintott (in prep) documents farther-reaching consequences on the volunteer community: On *Supernova Hunters,* the regular release of small datasets that were quickly completed created highly skewed volunteer contributions and may explain the overwhelmingly older-male demographic of the project. Thus, changing the classification methodology from a traditional survey design to a cascade filtering approach could impact data quantity and quality as well as the nature of volunteer communities contributing to a project.

In this paper, for the first time, we directly compare classification paper for two implementations of the same project: one using a ‘traditional’ design, which solicits a full set of answers, and the other utilizing a cascade design. While models based on gamification would suggest that the introduction of a repetitive, simple interface would increase participation (CITE), previous studies of complex citizen science projects have found that gamification of this kind may be problematic (Eveleigh et al. 2014). If volunteers are motivated directly by a desire to contribute to research, it seems possible that they will seek out experiences which seem to allow them to contribute most, providing all the information they can for each image.

We evaluate the effects of these different designs within a single Zooniverse project, *Snapshots at Sea* ([www.snapshotsatsea.org](http://www.snapshotsatsea.org)), which asks volunteers to identify humpback whale flukes (tails) from images contributed by tourists. By providing volunteers with a choice of workflows with which to classify, we were able to evaluate differential effects of the alternative workflows on the project efficiency and accuracy, as well as effects on the volunteer community. The results have implications for researchers relying on citizen science for data interpretation and for those making design decisions related to the structure of such tasks, whether in-browser or in-app and, especially when running such a project in tandem with machine learning.

**Methods**

***Data Collection***

*Snapshots at Sea* was originally launched as a cascade filtering project on June 9, 2016 to identify images containing humpback whale flukes from photographs taken by tourists and researchers on board cruise ships. The undersides of humpback whale flukes are individually identifiable by color pattern and shape, which allow for high resolution population monitoring and conservation research (TED – citations). Images containing humpback whale flukes with visible undersides are passed to another citizen science project, Whales as Individuals, where volunteers help prepare the photos for individual recognition algorithms.

For this study, we integrated the binary questions into a more traditional “survey” design that allowed volunteers to select species from a list of multiple options, which is standard for ecology projects (e.g. Swanson et al. 2015). To ensure that images on each workflow were similar in content and appeal, we acquired a set of 13,264 images taken by cruise passengers on board X, randomized them into two subsets of 6,632 images each, and assigned them to either the standard “Survey” workflow or the cascade filtering “Yes/No” workflow. Because each individual Yes/No question provides less information than the survey task, the cascade filtering approach ultimately requires more classifications to complete the same number of images (see Figure 2 for a detailed description of each workflow, retirement rules, and efficiency).

*‘Yes/No’ workflow*

The cascade filtering workflow was comprised of four separate binary (“Yes” or “No”) questions, of which only one question was active on the website at a time (detailed in Figure 2). Each image was seen by five volunteers prior to retirement at each stage. Following retirement of all images associated with one question, all images of interest were identified based on the percentage of volunteers who answered “yes” to that question. Images of interest were then taken forward, forming the subject set of the next question. All 6,632 images were passed to Q1; images with 75% agreement on the presence of an animal were passed to Q2 (n = 5,910); images with 75% agreement on the presence of a whale or dolphin were passed to Q3 (n = 5,020); images with 50% agreement on the presence of a fluke were passed to Q4 (n = 1,470). These thresholds were decided by the project lead during early data validation on *Snapshots at Sea*. The limits were devised to rapidly filter the dataset while minimizing false negatives of images that could eventually be used to identify whales; because Q3 and Q4 had higher rates of false negative error among raw classifications, the lower percent-agreement requirements ensured that “borderline” images were included in the final dataset.

*‘Survey’ workflow*

The survey workflow provided a more direct route to classification by presenting volunteers with a list of simultaneous options for every image: “Whale or Dolphin,” “Seal, sea otter, etc.,” “Bird,” “Fish,” and “No animal.” When volunteers select one of these options, a pop-up screen appeared with reference photos and additional details about their selection, and they could choose to select either “cancel” or “identify.” If volunteers selected “Whale or Dolphin” from the menu, the pop-up screen questions: “Is the underside of the tail visible?” (answers: “yes” or “no”), and “If YES, does the tail belong to a humpback whale?” (answers: “yes” or “no”). Volunteers could select and identify multiple species in an image. When they were done classifying on one image, they submitted their identifications by selecting “Done” below the species list and were then shown the next randomly selected image. Because the Survey workflow task was more complex and combined multiple questions into a single task, images were shown to 10 volunteers before being retired from circulation.

