The cost of coordination can exceed the benefit of collaboration in performing complex tasks

Vincent J Straub
Oxford Internet Institute, University of Oxford, Oxford, UK

Milena Tsvetkova
Oxford Internet Institute, University of Oxford, Oxford, UK

Department of Methodology, London School of Economics and Political Science, London, UK

Taha Yasseri (ORCID)
Oxford Internet Institute, University of Oxford, Oxford, UK

School of Sociology, University College Dublin, Dublin, Ireland

Geary Institute for Public Policy, University College Dublin, Dublin, Ireland

Alan Turing Institute for Data Science and AI, London, UK

Abstract
Humans and other intelligent agents often rely on collective decision making based on an intuition that groups outperform individuals. However, at present, we lack a complete theoretical understanding of when groups perform better. Here, we examine performance in collective decision making in the context of a real-world citizen science task environment in which individuals with manipulated differences in task-relevant training collaborated. We find 1) dyads gradually improve in performance but do not experience a collective benefit compared to individuals in most situations; 2) the cost of co-ordination to efficiency and speed that results when switching to a dyadic context after training individually is consistently larger than the leverage of having a partner, even if they are expertly trained in that task; and 3) on the most complex tasks having an additional expert in the dyad who is adequately trained improves accuracy. These findings highlight that the extent of training received by an individual, the complexity of the task at hand, and the desired performance indicator are all critical factors that need to be accounted for when weighing up the benefits of collective decision making.

Keywords
collective intelligence, competence, team, decision making, citizen science

Corresponding author:
Taha Yasseri, School of Sociology, University College Dublin, Dublin D04 V1W8, Ireland.
Email: taha.yasseri@ucd.ie

Significance Statement

Collaboration is one of the fundamental processes in humans' social lives. With the invention of digital communication technologies designed to further facilitate collaboration, identifying the keys to successful collaborations is more desirable than ever. Previous work in this area consists primarily of stylized tasks and therefore suffers from a lack of generalizability. In this work, we utilize a real-world citizen science platform to run experiments with subjects recruited from a diverse pool of non-expert participants. We show that when the task is complex, making decisions as an individual can be better than joint decision making in dyads, particularly where coordination requires time and effort. Our work can inform the design of collaboration platforms and advance the science of teamwork.

Introduction

Intelligent agents, be they natural or artificial, constantly have to make decisions to solve complex problems. Humans are no exception; decision making is omnipresent in socioeconomic life, occurring in households, classrooms, and, increasingly, online (Tsvetkova et al., 2016). The common belief is that group decisions are superior to decisions made by individuals. The proverb "two heads are better than one" captures the intuition that two (or more) people working together are indeed more likely to come to a better decision than they would if working alone (Bahrami et al., 2010; Koriat, 2012). To test this notion empirically, we experimentally study collective image classification tasks using the real-world task environment of a citizen science project, where participants classified pictures of wildlife, taken as part of a conservation effort, as accurately and quickly as possible.

Our study focuses on dyadic interaction and was designed to reveal to what extent task complexity and variation in learned expertise influence the accuracy of classifications. We ask three core questions: whether and when dyads perform better than an average individual, whether dyads with similar training perform better than dyads with different training, and how performance is mediated by the complexity of the task.

Surprisingly, we find little support for two heads being more accurate than one except for the most complex tasks, for which having an additional expert significantly improves performance upon that of non-experts. Our results show that pairs of individuals gradually improve in performance as they work together but tend not to experience a collective benefit compared to individuals working alone; rather, the cost of coordination to efficiency and speed is consistently larger than the leverage of having a partner, even if they are expertly trained. These findings highlight that the extent of training received by an individual, the complexity of the task at hand, and the desired performance outcome are all critical factors that need to be accounted for when weighing up the benefits of collective decision making.

Our findings stand in stark contrast to a vast literature of research on decision making showing that groups usually make better decisions than individuals; most of that research considers groups with more than two members. Beginning with the discovery of judgment feats achieved by large numbers of people, classically in point estimation tasks like guessing the weight of an animal (Galton, 1907), the idea of the "wisdom of the crowd" (Surowiecki, 2004) has become a prominent example. Building on other cases from stock markets, political elections, and quiz shows, evidence from other guessing tasks and problem-solving experiments (Kerr and Tindale, 2004; Grasso and Convertino, 2012) shows that the aggregate of many people's estimates often tends to be closer to the true value than all of the separate individuals.

Recent examples of research on the wisdom of the crowd include answering general knowledge questions (Navajas et al., 2018) and estimating political events (Becker et al., 2019). Yet, while some have found the effect holds for higher dimensional tasks involving spatial reasoning and combinatorial problems such as the traveling salesman task (Yi et al., 2012), in general, most are different to our study as they nevertheless continue to rely on numeric judgment tasks, such as dot estimation (Almaatouq et al., 2020). This difference in task context means most studies have focused simply on understanding the collective decision making through the pooling of personal information, often via simple averaging.

Importantly, as studies have turned their attention to understanding what determines the performance advantage of collective decision making, they have reached divergent conclusions on the importance of (i) diversity, in terms of individual team member attributes (Hong and Page, 2004; De Oliveira and Nisbett, 2018); (ii) group size and structure (Galesic et al., 2018; Navajas et al., 2018); (iii) incentive schemes (Mann and Helbing, 2017); (iv) the nature of the task, also referred to as the "task context" (used henceforth) or "task environment" (Mata et al., 2012); and, perhaps most contentiously, (v) the role of social influence, that is, whether interaction undermines (Lorenz et al., 2011; Muchnik et al., 2013) or improves (Becker et al., 2019; Farrell, 2011) collective performance.

With respect to pairs of individuals (dyads)—the most elementary social unit (Kozlowski and Bell, 2013) and focus of this paper—related recent experimental studies have tended to examine the role played by (1) individual confidence (Koriat, 2012; Bang et al., 2014) and (2) argument quality (Trouche et al., 2014) in determining the

success of collective decision making. When taking others' opinions into account on various numerical estimation tasks, studies have shown that people are known to rely on a "confidence heuristic" (Yaniv and Milyavsky, 2007) such that more confident opinions tend to weigh more. For instance, Koriat (2015) has shown that in the case of perceptual or even general knowledge questions, the outcome of group discussion can be emulated by aggregating the individual judgments of the group members weighed by their confidence. Meanwhile, studies of intellective tasks, problems, or decisions for which there exists a demonstrably correct answer within a verbal or mathematical conceptual system, have focused on the idea that arguments, more than confidence, explain the good performance of reasoning groups (Trouche et al., 2014). Blanchard et al. (2020) report that decision outcomes improved when people act in dyads compared with acting individually to complete a general knowledge test.

Recent studies have reported on other challenges in dyadic collaboration. The dyad members can suffer from an egocentric preference for personal information and fail to reap collective benefits by not listening to their expert counterparts (Yaniv and Milyavsky, 2007; Tump et al., 2018). Similarly, pairs may underperform due to people's tendency to give equal weight to others' opinions (Mahmoodi et al., 2015). Finally, individuals experience a collective benefit only when dyad members have similar levels of expertise and training, as this ensures members accurately convey the strength of their belief and the dyad can reliably choose the best strategy (Bang et al., 2014, 2017).

Based on the observation that such critical agent characteristics, particularly the relevance of prior personal knowledge (i.e., expertise), are not randomly distributed, our work builds on this line of work by directly operationalizing differences in the extent of task-relevant training between interacting individuals—in contrast to studies which have relied on proxies for fixed differences in decision experience and ability (Sella et al., 2018).

Crucially, most prior studies typically provide participants with immediate feedback, alongside treating collective benefit as a static event. However, this fails to take into account that actual human pairs often have to perform tasks under uncertainty, whilst simultaneously benefiting from continued individual and social learning. The present study narrows the gap between experimental control and realistic settings by examining collective decision making in the context of an established citizen science task—something that has not hitherto been tried for large-scale image classification platforms.

