Measuring Discrimination Abilities of Monk Parakeets Between Discreet and Continuous Quantities Through a Digital Life Enrichment Application

Fig. 1. A Monk Parakeet selecting the largest value out of two displayed, in heap mode.

Fig. 2. A Monk Parakeet selecting the largest value out of four displayed, in disk mode.

Ain et al. measured three African Grey (*Psittacus Erithacus*) parrot's discrimination abilities between discrete and continuous quantities. Some features of their experimental protocol make it difficult to apply to other subjects and/or species without introducing a risk for some bias, as subjects could read cues from the experimenter (even though the study's subjects probably did not). Can digital life enrichment techniques permit us to replicate their results with other species with less risk for experimental bias, with a better precision, and at lower cost? Inspired by previous informal digital life enrichment experiments with parrots, we designed and tested a web application to digitally replicate and extend Ain et al.'s experimental setup. We were able to obtain similar results to theirs for two individuals from a distinct species of parrots, Monk Parakeets (*Myiopsitta Monachus*), with increased guarantees against potential experimental biases, in a way which should allow to replicate such experiments at larger scale and at a much lower cost.

Additional Key Words and Phrases: Comparative Cognition Study, Continuous and Discreet Comparative Abilities, Digital Life Enrichment, Monk Parakeet

ACM Reference Format:

. 2022. Measuring Discrimination Abilities of Monk Parakeets Between Discreet and Continuous Quantities Through a Digital Life Enrichment Application. In *ACI 2022: Ninth International Conference on Animal-Computer Interaction 5-8 December 2022, hosted by Northumbria University, Newcastle-upon-Tyne, UK.*. ACM, New York, NY, USA, 17 pages. https://doi.org/XXXXXXX.XXXXXXX

Fig. 3. A Cockatoo playing the game "Candy Crush" (picture used with the authorisation of the author).

Fig. 4. Monk Parakeet playing the piano music application "Mini Piano Lite" in order to learn to use touchscreen interfaces with a wide active surface.

Fig. 5. Monk Parakeets playing the steel drum music application "Meditation Drum" in order to learn how to properly aim when using touchscreen interfaces.

1 INTRODUCTION

Al Aïn et al. [1] measured the discrimination abilities between discrete and continuous quantities of three African Grey parrots (*Psittacus erithacus*), showing that their accuracy in choosing between two small quantities was inversely correlated with the ratio of the smallest quantity to the largest one.

Generalizing the experimental protocol described and implemented by Al Aïn et al. [1] to other subjects or species present some difficulties. The fact that the experimenter knows which answer is expected from the subjects is not an issue in their study because it was previously verified that the three subjects were unable to read such cues from human experimenters, but it means that the replication of such protocol is limited to individuals (from the same or from other species) which inability to read cues has been previously demonstrated. Beyond such a weakness, the cost of the experimental set-up and of the analysis of the video recordings of the experiments reduces the probability that such a protocol will be replicated with other subjects from the same species, or with subjects from the many other species of parrots existing around the world.

Touchscreens have been successfully used for experiments in life enrichment [4, 10, 17] and in Comparative Psychology [6], with individuals from various nonhuman species. Could Digital Life Enrichment techniques allow to replicate Al Aïn et al. [1]'s results at a lower cost, but with a better precision, and with less potential experimental bias? Which additional advantages could such digital variants bring?

Inspired by previous informal Digital Life Enrichment experiments such as a Cockatoo playing the video game `Candy Crush` (Figure 3), or Monk Parakeets learning to use touch interfaces by playing music on it (Figures 4 and 5), we designed, tested and used a Digital Life Enrichment web application to digitally replicate and extend Al Aïn et al. [1]'s experimental setup. We obtained similar results to that of Ain et al. for two individuals of a distinct species of parrots, Monk Parakeets (*Myiopsitta Monachus*), using an experimental protocol with increased guarantees against potential experimental biases, at a lower set-up cost, with additional advantages brought by the digital context, such as automatic logging and increased subject's agency. After describing a selection of concepts and results in the research area of comparative psychology (Section 2), we describe the application (Section 3), an experimental protocol (including separate development, training and testing phases) based upon it (Section 4), an implementation of this protocol and an analysis of its results (Section 5), and we conclude with a recapitulation of our results, a discussion of their potential weaknesses and a perspective on future research (Section 6).