*Volunteer communication*

We launched the experiment on June 7, 2017, and announced it via newsletter on June 8, 2017. Because *Snapshots at Sea* already had an active volunteer community used to the cascade filtering approach, we sent one email to the existing community announcing new data and a new, alternative workflow, and one email to volunteers who had previously participated the now-retired, marine-focused project *Seafloor Explorer*.

When Question 1 of the Yes/No workflow was completed, we activated Question 2 and sent out another newsletter announcing both new data and the alternative survey workflow. This newsletter went to all volunteers who had classified on *Snapshots at Sea* before or during the experiment, thus including all newly recruited volunteers. No newsletters were sent for the launch of Question 3 or 4. After all four questions on the Yes/No workflow were completed, we de-activated the Yes/No workflow and sent out a final newsletter to the *Snapshots at Sea* community alerting them to the remaining data on the survey workflow. When both workflows were complete, we emailed all *Snapshots at Sea* volunteers asking them to complete a survey about their demographic information and workflow preferences (See Appendix S1 for the full questionnaire and responses).

***Data analysis***

*Classification rates and volunteer engagement*

We retrieved all classifications made on *Snapshots at Sea* during the study and associated metadata, including volunteer user name (recorded as “not-logged-in” if an unregistered volunteer), the date and time of the classification (as recorded in the server logs), and the *device* (defined from the volunteer’s browser metadata) used to classify. We also tracked volunteer activity on the associated *Talk* forum. We classified registered volunteers as *existing*, if they had classified on *Snapshots at Sea* prior to launching the survey workflow, and as *new*, if their first classification was during the experiment. We calculated the time to complete a classification (*duration*) as the time from when the image was loaded on the volunteer’s screen (as stored in the front end) to when the volunteer submitted the classification; classifications with *duration* > 2 minutes were excluded as these typically represent incidents when the volunteer had loaded a classification but was not actively classifying. We identified a *session* asperiod of continued classifications with less than 30 minutes of inactivity between consecutive classifications. All analyses were conducted in Program R 3.4.0 and code is available online at [www.github.com/aliburchard/snapsatsea](http://www.github.com/aliburchard/snapsatsea).

We evaluated skew among volunteer contributions by calculating Lorenz curves (package *ineq*, function *Lc*) and Gini coefficients (package *reldist*, function *gini*) for each workflow for *new* and *existing* volunteers. Lorenz curves show the proportion of classifications vs. proportion of volunteers; y = x represents hypothetical perfect equality, where every volunteer contributes the same number of classifications. The Gini coefficient is a measure of inequality, calculated as G = A/(A+B), where A is the area above the Lorenz curve and below the 45º line, and B is the area below the Lorenz curve). The Gini coefficient ranges from 0 to 1, where 0 represents completely equal contributions and 1 represents completely unequal contributions.

*Aggregation & Accuracy*

We evaluated agreement between volunteer answers and expert answers contributed by the project’s scientific lead, T.C. We calculated raw agreement as the number of classifications that agreed with the expert classification.

Because the high-levels of accuracy on Zooniverse projects relies on the aggregation of multiple volunteer classifiers, we also applied very simple aggregation algorithms to produce a single “consensus” answer for every image. For each binary question, we calculated the proportion of volunteers who answered “yes” for a given image and calculated the consensus answers as “yes” if the proportion exceeded 0.75 for Q1 and Q2, 0.45 for Q3, and 0.5 for Q4, which are the same requirements for passing images into subsequent workflows.

While more sophisticated approaches exist for aggregating survey-style tasks (Swanson et al. 2014), we applied an analogous approach to the survey task to facilitate comparison between the workflows. For every image, we recorded the proportion of volunteers who identified the presence of any animal, a whale or dolphin, a fluke, and whether or not the fluke belonged to a humpback, and calculated consensus answers according to the same thresholds as the Yes/No workflow.