Study design

*The WildCam Gorongosa task.* In the experiment, participants had a set amount of time to classify pictures taken by motion-detecting camera traps in Gorongosa National Park in Mozambique as part of a wildlife conservation effort (the trail cameras are designed to automatically take a photo when an animal moves in front of them). Specifically, we instructed participants, physically present in the lab and seated in front of a computer monitor, to use the WildCam Gorongosa website, hosted by Zooniverse, the world's largest online platform for citizen science (Cox et al., 2015). To increase the generality of the findings, the exact number and sequence of images are not controlled at the individual level, as is the case in many prior studies that have employed abstract, stylized tasks. Rather, to reflect the real-world nature of the tasks used in the experiment, we ensured participants classified images in a manner consistent with the experience of volunteering citizens visiting the WildCam Gorongosa site.

The WildCam Gorongosa task consists of five distinct labeling subtasks. For each image, participants in the experiment had to perform the following classifications, listed in increasing complexity: (1) detecting the presence of the animal(s), (2) counting how many animals there are, (3) identifying the behaviors exhibited, specifically, identifying whether the animal(s) is (a) standing, (b) resting, (c) moving, (d) eating, or (e) interacting (multiple behaviors may be selected), (4) recognizing whether any young are present, and (5) identifying the species type. The 52 possible species options include a "Nothing here" button but no "I don't know" option (see Figure 1(a) for example images and Supplementary Material Figures S1–S4 for further examples and screenshots of the online platform and instructions).

Prior work has demonstrated that object recognition is a task that requires context-dependent knowledge and various facets of our visual intelligence (DiCarlo et al., 2012); hence, citizen scientists primarily carry out tasks for which human-based analysis often still exceeds that of machine intelligence (Trouille et al., 2019). In the present case, the WildCam Gorongosa task is thus considered suitable for the purposes of analysis for the following two reasons. First, it has high ecological validity: as part of an established crowdsourcing platform, a type of "human-machine network" characteristic of our hyper-connected era (Tsvetkova et al., 2017), the site has engaged over 40,000 volunteers to date (Wildcam Gorongosa, 2020). Given the abundance of situations in everyday life where immediate outcomes are difficult, sometimes even impossible, to establish, the task can also be generalized to other contexts in that no feedback was provided to participants. Second, identifying various features in camera trap wildlife images is sufficiently difficult (Norouzzadeh et al., 2018) that it offers the possibility of collaborative benefits, especially when dyad members have received similar training for identifying particular features.

We operationalize task complexity experimentally by varying the number of information cues that subjects consider (De Cesarei et al., 2017; Groen et al., 2018). In

A   Examples of Wildcam Gorongosa classification tasks and difference in training conditions
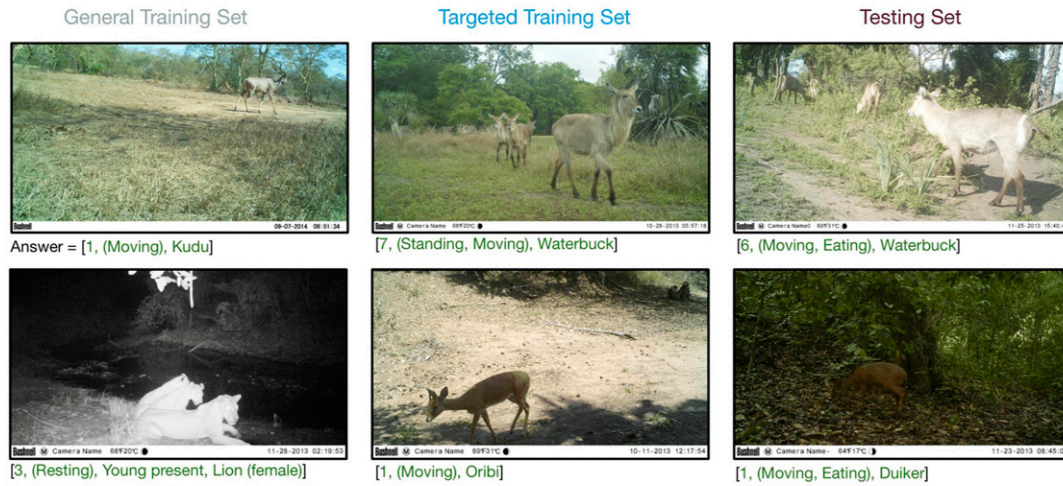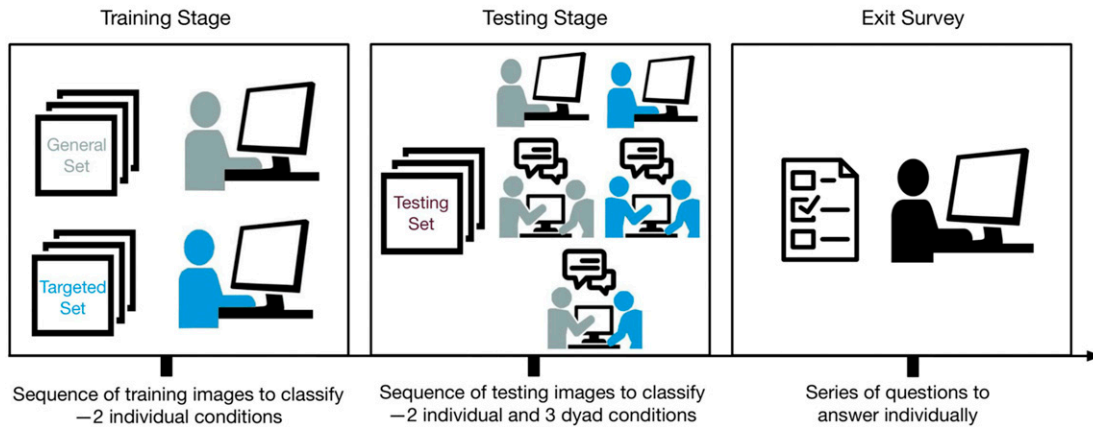

Figure 1. Experimental design. (A) Examples of the experimental task and illustration of the difference between the set of images seen in the general versus the targeted training condition. Instances of images classified in the testing stage are also provided. (B) Sequential schema of events in the Experiment. In the training stage T_1, every participant classified images individually. In the testing stage T_2, participants were additionally assigned to either an individual or dyad condition. WildCam Gorongosa imagery is reprinted under a CC BY 4.0 license with permission.

particular, as our definition means the complexity of each subtask is based upon the number of visual information cues that must be processed in order to succeed in the task, identifying the species is considered the most complex because it is expected to require processing the most inputs (Swanson et al., 2016; Parrish et al., 2018).

Whilst this conceptualization diverges from classical organizational behavior definitions (Wood, 1986; Hackman, 1969), which also consider the actor doing the task, the task context, and the behavioral pattern required to perform the task, it has parallels with the notion of "component complexity," part of the original framework proposed by Wood (1986). We thus nevertheless retain the original label "task complexity" because our definition retains a connection to the original concept. Differences in learned expertise are

meanwhile operationalized as differences in task-relevant training, which is taken to be the main way to facilitate the acquisition, enhance retention, and promote the transfer of relevant knowledge to contexts not encountered during training (Healy et al., 2014).

Experimental procedure. The experiment was divided into a training and a testing stage ($T_1 = 30$ min and $T_2 = 45$ min). A training stage was required in order to ensure select participants could receive task-relevant training and build up greater expert knowledge in performing the task at hand; this in turn ensured mixed-skill dyads could be formed for the testing stage where one member (the "expert") has higher learned expertise. The length of the training stage was determined by analyzing pre-existing citizen scientist-
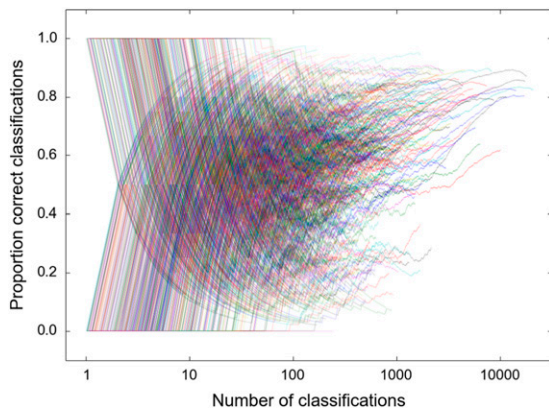
Figure 2. Individual learning trajectories of citizen scientists. Trajectories are defined as the change in the proportion of correct classifications averaged across each of the five WildCam Gorongosa tasks relative to the number of classifications made. Each learning trajectory is for an individual volunteer citizen scientist who accessed the WildCam Gorongosa site prior to the time of the experiment. Note the logarithmic scale on the x-axis; hence for a small number of classifications, the curves take exotic shapes.

generated data for the WildCam Gorongosa task. In particular, we analyzed the "learning trajectories" of users, and changes in performance over time, to determine the average saturation time point, defined as the point after which the majority of users no longer improve. As depicted in Figure 2, most users who classified the first image correctly (incorrectly) made an incorrect (correct) classification within 10 classifications; only a small subset maintained their winning (losing) streak. After determining the average number of classifications an individual made before no longer experiencing an increase in the proportion of classifications classified correctly and multiplying this count by the average time it took to make a single classification, this was found to be 30 minutes.