2 COMPARATIVE PSYCHOLOGY

Comparative psychology refers to the scientific study of the behavior and mental processes of non-human animals (referred to as "nonhumans" thereafter), especially as these relate to the phylogenetic history, adaptive significance, and development of behavior in many different species, from insects to primates. The abilities of nonhumans, traditionally less studied than that of humans, has been receiving more attention in the last half century. Such studies started with the nonhumans perceived to be "closest" to humankind, such as apes [4, 17], and has spread more recently to birds [1, 5, 16].

2.1 Discrimination Abilities in African Grey parrots

Al Aïn et al. [1] tested the discrimination abilities of African Grey (*Psittacus erithacus*) parrots on discrete and continuous amounts. More precisely, they investigated the ability of three African grey parrots to select the largest amount of food between two sets, in two types of experiments. In the first experiment type, the subjects were tested on discrete quantities via the presentation of two distinct quantities of sunflower seeds, between 1,2,3,4 and 5 seeds. In the second experiment type, the subjects were tested on continuous quantities via the presentation of two distinct quantities of parrot formula, with amounts between 0.2,0.4,0.6,0.8 and 1 ml. For each experiment, the two amounts were presented simultaneously and were visible at the time of choice. Albeit the subjects sometimes failed to choose the largest value, they always performed above chance, their performance improving when the difference between amounts was the greatest.

The experimental setup was completely physical. A permanent table was set-up in the aviary, and two black pieces of cardboard were used to present food item (sunflower seeds or parrot formula). For each experiment, different amounts of either seeds or parrot formula were placed on each piece of cardboard. The experimenter put the subject for 5 seconds in a position from which they could observe the two sets, then placed them on the table at equal distances from the two sets, letting them chose one set to it while removing the ignored set. The position of the sets (small and large) was pseudo-randomized: the larger set was never presented more than two times on the same side and was presented as often on the right side as on the left side.

In the experimental setup described by Al Aïn et al. [1], some subjects could potentially read involuntarily cues from the experimenter: even though the experimenter was standing behind the subject, at equal distances from each set, not pointing to it, looking at the subject, aiming to avoid communicating any cue to the subject, the experimenter *knew* where the largest quantity was. While it was not an issue with the specific subjects of Al Aïn et al. [1]'s study because the authors demonstrated in a previous study that they were not able to use any gazing cue, the protocol should not be applied as such to other subjects without verifying their inability to read such cues, adding to the cost of implementing such protocol.

Avoiding giving cues to the subject is hard even for a professionally trained experimenter [21]. Requiring either such training or a separate study to insure that the subject cannot read cues from the experimenter restricts the applicability of a protocol to laboratories. For example, in the context of *citizen science* projects [8] where non professional experimenters (such as zoo personal or simple citizen) would guide the experiments (see Section 6.3.4 for a short discussion of such potential project), a masked protocol where the experimenters *ignore* what the correct answer is (because they did not receive the information that the subject did) would be more robust against subjects reading cues from the experimenter. We describe in Section 3 an application allowing for such an alternate experimental setup which, if not exactly equivalent to that of Al Aïn et al. [1] (e.g. the reward is not proportional to the quantity selected), presents the advantage of being "experimenter-masked", inspired by some of the life enrichment experiences described in the next section.

2.2 Life Enrichment and Comparative Psychology

The study of the abilities of nonhumans and the use of life enrichment activities in general, and digital ones in particular, have been interconected from their very begining. In 1990, when Richardson et al. [19] describe a Computerized Test System to measure some habilities in a population of rhesus monkeys, they mention that "*the animals readily started to work even when the reward was a small pellet of chow very similar in composition to the chow just removed from the cage*", and that "*the tasks have some motivating or rewarding of their own*".

Furthermore, nonhuman subjects seem to enjoy participating in cognitive studies involving game-like digital applications. Washburn [22] describes, among various other anecdotes, how game-like application for apes were developed as early as 1984, and how the subjects "*chose to work on joystick-based tasks, even though they did not need to perform the game-like tests in order to receive food*", and "*opted for computer task activity over other potential activities that were available to them*". Lastly, he describes evidence that the subjects were not only motivated by food rewards, but also by the enjoyment of the tasks themselves: when given a choice between completing trials for pellets or receiving pellets for free but not being able to play the game-like tasks during the free-pellet period, the subjects chose to work for their reward.

2.3 Experimental Biases

The history of Comparative Psychology has been prone with fights about the validity of methodologies and results: Pepperberg [14] describes various such tensions between researchers about the psychology of animals, with some accusing other researchers in the field to be "*liars, cheats and frauds*", and she highlights how sign language researchers were accused of "*cuing their apes by ostensive signals*" and of "*consistently over-interpreting the animals' signs*".

