#### Supporting Information

##### Supporting Methods

###### Zooniverse project design interface and data management

Videos for this study were originally hosted on Zooniverse’s Ouroboros Application Programming Interface (API) (<https://github.com/zooniverse/Zooniverse>) but the current iteration is hosted on the Panoptes API (<https://github.com/zooniverse/panoptes>). The methods for project design and management however remain fundamentally the same. Using custom R and Python scripts (<https://github.com/Chimp-and-See/ChimpAndSee_video_preparation_scripts>), we split the 1-minute videos into 15-second clips, and then downsize the clips such that the final size of all clips is between 900kb and 1MB using ffmpeg (https://ffmpeg.org/). If a clip is already below 1MB before downsizing, that clip is left unchanged. The 15-second clips are then uploaded to Zooniverse via the project builder interface (<https://help.zooniverse.org/getting-started>), where each site has its own collection of clips (subject set). We then use the Zooniverse workflow creation process to make site-specific classification workflows, which includes specifying the questions to be asked, the possible answers for each question, and any pictures that should be associated with the workflow. This information is detailed in.csv files that we create separately for each workflow. Each workflow is then connected to the corresponding subject set. Before the workflow becomes active, we specify the desired retirement conditions for all clips (<https://help.zooniverse.org/next-steps/caesar-realtime-data-processing/>). For example, a clip should be retired and taken out of circulation if the first three classifications are blank, or if a total of nine people classify it. At any time while the workflow is live or after all clips have been retired, data exports can be requested. The most useful exported files contain information on classifications, hashtags, comments, or the clips that were uploaded. The hashtag and comment information is exported as .json files, while the classification and clip information is exported as .csv files.

###### 

###### Data cleaning

In the case where an animal was impossible to identify in the PE data set, the PE annotated the video as “unidentifiable”. Similarly, for those videos where a duiker was present, but a more specific identification was not possible, the PE annotated the video as “duiker”.

As summarised in Figure 2a, first, we limited the PE data set to videos that were deemed to contain only zero or one species by our PE classifiers to avoid comparison issues with multispecies videos (N=264 removed). That is, our analysis workflow with PE annotations at the 1-minute level cannot differentiate between 1-minute videos with two different species in two different 15-second clips and those with two different species in the same 15-second clip. As such, we removed them from the analysis. We removed data corresponding to videos that the PE broadly classified as “duiker” (N=117). We did this as the classification options in Chimp&See are at a more specific level for duikers, and an exact comparison with the PE data set was not possible. We also removed videos that were marked “unidentifiable” (N=55) in the PE data set, as any Chimp&See classification would be a trivial disagreement. FInally we translated all PE annotations into Chimp&See classification categories (e.g.: a PE annotation of civet would equate to a CS classification of other (non-primate)). We then cleaned the CS clip classification data using an R script (<https://github.com/Chimp-and-See/ChimpandSee_Species_ID_MS_2024/blob/main/prepare_datasets_CandS_Species_ID_MS.r>) to address occasional bugs in the classification process. Starting with 52,380 clips (13,095 1-minute videos), first, we removed 156 videos as they contained clips with more than 15 classifications (and not all were blank). Further, when a clip was shown to a single community scientist on multiple occasions, we discarded all classifications except those from the first occasion. We also aggregated classifications into one line of data where a community scientist annotated a single species multiple times and summed the number of individuals across those annotations. In cases where a clip then had only one or two classifications, all of which were blank, we removed the videos from the analysis (N= 982). Clips with three or fewer non-blank classifications (classified as containing an animal of interest) are considered as “presence undetermined” and were removed from the data set. Two videos were made up of four “presence undetermined” clips and were removed from the data set. We also removed videos where community scientists may have observed multiple species as that analysis is beyond the scope of this paper (see above). To do so, we determined the number of species community scientists reported were present in each clip as the median number of species across all classifications. Videos were removed from consideration if they contained one or more clips where the median number of species was above one (n=61). This resulted in a data set of 45,821 15-second clips representing 11,894 single-species or blank 1-minute videos (Figure 2a).

###### 

###### Clip consensus

We first determined community scientist classification consensus at the 15-second clip level. A 15-second clip is considered to have a consensus on blank only when all classifications for the clip are blank. We considered clips with four or more non-blank classifications to be non-blank. Further, we considered annotations to have reached a consensus on an animal category if the proportion of classifications for that category out of the total number of non-blank classifications met or exceeded a consensus threshold (50% unless otherwise noted) and it was the only category to reach that threshold. If these criteria were not met, then the clip did not come to a consensus. For example, consider a clip with the following classifications: red duiker, dark duiker, red duiker, red duiker, warthog and the rest as blanks. This clip’s consensus classification is therefore red duiker as it represents 60% (3/5) of the non-blank classification responses. However, as a second example, if a clip contained classifications of red duiker, red duiker, dark duiker, dark duiker and the rest as blank, the clip would not come to a consensus. This is because, despite red duiker having 50% of the responses and meeting the consensus threshold, dark duiker also meets the consensus threshold with the same proportion of classifications (50%). In this latter example however, when using the umbrella designations (see below), the video would be considered as reaching a consensus on duiker (100%).