For the T_1 period, participants were randomly assigned to two different training conditions, "General" and "Targeted," and asked to individually classify a sequence of approximately 50 images, where the content of the images was varied between conditions. As T_1 was designed to provide selective training to participants in the Targeted treatment condition, the set of images seen by these participants (Targeted set) consisted of pictures sharing specific features, sampled from a predetermined subset of all the images available to classify on WildCam Gorongosa. Specifically, the Targeted set was restricted to pictures containing antelope species: bushbuck, duiker, impala, kudu, nyala, oribi, reedbuck, and waterbuck. These animals were chosen as they look visually similar, share a number of morphological features, and exhibit similar behaviors, thus making them relatively harder to distinguish (Norouzzadeh et al., 2018). For the General condition, pictures shared less specific

features and were instead sampled from a much broader predetermined subset containing a more diverse set of animals other than the ones mentioned above which are easier to distinguish, including baboons, warthogs, and lions, among many others (see Supplementary Text S2.3.1 for details). The analysis presented below confirms the effectiveness of this treatment in producing significant differences between different groups.

T_2 was designed to assess the effect of differences in task-relevant training as well as the effect of working alone versus collectively. For this reason, participants were further randomly assigned to either a "Solo" or "Dyad" condition, with participants in the dyad condition having to make a joint decision by reaching a consensus (as opposed to voting or relying on an averaging procedure); the dyad members worked together, and one of them was randomly assigned to input decisions on behalf of the pair. When taking into account the level of training, the experimental procedure resulted in 2 individual testing conditions, "General Solo" and "Targeted Solo," and 3 dyad testing conditions, "General Dyad," "Targeted Dyad," and "Mixed Dyad." The set of images seen by all participants, regardless of testing condition, thus consisted of pictures sampled from the Targeted set (see Supplementary Text S2.3.2 for details). Figure 1 illustrates the overall experimental design.

Each decision process that we measure can be broken down into two phases: (1) individual or (in the case of dyads) joint deliberation, deciding what decision to make, and (2) execution, confirming and inputting the decision using the computer mouse. Because we consider collective decision making to be primarily concerned with the process outcomes of interaction, we are interested in the former phase: how effectively dyads deliberate (i.e., share social information). In the execution phase, dyads may be expected to be, at least initially, slower in inputting their decision as they have to coordinate their classification. However, we consider this to have a negligible impact on their overall performance and expect it to diminish over time as they learn to coordinate. Instead, we expect their performance to be determined primarily by how effectively they interact during deliberation, so we do not try to arbitrarily discount the effect of the execution phase on performance via any type of weighting.

To measure performance, three metrics were in turn computed for every individual or dyad i: *pace*, defined as the number of images classified by i per minute (referred to as *volume* when considering the absolute number); *accuracy*, defined as the number of correct guesses made by i as a proportion of volume; and *efficiency*, defined as the number of correct guesses made by i per minute. These metrics are chosen as they have been widely used as measures of performance in decision science (Maloney, 2002). Except for pace, which is an aggregate measure—and thus the same across tasks—accuracy and efficiency were computed
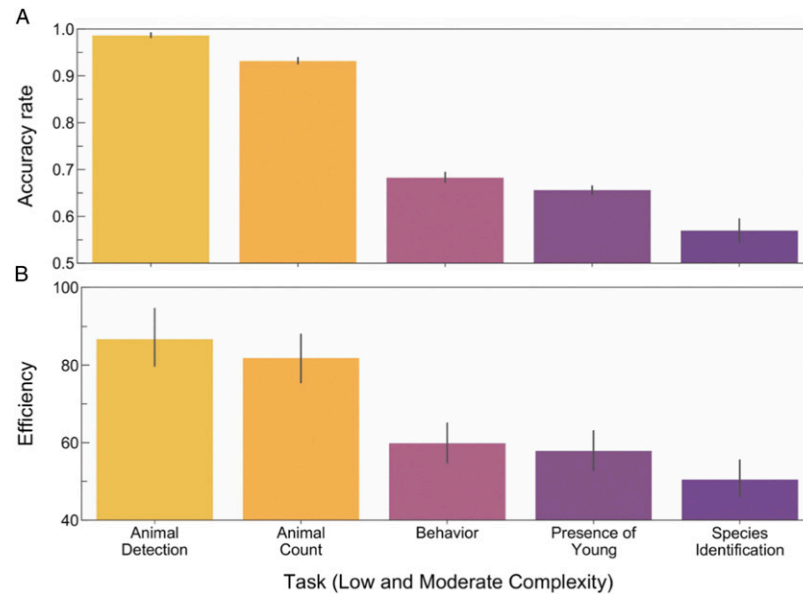
Figure 3. Task complexity mediates performance. Data are combined across all solo and dyad grouping conditions for the testing stage. The more "complex" the task, the greater the reduction in average accuracy (**A**) and average efficiency (**B**). The difference in experienced difficulty between the task with the lowest and highest complexity is very large: the average accuracy score dropped by nearly 50%. Error bars indicate the 95% confidence intervals.

separately for the five tasks (see Supplementary Text S3.2 for details).

Results

Figure 3 shows the overall accuracy and efficiency of participants for different tasks. Tasks vary in their complexity, and therefore in the efficiency and accuracy of classifications. Based on these results we categorize detection of an animal and animal count as *low complexity* tasks and identifying the behavior, recognizing young animals, and identifying the species as *moderate complexity* tasks.

The effect of training on classifications

Focusing first on the effect of training on performance across the different tasks of varying complexity, we assess the difference in performance changes over the course of the entire experiment where change is relative to an individual's initial performance (see Materials and Methods and Details of Analysis in the Supplementary Information for details). T_1 and T_2 were separately split into three equally spaced and non-overlapping time intervals. The average change in performance was then estimated separately for each interval by computing the difference in performance compared to the performance of the respective individual(s) in the first interval of the training stage, which can be thought of as the natural baseline prior to training. The rationale is that

benchmarking against this initial performance allows for a more precise estimation of whether individual training, interaction (in the case of dyads), or both provide performance gains (Almaatouq et al., 2021).

Figure 4 depicts the results, showing the average change in efficiency across each of the five WildCam Gorongosa tasks during the training (T_1) and testing stages (T_2) for participants in the General Solo (Figure 4(a)) and Targeted Solo (Figure 4(b)) groups. Relative differences in learning rates and performance changes indicate that individual performance is mediated both by the level of training received and the complexity of the task (Saffell and Matthews, 2003; Mao et al., 2016; Mata et al., 2012).