2.3.1 Selective Reporting Bias. Selection biases occur in a survey or experimental data when the selection of data points is not sufficiently random to draw a general conclusion. Selective reporting biases are a specific form of selection bias whereby only interesting or relevant examples are cited. Cognitive skills can be particularly hard to study in nonhumans, requiring unconventional approaches but often presenting the risk of such biases. For example, an experimenter who would present a subject repeatedly with the same exercise could be tempted to omit or exclude bad performances (eventually attributing them to a "bad mood" of the subject, which stays a real possibility) and report only on good performances, creating a biased representation of the abilities of the subject, a selective reporting bias. We describe how to use a digital application to systematically log the result and easily avoid such bias in Section 3.

2.3.2 "Clever Hans" effect. Among such methodological issues resulting in experimental biases, the most iconic one might be the case of the eponymous horse nicknamed "Clever Hans" which appeared to be able to perform simple intellectual tasks, but in reality relied on involuntary cues given by not only by their human handler, but also by a variety of human experimenters. It is possible to avoid the confusion between a subject's ability to read cues from the experimenter from its ability to answer the tests presented to them by such an experimenter. The principle is quite simple: make sure that the experimenter does not know the test, by having a third party out of reach from the subject's reading to prepare the test. Whereas such experimental setup was historically referred to as a "Blind Setup" or a "Blinded Setup", we follow the recommendations of Moris et al. [12] and prefer the term of "masked" to the term "blind" when describing the temporary and purposeful restricted access of the experimenter to the testing information. In the next section, we describe an application designed so that to facilitate a type of "masked" experimental set-up, in

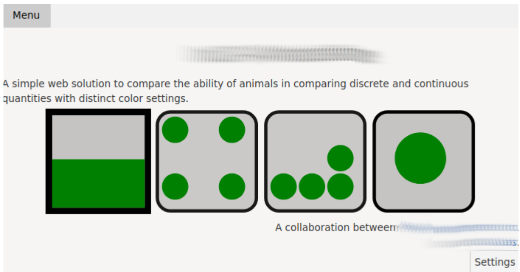

Fig. 6. The main menu of the application is designed so that the subject can choose in which visualisation mode it wishes to play, in the hope to support a sense of agency. The name of the application and the collaborations were blurred to protect the anonymity of the submission.

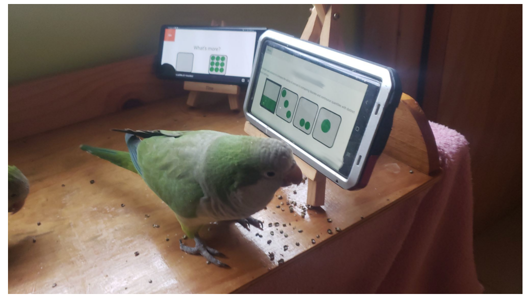

Fig. 7. Both subjects quickly learned to select a display mode to start a game, but did not seem to show a preference for a display mode in particular. The name of the application and the collaborations were blurred to protect the anonymity of the submission.

which it is guaranteed that the ability of the subject to read cues from the experimenter does not affect the result of the experiment, as the experimenter himself ignores the question (and hence its correct answer) being asked to the subject.

3 APPLICATION

3.1 Application's Structure

The web application is composed of four views. The first two, the `Main Menu` (described in Figures 6 and 9) and the `Gaming View` (which can be seen in Figures 12 and 17 among others), are especially designed to be navigable by nonhuman subjects. The access to the two others, the `settings` (see Figures 8 to 10) and the `information` views are tentatively restricted to the experimenters by requesting the long pressing of a button.

3.2 Masked Experimental Setup

Among other features, the web application was designed to facilitate digital experiments similar to that performed by Al Aïn et al. [1] but in a way such that the experimenter does *not* know where the "correct" answer is, a masked experimental setup (see Figures 11 to 13 for an illustration and pictures illustrating such a setup). This insures that the subject cannot receive any voluntary or involuntary cue from the experimenter. Such a purpose is achieved through the extensive audio feedback system, which aims at notifying the experimenter about any event which requires their intervention (e.g. rewarding or encouraging the subject, or acknowledging that the subject does not want to play this game any more), so that they do not need to check the screen of the device at any point.