**Results**

*Classification rates & volunteer engagement*

Volunteers completed the Yes/No workflow twice as quickly as the Survey workflow, contributing the 95,195 classifications needed in only 21 days, compared to the 41 days required to acheive the 66,320 classifications needed to complete the Survey workflow (Table 2, Figure 3). Classification rates following newsletters spiked for both workflows, but the increases were three times higher for Yes/No questions: the Yes/No workflow reached a maximum hourly classification rate of 3,000 classifications/hour, compared to 1,000 classifications/hour on the Survey workflow.

Daily per volunteer contribution was higher on the Yes/No workflow (9.3 classifications per person per day) than on the Survey workflow (2.75 classifications per person per day). The Yes/No classifications were much faster to complete (Figure 4): the median duration for a classification ranged from 8.2 – 10.7 seconds for survey tasks and 1.0 – 3.7 seconds for a Yes/No task, depending on the device used to classify. Although individual classifications were faster, volunteers spent more time classifying when working on Yes/No tasks: median session length for volunteers classifying on exclusively Yes/No questions was 29 minutes compared to 21 minutes when classifying on both tasks and 16.7 minutes when classifying exclusively on the Survey task.

There was greater inequality in volunteer contributions on the Yes/No workflow than on the survey workflow (Figure 5a), regardless of whether the volunteers were *existing* volunteers (those who had classified prior to the experiment) or *new* volunteers (who joined *Snapshots at Sea* during the experiment). Gini coefficients ranged from 0.762 (new users) to 0.745 (existing users) for the survey task and 0.825 (new users) to 0.828 (existing users) on the yes/no questions. The increased skew on the Yes/No workflow classifications appears to be driven primarily by increases in the number of users contributing >100 classifications (Figure 5b, Table 3), though also reflect increases in the number of users contributing <10 classifications.

For the 31 days in which both workflows were available on the website, a similar number of volunteers contributed to both workflows as contributed to Yes/No only and Survey only (Table 2). However, volunteers who contributed to both workflows classified many more images and participated more actively on the discussion forums than those who only classified on one workflow or the other: 14% of volunteers who classified on both workflows commented on *Talk*, while only 4.7% of Survey-only and 3.7% of Yes/No-only volunteers commented on *Talk*.

*Accuracy*

Accuracy was similar across workflows (Table 4). The accuracy of raw volunteer classifications varied dramatically across questions, ranging from 69-99% agreement with expert data. The aggregated answers were much more consistent, ranging from 90-100% agreement. Note that most of this error came from false positives, e.g. the incorrect identification of an animal where none was present, which would be easily filtered out of the final dataset by subsequent questions. Altering the number of required classifications per subject (currently set n=5 for Yes/No and n=10 for Survey) or the threshold agreement levels within the aggregation algorithms could reduce false negatives or otherwise improve overall accuracy. However, our goal here was not to improve the accuracy of this project, but rather to test whether there was a difference in accuracy between the different project designs.

*Follow-up survey*

We received 128 responses to the follow-up survey. According to the survey responses, *Snapshots at Sea* volunteers were twice as likely to be female as male, and 50% of volunteers had at least a university degree (Figure 6, Supplementary Table 1). Overall, 40% of volunteers preferred the Yes/No workflow and only 20% preferred the Survey workflow, but the proportion of volunteers who preferred each workflow did not vary significantly across age, gender, or education levels.

**Discussion**

Cascade filtering provides an effective and efficient route for engaging volunteers in data interpretation. Here we demonstrate that this alternative approach to more traditional, complex citizen-science projects engages volunteers in both classification and discussion, quickly producing accurate data for scientific use. We recommend that researchers consider implementing cascade filtering for their own citizen science projects, but raise some points for consideration when deciding on project design.

*Does cascade filtering work?*

Volunteers completed the Yes/No workflow in half of the time it took to complete the Survey workflow (Figure 3). This increase was driven by dramatically higher rates of classification per volunteer, as opposed to a greater number of volunteers: volunteers on the Yes/No workflow spent twice as long classifying in a single sitting (Table 2) and contributed three times as many classifications per day (Table 1).

Importantly, this acceleration did not come at the cost of data quality: levels of accuracy were similar across both workflows (Table 4). Individual accuracy, however, varied dramatically across the questions within each workflow; it was especially variable on the Survey workflow, ranging as low as 69% agreement on the presence of a humpback. This could suggest that the difficult tasks are easier to answer when isolated, perhaps because the task of learning a complex interface could interfere with the accuracy of the answers given (CITE).