Individuals in the General Solo group continuously improve across each task during T_1, becoming most efficient during the last third of training, as the benefits of general training gradually subside, before experiencing a sharp decline during the first phase of testing. Although they are able to recover by the end of T_2 with respect to animal detection, count, and behavior—tasks of lower complexity—the consistency of the decline across tasks indicates that individuals with general training are initially less efficient when confronted with new task stimuli. In contrast, individuals in the Targeted Solo group continuously improve upon their T_1 performance throughout the course of T_2, experiencing greater relative gains in efficiency than General Solo for each time interval (Figure 4(b)), suggesting that targeted training provides consistent accumulative performance benefits.
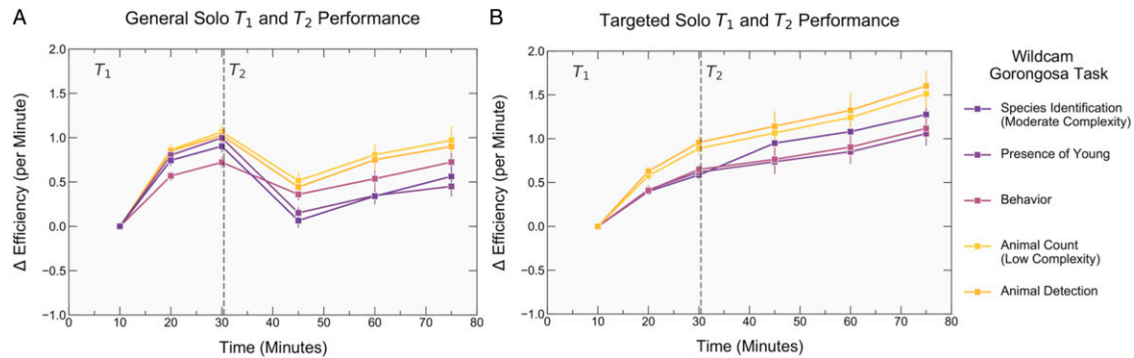
Figure 4. Individual training effectiveness. Performance changes in terms of the average change in efficiency across each of the five WildCam Gorongosa tasks during the training (T_1) and testing stage (T_2) for General Solo (**A**), individuals who received general training during T_1, and Targeted Solo (**B**), individuals who received selective training. The dashed vertical line falling in T_2 separates both stages, which each have 3 data points. Error bars indicate one standard error of the mean.

Individual versus collective classifications

We now turn to our primary research questions: whether dyads outperform individuals and how that the performance of dyads depends on task complexity and individual ability. We analyze differences in performance change during the final phase of the testing stage, when the benefits of interaction for dyads can best be separated from additional effects, such as initial improvements in coordination. As a secondary analysis, we also examine differences over the course of the entire testing stage; this allows us to see how these changes evolve, which in turn allows us to detect when and why coordination problems may arise, for instance, as a result of an initial conceptual shift.

The performance of Mixed Dyads is computed in two ways: using a "General benchmark" (the performance set by the member with general training) when comparing Mixed Dyads to General Solo and General Dyads, and using a "Targeted benchmark" (the performance set by the member with targeted training) when comparing Mixed Dyads to Targeted Solo and Targeted Dyads (see Material and Methods).

Figures 5(a)–(b) depicts the results for pace during the last time interval of T_2 for the species identification task (most complex task), indicating that participants in an individual condition were on average faster in classifying images belonging to the test set—regardless of whether or not they received targeted training. Moreover, only Targeted Dyads and, to a lesser extent, Mixed Dyads still improved in terms of pace during the final phase; however, both General Dyads and Mixed Dyads benchmarked against individuals receiving general training saw no or close to no further improvements on average. Statistical tests of the differences provide further evidence that participants in the solo conditions outperform participants in the respective dyad

conditions ($p < 0.05$ for pairwise comparisons; see Supplementary Material, Sec. S3.3).

Moving on to task-specific performance differences in accuracy and efficiency for the species identification task, considered the most complex task, General Dyads experience no interaction benefit. Yet, pairing an individual with general training and an individual with targeted training strongly counteracts this result: Mixed Dyads improve in accuracy by 37% more than General Solo ($p = 0.036$, Cohen's $d_s = 0.616$, 95% CI for Cohen's d_s: [0.034, 1.199]) and 55% more than General Dyads ($p = 0.018$, Cohen's $d_s = 0.895$, 95% CI for Cohen's d: [0.298, 1.492])—indicating that participants with general training can improve upon their expected individual performance by collaborating with a selectively trained partner who can supply task-specific expertise.

Targeted Solo individuals meanwhile experienced a significantly greater average improvement in accuracy during the last interval of T_2 (41%; Figure 5(d)) compared to Targeted Dyads (24%; $p = 0.026$), and to Mixed Dyads (29%; $p = 0.134$), although to a less evident extent. As a result, they appear to consistently achieve the greatest efficiency (Figure 5(f)), albeit only slightly. In contrast, whilst General Solo achieved greater efficiency than General Dyads as a result of increased speed (Figure 5(a)), achieving 0.78 more classifications per minute ($p = 0.002$, Cohen's $d_s = 0.934$, 95% CI for Cohen's d_s: [0.322, 1.546]), the slight difference between General Solo and Mixed Dyads (Figure 5(e); $p = 0.557$) implies that individuals without selective training experience a speed advantage when working alone but derive an accuracy benefit when collaborating with a selectively trained teammate.

Surprisingly, across all other tasks, there was no clear indication of substantive differences between individuals and dyads in terms of accuracy during the final phase of testing (see Supplementary Material).
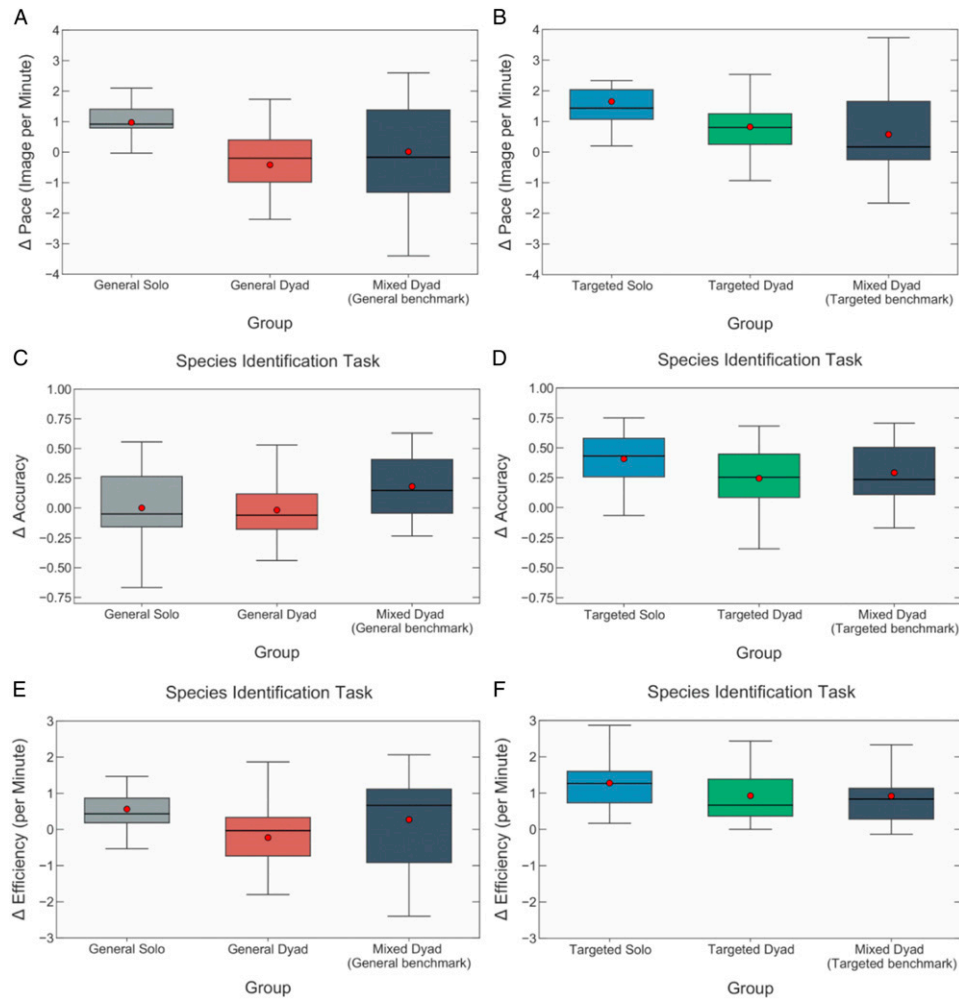
Figure 5. Individual and collective performance. Performance differences between individuals and dyads for the last third of T_2 in terms of the distribution and average change in pace (**A–B**), alongside accuracy (**C–D**) and efficiency (**E–F**) for the species identification task, considered the most complex. General Solo individuals are compared to General Dyads and Mixed Dyads (left panels), and Targeted Solo individuals are compared to Targeted Dyads and Mixed Dyads (right panels). Error bars indicate one standard error of the mean. Whiskers are 1.5 times the interquartile range. The solid black line inside each box indicates the median and the red circles do the mean.

Classifications over time

Figures 6 and 7 show how accuracy changes over time for different tasks (grouped into low and moderate complexity tasks, respectively).