3.3 Logging structure

In traditional, non digital experiments in comparative psychology, the experiments is usually recorded on video so that the video recording can be later processed in order to generate an extensive log of the interactions of the subject during the experiment. Such a task is long and tedious, and no video processing software is yet able to automatize such a process. An important advantage of a digital experimental set-up, such as that allowed by the software that we developped, is the ability to *automatically* log the interactions of the subject with the application. The software generates logs with data to be analyzed by researchers, including information on both the test performed and the subject's performance (see Figure 15 for a short extract).

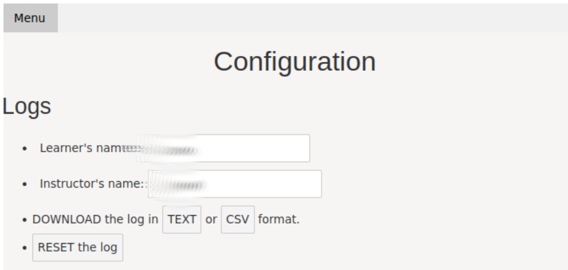

Fig. 8. The logs are exported in the top part of the setting page of the application. The names were blurred to protect the anonymity of the submission.

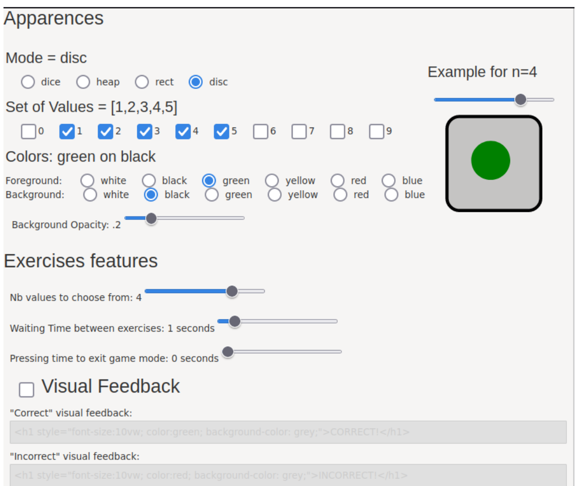

Fig. 9. The part of the setting page dedicated to the appearance and difficulty of the exercises.

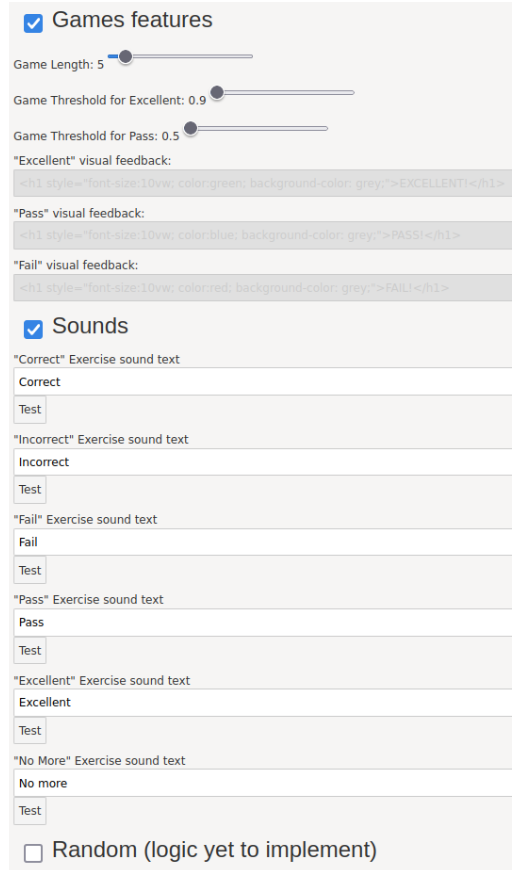

Fig. 10. The part of the setting page dedicated to the game features and sound feedback.

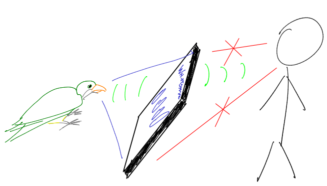

Fig. 11. An artistic rendition of a masked experimental setup. The subject (left) can see the display and hear the device (center), but the experimenter (right) can hear the device but not see its display.

Fig. 12. Example of masked experimental set-up: the experimenter can hear the instructions from the device and encourage and reward the subject, but cannot give any cue about correct answers.