Cascade filtering did not have a clear impact on deeper volunteer engagement. Volunteers who participated exclusively on the Yes/No workflow were just as likely to engage on the discussion forum as those who participated on the Survey workflow (Table 2), though they made far fewer (one-third as many) comments. The implications of this difference are difficult to distill because volunteers who classified on *both* workflows were three times more likely to participate on *Talk* and commented an order of magnitude more frequently than those who participated exclusively on either workflow; unfortunately, it was impossible to track which workflow they were classifying on when they chose to comment.

The speed and popularity of the collaborative filtering approach likely derives from the simplicity of the task and the speed and ease with which volunteers are able to make a tangible contribution. The average Yes/No task took anywhere from one-tenth to one-half the time of a Survey task (Figure 4), depending on the complexity of the image and the device used. These tasks were also more predictable: because only one question was active at a time, volunteers knew whether they would be asked to identify the presence of an animal (generally easy) or identify a tail as belonging to a humpback (generally somewhat difficult).

Peoples’ preference for simple tasks is well documented in the gaming literature (CITE) and has been demonstrated on other Zooniverse projects such as Snapshot Serengeti. [1 sentence or so on something about “getting in the flow” – this being simple and addictive] A reluctance to “break the flow” could also explain why volunteers comment less often from the Yes/No workflow.

*Should you implement cascade filtering on your citizen science project?*

We recommend that research teams consider cascade filtering as a potential project design for their citizen science projects. Beyond engaging volunteers to producing rapid results, cascade filtering also produces standardized binary classifications that are significantly easier to parse and analyze. Thus, rather than designing project-specific algorithms that have to accommodate deal with long, nested lists of complex data types, projects can leverage simple, generalizable code. The standardized format of binary classifications also enables easier integration of machine-learning routines at every stage within a cascade filtering workflow. This step would be critical for projects with very large data sets wishing to implement significant acceleration of classification results while retaining high levels of per-species classification completeness as well as purity (low levels of false negatives).

However, cascade filtering is not a panacea, and researchers should consider several limitations when deciding whether to implement this approach. First, we have only demonstrated the efficacy of cascade filtering for relatively simple projects with a small set of binary questions. With only four binary questions, *Snapshots at Sea* doubled the speed of data completion; and crude extrapolation of the rates from this experiment suggest that projects would cease to see faster data completion times by ~8 binary questions. Many traditional Survey-style projects (e.g. [www.snapshotserengeti.org)](http://www.snapshotserengeti.org)) ask volunteers to differentiate from more than 50 different species, so the sheer number of binary questions required to convert a standard survey project into an exclusively cascade filtering approach could outweigh any increases in efficiency.

Second, project design can impact the volunteer community (e.g. Spiers & Lintott in prep) and cascade filtering produced more unequal volunteer contributions than the Survey workflow (Figure 5). This was true regardless of whether or not volunteers had previously classified on *Snapshots at Sea*, suggesting that the skew resulted from the nature of the task and was not driven by volunteers seeking out what was familiar. Thus, changing the workflow design has a measurable impact on the volunteer community, though levels of skew on both *Snapshots at Sea* workflows were well within the bounds of other successful Zooniverse projects, which range from 0.7 - 0.9 (Cox et al. 2015, Spiers & Lintott in prep).

We should also note, that while it seems obvious that a citizen science project aimed at engagement should seek to minimize inequality (CITE?), such skew is not intrinsically bad. Examining the underlying volunteer contributions (Figure 5b) reveals that this inequality appears driven more by an increase in “super users” (who contribute many classifications) than an increase in “visitors” (who contribute very few). In general, increasing the amount of time that volunteers participate on projects leads to a variety of benefits. For example, volunteer accuracy in well-designed projects tends to correlate with the number of classifications (e.g. SpaceWarps, Marshall et al. 2015; Snapshot Serengeti, Swanson et al. 2016). Increased participation also correlates with increased understanding of project-specific content, suggesting that people are actively learning about science through their participation (Masters et al. 2016).