When comparing the change in accuracy over the course of the entire testing stage, there is no clear indication of significant or consistent differences between individual and dyads in terms of accuracy for less complex tasks; Figure 6 illustrates that this holds across the entire testing stage. When comparing changes in accuracy between General Solo and the respective dyad conditions for the presence of young task, considered to be of moderate complexity (Figure 7), a similar pattern can nevertheless be observed to changes in accuracy across the species identification task;

notably, individuals and dyads both fail to experience any accuracy improvements. Moreover, when comparing Targeted Solo and the respective dyads conditions, Mixed Dyads appear to experience the greatest improvements in accuracy during the first two intervals of T_2. However, the slight variations suggest that tasks of low complexity are most probably too simplistic for practically relevant performance differences to emerge.

A decline is to be expected given the coordination issue, the extent of the decrease in performance experienced by dyads depends on the distribution of expertise among dyad members and has a different effect depending on this distribution. As seen in Figure 8, whilst General Dyads and Mixed Dyads unsurprisingly undergo a decline in accuracy, pace, and consequently efficiency, Targeted Dyads and
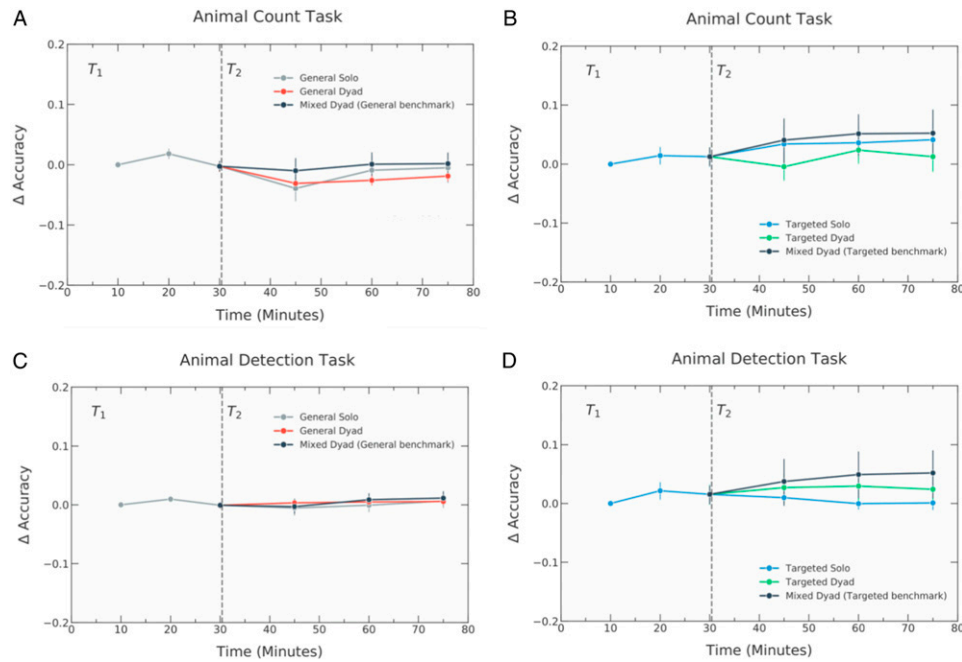
Figure 6. Change in accuracy over time. Performance differences between individuals and dyads over the course of T_2 in terms of the average change in accuracy for low complexity tasks. Error bars indicate one standard error of the mean.

Mixed Dyads only experience a fall in pace but not accuracy; hence, the slight dip in efficiency during the first time interval of T_2 (Figure 8(d)) is primarily the result of classifying images more slowly. More specifically, individuals and dyads with general training respectively experienced a reduction of 12% and 13% in accuracy during the first interval of T_2.

Although both groups on average still gradually improved over time, they do not manage to recover to the performance benchmark set by General Solo during the first phase of training. As a consequence, the finding that General Dyads experience no interaction benefit at all remains consistent beyond the initial phase of testing. The observation that pairing an individual with general training and an individual with targeted training causes this decline is meanwhile consistently strong but becomes most pronounced in the final time interval. This not only implies that participants with general training can improve upon their expected individual performance by collaborating with a selectively trained partner but also further suggests that the greatest interaction benefits among dyads with mixed expertise occur after overcoming initial coordination issues.

Regarding efficiency, Targeted Solo individuals appear to consistently achieve the greatest performance gains, as compared to Targeted Dyads and Mixed Dyads, albeit only slightly. In contrast, whilst General Solo achieved greater efficiency than General Dyads as a result of increased speed reflected in pace ($p = 0.018$), the slight difference between General Solo and Mixed Dyads ($p = 0.557$) implies that

individuals without selective training experience a speed advantage when working alone but derive an accuracy benefit when collaborating with a selectively trained teammate (see also Supplementary Material Fig. S7, which shows that using different time intervals does not significantly change any of the reported findings).

Discussion

Our findings, in contrast to empirical studies of larger groups, demonstrate that the benefits of dyads are contingent. On the one hand, the stepwise performance improvements experienced by dyads in terms of pace and accuracy for the species identification task are consistent with prior studies showing that collective performance increases gradually for complex tasks in the absence of feedback (Bahrami et al., 2012; Mao et al., 2016). On the other, these findings imply that switching to a dyadic context for a task after training in an individual context in most cases creates a coordination problem that hurts performance.

Although the observation that dyads do not consistently experience a collective benefit across tasks goes against prior results that indicate an interaction benefit in the context of low-level visual enumeration and numeric estimation tasks (Koriat, 2015; Wahn et al., 2018), this finding is nevertheless in line with the literature that has found no pair advantage when the task context consists of general knowledge-based tests involving discrete choice decisions
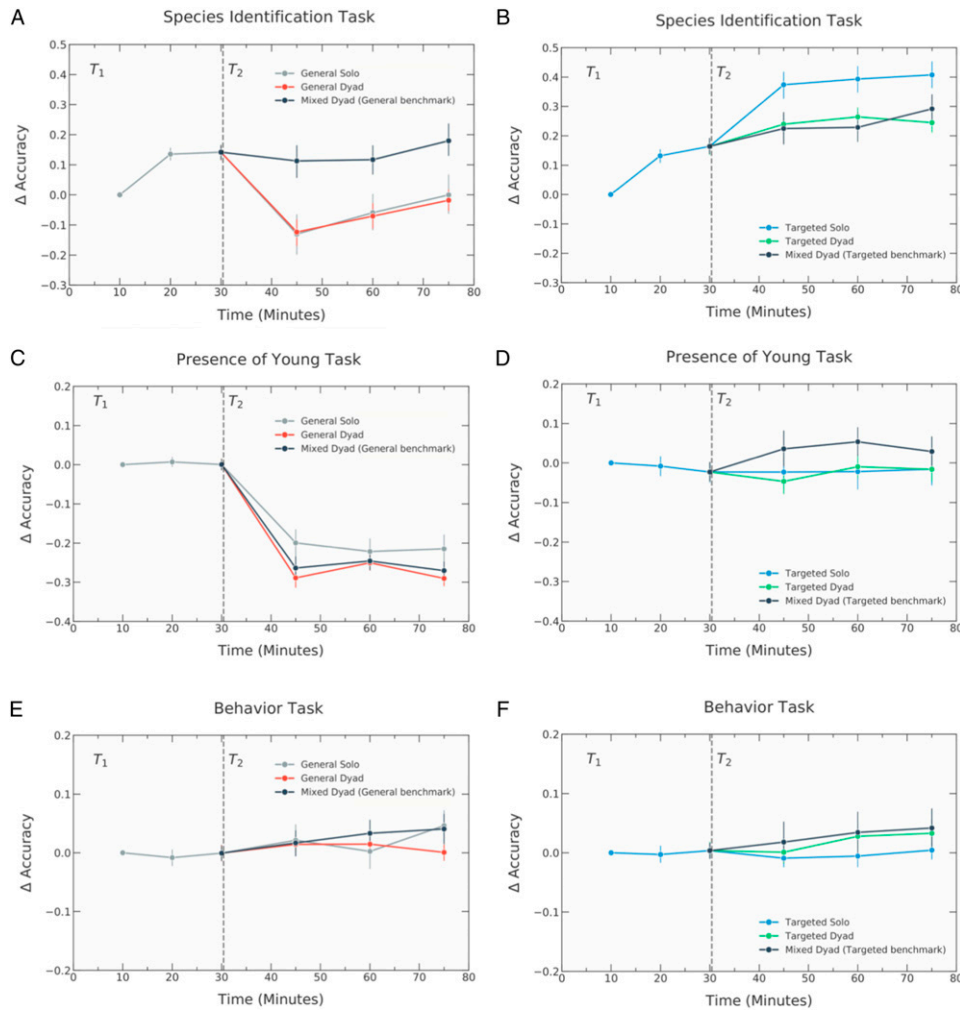
Figure 7. Change in accuracy over time. Performance differences between individuals and dyads over the course of T_2 in terms of the average change in accuracy for moderate complexity tasks. Error bars indicate one standard error of the mean.