###### 

###### Aggregating clips into consensus videos

We aggregated the 15-second clip results and produced a final video-level classification from the CS data so that it would be comparable to the 1-minute long videos of the PE data set.

We assigned videos as blank only when all constituent 15-second clips were determined to be blank.

For non-blank videos, we first selected the constituent clips that reached a consensus. If there was only one consensus animal category represented among the chosen clips, the video was assigned that animal category. If multiple clips reached a consensus on different animal categories, we selected the category with the maximum consensus proportion. We refer to these videos as “consensus videos”. For example, consider a video classified as dark duiker by the PEs with four clip consensus classifications by community scientists as follows: One clip with no consensus, one clip with a consensus of leopard (consensus proportion=80%), another one with red duiker (consensus proportion=75%), and one with dark duiker (consensus proportion=75%). The highest consensus proportion among these clips is 80% and this is reached by the second clip, which reached an (incorrect) consensus on leopard. The consensus video is thus assigned an animal category classification of leopard and a consensus proportion agreement of 80%. If the video contained only clips that did not reach consensus, or if the maximum consensus proportion was shared by multiple clips that came to a consensus on different categories, then the video was designated a non-consensus video.

###### Incorporating hashtags into the classification data

For this analysis, we considered only hashtags that were correctly formatted, thus a video that contained only misspelled or unofficial hashtags was said to contain no hashtags. Videos that contained the "vocal-only" or "off\_vocalization" hashtags were omitted as these hashtags are used when an animal is heard but not necessarily seen (n=22). We also omitted videos that contained a "multispecies" hashtag (n=25) as multi-species analyses are beyond the scope of this paper.

First, the community scientist clip classifications were summarized as previously described. Then, the hashtags were incorporated depending on the clip results within each video. If the video did not have associated hashtags, the consensus species from the clips was retained. If the video clips had hashtags, a consensus hashtag was assigned to the video. In most cases only a single hashtag, or group of hashtags representing the same species at different levels of specificity (e.g. #duiker, #bay\_duiker), are present per clip. To determine the consensus hashtag at the video level, similar consensus rules applied as for classifications. The notable differences were that the hashtags were collectively considered at the video level instead of at the clip level, and there was no minimum number of hashtags needed to reach a consensus. For example, if there are three hashtags that correspond to red duiker and two that correspond to dark duiker, the video would reach a consensus on red duiker. However, if two or more species were equally supported by the hashtags, or if the most-supported category did not come from at least 50% of the hashtags, then the video did not reach a hashtag consensus.

When the video contained at least one consensus clip and there was only one consensus species represented across the consensus clips, then the hashtag data was incorporated to either accept or reject the consensus species. When the consensus classification and consensus hashtag data showed the same category, the video retained the classification consensus. If the hashtag and classification consensus data did not agree, then the video did not reach a consensus.

When all non-blank clips were non-consensus clips, then the hashtags were checked for the possibility of introducing a consensus species or category for the video based on the hashtag consensus alone.

When the video contained at least two consensus clips but did not reach a consensus originally because multiple consensus categories were supported, then the hashtags were incorporated to either accept one of the classification consensus categories or reject them all. The video was assigned a consensus category if any of the clip consensus categories matched the hashtag consensus category. If the hashtag consensus was a category that was not represented in any of the classification consensus clips for that video, then the video did not reach consensus.

Finally, if the video had been classified as blank, then the hashtags were checked for the possibility of introducing a consensus species. If the hashtags produced a consensus category, that category was then used for the video. If there were multiple species supported by the hashtags, we assigned the video as non-consensus. If a blank video had no hashtags, the video remained classified as blank.

The scripts used to process the data and determine the consensus categories at both the clip and video level as described above are available at https://github.com/Chimp-and-See/ChimpandSee\_Species\_ID\_MS\_2024.