Third, volunteers show clearly divergent preferences for different styles of contribution. The follow-up survey revealed that the 40% of volunteers who preferred the Yes/No tasks did so mostly because the questions were quick, easy, and immediately rewarding, while the 20% of volunteers who preferred the Survey workflow did so because they were able to contribute a more robust and meaningful classification. These divergent preferences were not explained by differences across age, gender, or education level (Figure 6), but do highlight the diversity of preference in the volunteer community. While cascade filtering may, on one hand, lower the barrier to entry by providing a simple route to engagement (CITE), the binary questions might not be as enriching or provide as many learning opportunities. Changing the *way* in which volunteers spend their time on a project could change learning outcomes in unintended ways, although longer-term studies would be needed to evaluate these impacts.

Clearly, not all citizen science projects can or should attempt to implement and exclusively cascade-filtering approach. However, integrating both approaches into a project could simultaneously increase data throughput and provide multiple paths to engagement. The Zooniverse project builder ([www.zooniverse.org/lab)](http://www.zooniverse.org/lab)) makes it straightforward to build multiple routes to classification for a single project. Since completing this test on *Snapshots at Sea*, this strategy has been implemented via the introduction of two binary workflows on the *Panthera* conservation organization’s *Camera CATalogue* project ([www.cameracatalogue.org)](http://www.cameracatalogue.org)). Images are first passed through an “Empty or Not” workflow, and then to a “Vehicle or Not” workflow, before continuing on to the multi-species survey task. Images that are empty or contain only vehicles represent roughly half of the 30-million image dataset (Ross Pitman, pers comm).The binary filtering steps are completing at X times faster than the full survey task, potentially reducing the project’s time to completion by X%.

While more work needs to be done to assess potential long-term impacts to the volunteer community, this study demonstrates that cascade filtering provides a rapid, efficient, and accurate alternative to traditional project design, and provides clear benefits for both research teams and the volunteer community.

**Tables and Figures**

**Figure 1:** Growth of Zooniverse projects through time. Each point and vertical line reflect the launch of a new project. Gray and black dashed line denotes the launch of the Zooniverse project builder platform, which enables research teams to build their own Zooniverse project without custom web development.



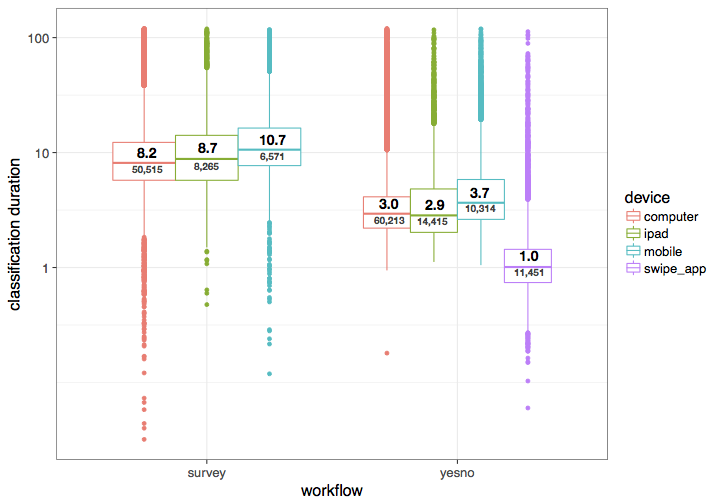
**Figure 2:** *Snapshots at Sea* workflows and retirement rules.



**Figure 3:** Instantaneous (a) and Cumulative (b) classifications through time for each approach. Cumulative classifications are plotted as the proportion of total required to complete the dataset (66,320 for the Survey workflow and 95,195 for the Yes/No workflow). Data additions are indicated in gray dashed vertical lines, newsletters in blue. The completed Yes/No workflow was completed on June 27 and de-activated (i.e. removed from the website) on July 5.



**Figure 4:** Classification duration in seconds (calculated as time that an image loaded into a volunteer’s browser to the time that “Done” was selected) for each workflow across different devices. Median durations given in black text above the median line; sample sizes given in gray text below the median line. Note that the Y-axis is on a Log10 scale.



**Figure 5:** (a) Lorenz curves for each workflow. Registered volunteers (i.e. those with Zooniverse accounts) were categorized as *existing* if they had classified prior to the experiment and *new* if they only began classifying during the experiment. (b) Probability density distributions for classifications per registered volunteer for each workflow (note that the x-axis is plotted on a log-scale).



**Figure 6:**





****

****

****