(Schuldt et al., 2017; Blanchard et al., 2020). This underscores the extent to which performance greatly depends on overcoming any initial coordination issues and the task context. Whilst some previous studies have suggested that individuals outperform dyads due to faster skill acquisition (Crook and Beier, 2010), the greater accuracy gains achieved by Mixed Dyads when compared to General Dyads and General Solo for the species identification task show that, at least for complex tasks, this may further depend on the level of task-relevant training among dyad members. Additionally, our results highlight that task-relevant training, particularly targeted training, provides significant performance improvements, regardless of working individually or in a pair.

Our work stresses that existing experiments and theories of collective decision making, such as the "confidence matching" model stating that only pairs similar in skill or social sensitivity experience a collective benefit (Bang et al.,

2017), do not adequately stipulate if, and under what conditions, dyads will outperform individuals.

The immediate contribution of this study is to demonstrate that in the context of a real-world task context, namely, Zooniverse's WildCam Gorongosa citizen science project, the heads of two experts, understood here as individuals with targeted training, are not always better than one. Rather, one expert is more *efficient* than two experts or a mixed ability dyad, indicating that the cost of coordination to efficiency is consistently larger than the leverage of having a partner—even if that partner is also specially trained for the task at hand.

Moreover, we also show that a non-expert, an individual with basic training, works faster than a dyad, even if one of the dyad members is an expert, thereby giving further support to the theory that dyads face a coordination problem which costs time and suggesting that individuals exert less effort when working in a pair. However, when it comes to
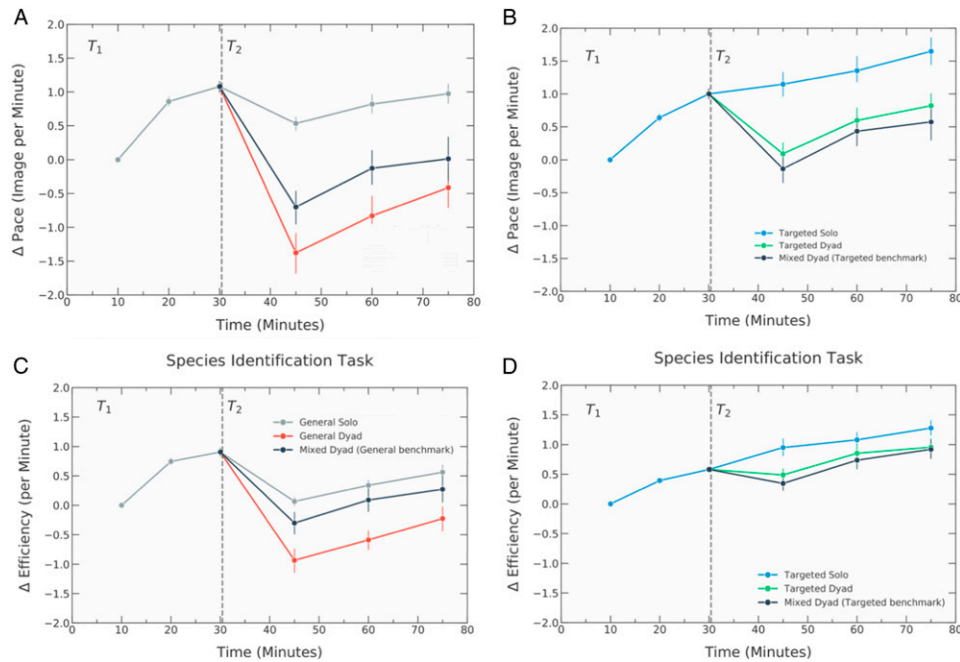
Figure 8. Change in pace over time. Differences between individuals and dyads over the course of T_2 in terms of the pace of task completion. Error bars indicate one standard error of the mean.

accuracy in the most complex task, having one expert in a dyad significantly improves performance compared to that of individual non-experts or dyads of non-experts; importantly, two expertly trained individuals working together do not appear to be more accurate than one. This suggests interaction, whereby dyads are allowed to exchange personal information, may inherently give rise to as-of-yet unexplored psychological biases that hurt performance, even when there is equality in the distribution of knowledge.

To explain our findings, we theorize the following scenarios:

(a) Process losses due to social dynamics: Our findings are in line with the literature that has found no pair advantage when the task context consists of general knowledge-based tests involving discrete choice decisions. Theories of social dynamics that could help explain why this is the case, relate to the "process losses" that can plague group settings (e.g., groupthink). One possible explanation for the differential gains in confidence observed across dyad types is with regard to the ability of group settings to reduce individual feelings of uncertainty. Social comparison theory (Festinger, 1954) posits that individuals are motivated to assess the validity of their opinions by comparing them to those held by others, in the absence of other non-social, "physical" means for doing so. Social comparison processes are thus more likely to operate in groups

performing tasks that are more judgmental, as compared with intellective, in nature (Laughlin and Earley, 1982), as when individuals in a jury scenario adopt a higher threshold for finding a suspect guilty ("beyond a reasonable doubt") after participating in a group discussion (Magnussen et al., 2014; Schuldt et al., 2017).

(b) Overconfidence/equality bias: Overconfident (but inaccurate) people convince less confident (but accurate) people to change their opinions toward the wrong decision; this idea is regularly invoked in the field of collective decision making and is advanced, for instance, by Valeriani et al. (2017). Blanchard et al. (2020), for instance, argue that overconfidence may be responsible for the failure to improve group decisions, as they find that, on average, dyads became more confident than individuals despite no increase in accuracy. The reasons they give to explain this relate to the metacognitive constructs (i.e., trait-confidence and bias), that is, the presence of one cautious individual (i.e., lower trait-confidence and underconfident) may cause unjustified increases in confidence displayed by dyads composed of overconfident and potentially higher trait-confidence individuals. In a different but related vein, Mahmoodi et al. (2015) argue that the reason individuals in a dyad weigh each other's opinions equally, even if one individual is unjustly confident, is because there is an "equality bias"

across cultures, meaning when making decisions together we tend to give everyone an equal chance to voice their opinion. In our experiment, this may be especially useful for explaining why mixed dyads performed poorly.

(c) Insufficient time for discussion/social learning theory: In our experiments, time was also a practical constraint. The time pressure of completing the task in a 45-minute window meant dyads failed to properly discuss and instead the more confident/dominant individual was able to push their opinion through more easily. The more theoretical reason that we could use to explain why time matters, relates to social learning theory (Toyokawa et al., 2019)—"learning that is facilitated by observation of, or interaction with, another individual or its products, is known as 'social learning'" (Kendal et al., 2018). Despite the positive connotation of the term learning, another face of social learning is herding effects where the collective decision gradually moves toward a less optimal solution (Vives, 1996).

In a situation where a general manager is deciding whether or not to rely on a qualified expert or a pair of untrained team members for solving various tasks effectively, the results of the present study would suggest that, at least for complex classification tasks, overall performance can be maximized by relying on a pool of expertly trained individuals working alone. Similarly, if the speed at which the decision needs to be made is key, it is also best to rely on individuals working alone. Yet, if accuracy carries more practical importance than productivity, and it is not possible to provide specialized training, then recruiting experts and building mixed-ability teams may be most effective.

Providing specialized individual training to all or at least some workers and relying on group work only when accuracy matters may be the most effective strategy, whilst relying on a dyad of trained experts will likely represent a waste of resources, as it does not provide any additional performance gains. Taken together, these results both complement prior findings (Schuldt et al., 2017; Bahrami et al., 2012) and provide new insights, highlighting that the extent of training received by an individual, the complexity of the task at hand, and the desired performance outcome are all critical factors that need to be accounted for when weighing up the benefits of collective decision making.