###### 

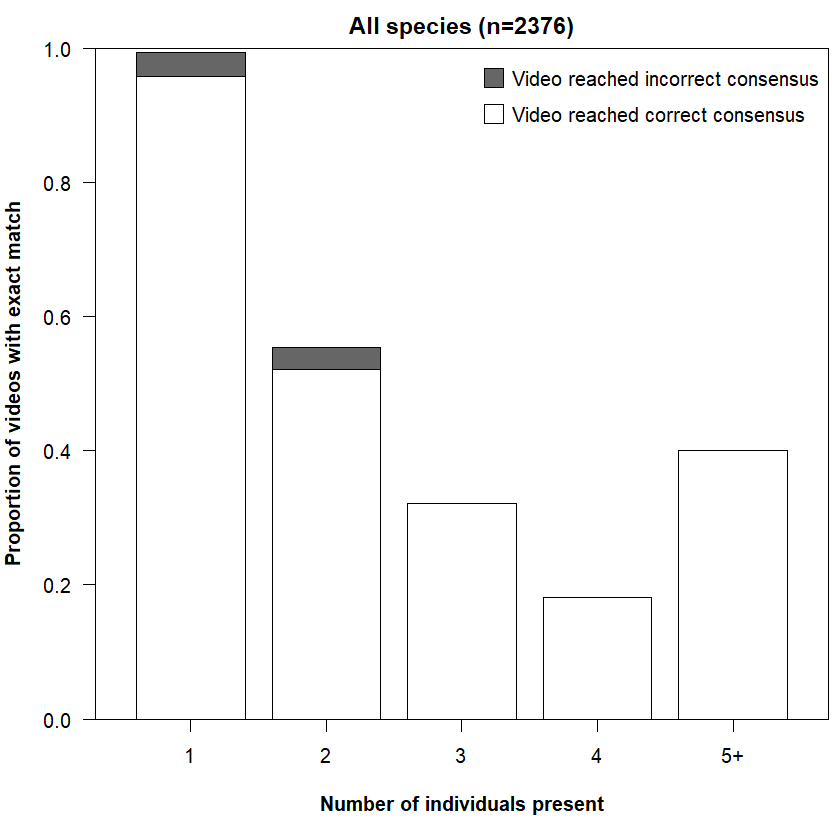
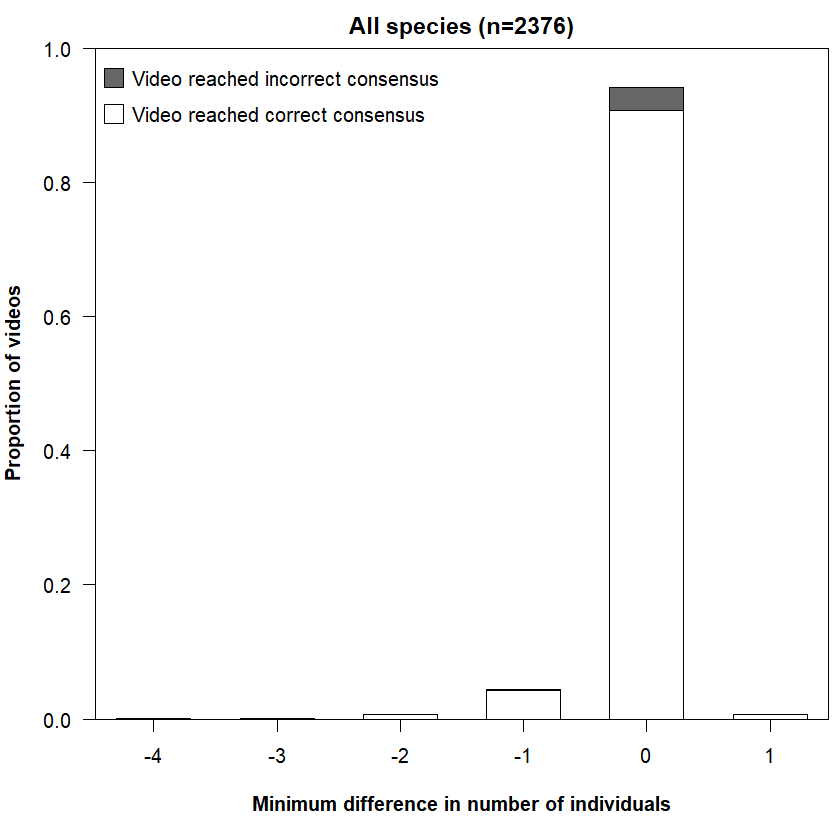
###### Analysis of counts

To analyse how well community scientists agreed with the number of individuals present in a video, we used a restricted data set, taken from the full data set described above, comprising only videos that met the following criteria: the video truly contained an animal, community scientists reported species in only one of the four clips (to avoid comparison issues with different individuals in different clips), community scientists came to a consensus on a species, the consensus was not “human” (as community scientists were not given the option to choose a number of humans present), and there were four or more community scientists who agreed with the ultimate consensus category. To assign a final count for each video, we determined a consensus number in the same manner as the consensus category classification, using only the count data from community scientists who agreed with the clip's consensus category (N=2,395). Count classifications can be 1, 2, 3, 4 or 5+. For these analyses all classifications of 5+ were standardised to 5.

##### 

##### Supplemental Figures

###### Supplemental Figure 1

a) b)

**Supplemental Figure 1. Difference in number of animals counted by the community scientists and the professional ecologist (PE).** White indicates the consensus classification category matched that of the annotator, grey indicates the wrong consensus animal was found by community scientists, even if the correct number of individuals was identified. a) Distribution of the difference in counts between the consensus count and the PE count (i.e. consensus count minus PE count) while using the original classification categories. Positive numbers indicate the community scientists classified a larger number of individuals than the PE. All classifications of 5+ were changed to 5 and therefore represents the minimum difference between the community scientists and the PE. b) Proportion of videos containing the indicated number of individuals that reached a consensus on that number exactly (i.e. difference of 0 between consensus count and PE count).