Despite the fact that collective decision making has been studied extensively (Heath and Gonzalez, 1995; Bose et al., 2017), prior studies have relied mostly on artificial or comparatively simple tasks (e.g., number line or dot estimation), which do not reflect the nature of human interaction in daily life nor account for the uncertainty and limited task-relevant knowledge that individuals often possess (see Supplementary Material Fig. S8(a), which demonstrates that, in the present case, 86% of participants had close to none, basic, or average zoology knowledge and zero had any expertise). The WildCam Gorongosa task studied here, however, is not only an established citizen science challenge engaging thousands of participants but also a task that shares significant features with other forms of online collaboration and information processing activities more broadly, given it requires both perceptual ability and general knowledge. Thus, it can be expected that the present results can be generalized to similar environments where collaboration may be substituted or complemented with specialized training to improve outcomes.

The finding that individuals and dyads with similar training do not perform significantly differently suggests that the pair advantage observed in highly controlled experimental studies employing one-dimensional tasks may be less observable in many multi-dimensional contexts. Yet, given some remaining shortcomings, this study also provides building blocks for future research that can help resolve some of the gaps in our understanding of the relationship between task complexity and the performance benefits experienced by interacting individuals.

A limitation of this study is the fact that only one type of task context was studied. Nevertheless, by focusing on a real-world task and advocating for a "solution-oriented" approach (Watts, 2017), we hope that our approach will inspire further research in computational social science on decision making using externally valid domain-general task contexts. Moreover, we note that the use of a preexisting citizen science task resulted in a significant proportion of participants reporting an interest in contributing further to citizen science, independent from working alone or in a dyad (see Supplementary Material Fig. S9 which shows that 68% of participants reported it as either somewhat or extremely likely that they would contribute to citizen science in the long term), thereby demonstrating the additional benefit to adopting such a task, that is, engaging participants as well as advance our understanding of collective decision making.

We believe our study will spur further experimental research using citizen science platforms; we especially encourage researchers to build on our task complexity definition and develop a standardized framework to enable reliable comparisons across citizen science task environments. In the context of collective decision-making research, in particular, when considering individuals with similar expertise who disagree, future studies will advance our understanding of the value of deliberation. Such studies might also vary the time that individuals and groups have to make classifications. Another way of improving on this work is to consider and monitor different strategies that team members take or the difference in features on which

they base their decisions. Alignment or misalignment between these could significantly influence team performance.

Finally, we welcome recent calls for the field to take greater inspiration from animal research (O'Bryan et al., 2020) and for researchers to consider how technologies and machines affect human interaction in online collaborative tasks (Rahwan et al., 2020) as additional avenues of research to explore and further advance our collective knowledge of collective intelligence.

Materials and Methods

This study was reviewed and approved by the Oxford Internet Institute's Departmental Research Ethics Committee (DREC) on behalf of the Social Sciences and Humanities Inter-Divisional Research Ethics Committee (IDREC) in accordance with the procedures laid down by the University of Oxford for ethical approval of all research involving human participants. Participants ($n = 195$) were recruited from the general public via the Nuffield Centre for Experimental Social Science (see Supplementary Material, Sec. S2).

Zooniverse answers

As all images seen by participants have already been evaluated on the Zooniverse website, Zooniverse estimation data were used to assess the performance. These data consist of the aggregated guesses of citizen scientists who used the platform prior to when the experiment was conducted. Whilst volunteer estimates have consistently been found to agree with expert-verified wildlife images (Swanson et al., 2016), it is noted that the ground truth data could still be missing correct attributes that only experts could identify. However, it is by definition impossible for any participant to have performed better than the citizen scientist-provided estimates; hence, we refer to these data as the "ground truth" throughout (see Supplementary Material, Sec. S3.1 for details on the content of the ground truth).

Performance analysis

To comprehensively compare individual versus collective classifications, we assess performance at a specific point and as a function of time. As our primary analysis, we analyze differences in performance change during the final phase of the testing stage. As a secondary analysis, we also examine how any differences evolve over the course of the entire testing stage. In order to do so, we split T_1 and T_2 into three equally spaced and non-overlapping intervals of length 10 and 15 minutes, respectively. The reason we opted for computing intervals, as opposed to examining performance trial by trial, for example, was because the aim of the study was not to assess the absolute improvements in performance

but to compare any change against initial performance during training—before participants were able to acquire any learned expertise.

Benchmarking changes in performance against average performance during this initial interval in turn ensures we can more precisely estimate whether individual training, interaction, or both provide performance gains, and how this changes across tasks and over time. We used two different time intervals because each stage was different in length and so we chose the highest factor that would still produce enough data points per stage to allow for at least two comparisons. Next to fulfilling the above criteria, the specific interval sizes were in turn chosen for clarity of presentation, as the results did not significantly change any of the reported findings when changing the size of the interval (see Supplementary Material Fig. S7).

The average change in performance was estimated separately for each interval by computing the difference in performance compared to the performance of the respective individual(s) in the first interval of T_1, which can be thought of as the natural baseline prior to training. In comparing individuals to dyads, two benchmarks are used: (a) a "non-interacting nominal dyad," defined as the average of the initial performance of both dyad members working individually (the sum is used for volume and efficiency, as these are additive processes), and, in the case of Mixed Dyads, (b) a "comparably trained individual," defined as the average initial performance of the dyad member with the same level of training as the individual in the respective solo condition (multiplied by two when considering volume and efficiency).

Given the sample size, we opted to employ analysis of variance (ANOVA) to statistically qualify the results, instead of applying growth curve modeling, for instance. Similarly, given the length of the experiment, as well as the non-stationary nature of our temporal data, we considered other time-based analysis methods, such as a sliding window approach using a distinct or overlapping time interval, unsuitable as most assumptions would not be met.

Statistical tests

All statistical tests were two-tailed, and non-parametric alternatives were planned if the data strongly violated normality assumptions. For omnibus tests, the significance of mean differences between groups was analyzed via planned pairwise comparison tests. Given the fact that all pairwise comparisons were planned, and taking into account the sample size, length of the experiment as well as the nature of the data, we follow the recommendation that reporting actual p-values leads to fewer errors of interpretation (Rothman, 1990; Saville, 1990). This also means we avoid any arbitrary significance cutoff points (Vidgen and Yasseri, 2016). As a result, we do not correct p-values

but instead ensure that they are situated within the overall findings of the study. Effect size values were meanwhile interpreted in simple and standardized terms according to Cohen (1988), when no previously described values are available for comparison, and reported using conventions recommended by Lakens (2013). Details of the statistical tests are in Supplementary Material, Sec. S3.3.

Author's contribution

V.J.S. analyzed the data and drafted the manuscript. M.T. designed and performed the experiments. T.Y. designed and performed the experiments, designed the analysis, secured the funding, and supervised the project. All authors contributed to writing the manuscript and gave final approval for publication.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

This article earned Open Data and Open Materials badges through Open Practices Disclosure from the Center for Open Science. The Gorongosa Lab educational platform used to run the "WildCam Gorongosa task" is hosted by Zooniverse (www. zooniverse.com). Replication data and code are available at the Open Science Framework: https://osf.io/6rcgx/.

ORCID iD

Taha Yasseri https://orcid.org/0000-0002-1800-6094

Supplemental Material

Supplemental material for this article is available online.

References

Almaatouq A., Alsobay M., Yin M., et al. (2021) Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences of the United States of America* 118(36): e2101062118.

Almaatouq A., Noriega-Campero A., Alotaibi A., et al. (2020) Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America* 117(21): 11379–11386.

Bahrami B., Olsen K., Bang D., et al. (2012) Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology. Human Perception and Performance* 38(1): 3–8.

Bahrami B., Olsen K., Latham P. E., et al. (2010) Optimally interacting minds. *Science* 329(5995): 1081–1085.

Bang D., Aitchison L., Moran R., et al. (2017) Confidence matching in group decision-making. *Nature Human Behaviour* 1(6): 0117–0118.

Bang D., Fusaroli R., Tylén K., et al. (2014) Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition* 26(1): 13–23.

Becker J., Porter E. and Centola D. (2019) The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences of the United States of America* 116(22): 10717–10722.

Blanchard MD., Jackson SA. and Kleitman S (2020) Collective decision making reduces metacognitive control and increases error rates, particularly for overconfident individuals. *Journal of Behavioral Decision Making* 33(3): 348–375.

Bose T., Reina A. and Marshall J. A. (2017) Collective decision-making. *Current Opinion in Behavioral Sciences* 16: 30–34.

Cohen J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cox J., Oh E. Y. and Simmons B., et al. (2015) Defining and measuring success in online citizen science: A case study of zooniverse projects. *Computing in Science & Engineering* 17(4): 28–41.

Crook A. E. and Beier M. E. (2010) When training with a partner is inferior to training alone: The importance of dyad type and interaction quality. *Journal of Experimental Psychology. Applied* 16(4): 335–348.

De Cesarei A., Loftus G. R. and Mastria S., et al. (2017) Understanding natural scenes: Contributions of image statistics. *Neuroscience and Biobehavioral Reviews* 74: 44–57.

De Oliveira S. and Nisbett R. E. (2018) Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences of the United States of America* 115(9): 2066–2071.

DiCarlo J. J., Zoccolan D. and Rust N. C. (2012) How does the brain solve visual object recognition? *Neuron* 73(3): 415–434.

Farrell S. (2011) Social influence benefits the wisdom of indi-

viduals in the crowd. *Proceedings of the National Academy of Sciences of the United States of America* 108(36): E625.

Festinger L. (1954) A theory of social comparison processes. *Human Relations* 7(2): 117–140.

Galesic M., Barkoczi D. and Katsikopoulos K. (2018) Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision* 5(1): 1–15.

Galton F. (1907) Vox populi. *Nature* 75: 450–451.

Grasso A. and Convertino G. (2012) Collective intelligence in organizations: Tools and studies. *Computer Supported Cooperative Work* 21(4–5): 357–369.

Groen IIA, Jahfari S., Seijdel N., et al. (2018) Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS Computational Biology* 14(12): e1006690.

Hackman J. R. (1969) Toward understanding the role of tasks in behavioral research. *Acta Psychologica* 31: 97–128.

Healy A. F., Kole J. A. and Bourne L. E. Jr (2014) Training principles to advance expertise. *Frontiers in Psychology* 5: 131.

Heath C. and Gonzalez R. (1995) Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes* 61(3): 305–326.

Hong L. and Page S. E. (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America* 101(46): 16385–16389.

Kendal R. L., Boogert N. J., Rendell L., et al. (2018) Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences* 22(7): 651–665.

Kerr N. L. and Tindale R. S. (2004) Group Performance and Decision Making. *Annual Review of Psychology* 55(1): 623–655.

Koriat A. (2012) When are two heads better than one and why? *Science* 336: 360–362.

Koriat A. (2015) When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology. General* 144(5): 934–950.

Kozlowski S.W. J. and Bell B. (2013) Work groups and teams in organizations: Review update. *Handbook of Psychology.*

Lakens D. (2013) Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology* 4(NOV): 863.

Laughlin P. R. and Earley P. C. (1982) Social combination models, persuasive arguments theory, social comparison theory, and choice shift. *Journal of Personality and Social Psychology* 42(2): 273–280.

Lorenz J., Rauhut H., Schweitzer F., et al. (2011) How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America* 108(22): 9020–9025.

Magnussen S., Eilertsen D. E., Teigen K. H., et al. (2014) The probability of guilt in criminal cases: Are people aware of being 'beyond reasonable doubt. *Applied Cognitive Psychology* 28(2): 196–203.

Mahmoodi A., Bang D., Olsen K., et al. (2015) Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences of the United States of America* 112(12): 3835–3840.

Maloney L. T (2002) Statistical decison theory and biological vision. In: Heyer D and Mausfeld R (eds), *Perception and physical world*, p. 145–189. New York, NY: Wiley.

Mann R. P. and Helbing D. (2017) Optimal incentives for collective intelligence. *Proceedings of the National Academy of Sciences of the United States of America* 114(20): 5077–5082.

Mao A., Mason W., Suri S., et al. (2016) An experimental study of team size and performance on a complex task. *PLoS ONE* 11(4): e0153048.

Mata R., Pachur T., von Helversen B., et al. (2012) Ecological Rationality: A Framework for Understanding and Aiding the Aging Decision Maker. *Frontiers in Neuroscience* 6(FEB): 19.

Muchnik L., Aral S. and Taylor S. J. (2013) Social influence bias: A randomized experiment. *Science* 341(6146): 647–651.

Navajas J., Niella T., Garbulsky G., et al. (2018) Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* 2(2): 126–132.

Norouzzadeh M. S., Nguyen A., Kosmala M., et al. (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America* 115(25): E5716–E5725.

O'Bryan L., Beier M. and Salas E. (2020) *How Approaches to Animal Swarm Intelligence Can Improve the Study of Collective Intelligence in Human Teams.*

Parrish J. K., Burgess H., Weltzin J. F., et al. (2018) Exposing the science in citizen science: fitness to purpose and intentional design. *Integrative and Comparative Biology* 58(1): 150–160.

Rahwan I., Crandall J. W. and Bonnefon J. F. (2020) Intelligent machines as social catalysts. *Proceedings of the National Academy of Sciences of the United States of America* 117(14): 7555–7557.

Rothman K. J. (1990) No adjustments are needed for multiple comparisons. *Epidemiology* 1: 43–46.

Saffell T. and Matthews N. (2003) Task-specific perceptual learning on speed and direction discrimination. *Vision Research* 43(12): 1365–1374.

Saville D. J (1990) Multiple comparison procedures: the practical solution. *The American Statistician* 44(2): 174–180.

Schuldt J. P., Chabris C. F., Woolley A. W., et al. (2017) Confidence in dyadic decision making: The role of individual differences. *Journal of Behavioral Decision Making* 30(2): 168–180.

Sella F., Blakey R., Bang D., et al. (2018) Who gains more: Experts or novices? The benefits of interaction under numerical uncertainty. *Journal of Experimental Psychology. Human Perception and Performance* 44(8): 1228–1239.

Surowiecki J. (2004) *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes*

Business, Economies, Societies, and Nations. New York, NY: Doubleday Books).

Swanson A., Kosmala M., Lintott C., et al. (2016) A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology: The Journal of the Society for Conservation Biology* 30(3): 520–531.

Toyokawa W., Whalen A. and Laland K. N. (2019) Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour* 3(2): 183–193.

Trouche E., Sander E. and Mercier H. (2014) Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology. General* 143(5): 1958–1971.

Trouille L., Lintott C. J. and Fortson L. F. (2019) Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human-machine systems. *Proceedings of the National Academy of Sciences of the United States of America* 116(6): 1902–1909.

Tsvetkova M., Garciá-Gavilanes R. and Yasseri T. (2016) Dynamics of disagreement: Large-scale temporal network analysis reveals negative interactions in online collaboration. *Scientific Reports* 6(November): 36333–36410.

Tsvetkova M., Yasseri T., Meyer E. T., et al. (2017) Understanding human-machine networks: A cross-disciplinary survey. *ACM Computing Surveys* 50(1): 1–35.

Tump A. N., Wolf M., Krause J., et al. (2018) Individuals fail to reap the collective benefits of diversity because of over-reliance on personal information. *Journal of the Royal Society Interface* 01(55): 11.

Valeriani D., Cinel C. and Poli R. (2017) Group augmentation in realistic visual-search decisions via a hybrid brain-computer interface. *Scientific Reports* 7(1): 7772–7812.

Vidgen B. and Yasseri T. (2016) P-Values: Misunderstood and Misused. *Frontiers in Physics* 4(March): 10–14.

Vives X. (1996) Social learning and rational expectations. *European Economic Review* 40(3–5): 589–601.

Wahn B., Czeszumski A. and König P. (2018) Performance similarities predict collective benefits in dyadic and triadic joint visual search. *PLOS ONE* 13(1): e0191179.

Watts D. J. (2017) Should social science be more solution-oriented? *Nature Human Behaviour* 1(0015): 0015.

Wildcam Gorongosa (2020) *Wildcam Gorongosa*.

Wood R. E. (1986) Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37: 60–82.

Yaniv I. and Milyavsky M. (2007) Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* 103(1): 104–120.

Yi S. K. M., Steyvers M., Lee M. D., et al. (2012) The wisdom of the crowd in combinatorial problems. *Cognitive Science* 36(3): 452–470.