Fig. 13. The masked experimental set-up as viewed by the experimenter, with two subjects participating in the experiment at the same time, each with its own device.

```
Test no, Test Name, Learner, Trainer, C_0, C_1, C_2, C_3, C_4, Value selected , Correction , Date, Answering Time (ms), Other Parameters
1, dice, Subject, Experimenter, 1,4,,,, 4,true, [2022-05-19 17:02(25.981)], 7946, background black, foreground green, bg opacity .2, Value Set [1,2,3,4,5]
(...)
81, rect, Subject, Experimenter, 4,2,3,,, 3,false, [2022-05-19 17:26(55.124)], 4655, background black, foreground green, bg opacity .2, Value Set [1,2,3,4,5]
(...)
180, heap, Subject, Experimenter, 3,2,1,,, 2,false, [2022-05-19 17:35(06.6)], 926, background black, foreground green, bg opacity .2, Value Set [1,2,3,4,5]
```

Fig. 14. A short extract showing four selected lines of the log generated by the application for the afternoon session of the 19th of May 2022 (deleted blocks of lines are marked by "`(...)`"). See Figure 15 for a more readable reformatting of the same extract. Log entries such as "`background black, foreground green, bg opacity .2`" refer to visualisation options, not used in this work.

Test no	Test Name	C0	C1	C2	C3	C4	Value selected	Correction	Date	Other Parameters
1	dice	1	4				4	true	[2022-05-19 17:02(25.981)]	Value Set [1,2,3,4,5]
81	rect	4	2	3			3	false	[2022-05-19 17:26(55.124)]	Value Set [1,2,3,4,5]
180	heap	3	2	1			2	false	[2022-05-19 17:35(06.6)]	Value Set [1,2,3,4,5]

Fig. 15. A more readable format of the log extract from Figure 14, with less relevant columns removed for readability. Observe that the subject was offered to choose the largest between 2 (on the first test) and 3 (on the 81st and 180th tests) values, represented as `dice` (first test), `rect` (81st test) and `heap` (180th test), and that the subject chose once correctly, and two times incorrectly, in games where the values were taken from the set $\{1, 2, 3, 4, 5\}$, with the precise time and date of each answer duly recorded. The columns labeled C3 and C4 are empty because no test was performed requesting the subject to choose the maximal value between 4 or 5.

4 EXPERIMENTATION PROTOCOL

4.1 Phases of the protocol

The protocol was implemented in three phases: a phase of *development* (of the software) with only one subject (the first one) interacting with the application, a phase of *training* with two subjects interacting with the application in a mix of unmasked and masked protocols, and a phase of *testing* using exclusively the masked protocol and collecting data with both subjects.

4.2 Experimentation Subjects

We experimented with two subjects, both Monk Parakeets (*Myiopsitta Monachus*), the first one participating in the development phase, and the second one, younger, sometimes voluntarily joined during the training and testing phases. Both can be considered as pets (as opposition to research animals used in professional laboratories). The first subject is 6 year old, male, and has been trained since he was 1 year old. The second subject is 3 year old, female, and has been trained since she was 1 year old. Both are fully able of flight, were previously trained to perform tricks (e.g. sorting coins by size, putting small basket balls in small suspended hoops, etc.), and to use touch screens via the use of music applications, first with a piano application making the whole surface of the screen active (see Figure 4 for a picture), then with a steel drum application reducing the active surfaces of the screen to a few circles (see Figure 5 for a picture).

4.3 Ethical Precautions

4.3.1 Physical settings. At no point were the subjects food or water deprived: at any point they could fly to their housing space, where food and water was available.

4.3.2 Application Usability. In order to minimize the potential frustration of the subjects when facing inadequate answers from the application, each version of the application was systematically tested by two human subjects, and any issue detected during such a phase corrected, before being presented to the nonhuman subjects. During the phase of software development, when a feature of the application (whether due to an error or to an setting proved to be

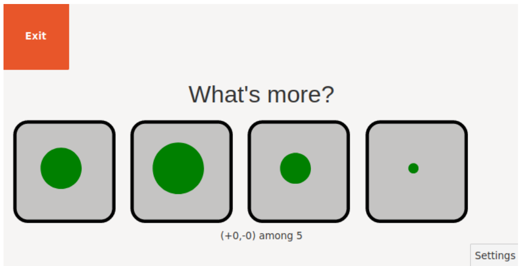

Fig. 16. A screenshot of the game view of the application, asking to choose the largest disk out of four. Top left is the orange "Exit" button actionable by the subject. Bottom right is the setting button requesting a long pressure to be activated. Bottom center is a summary of the game's score.

Fig. 17. Monk Parakeet selecting the largest disc out of four.

inadequate) was encountered to frustrate the subjects, the use of this application was replaced by another activity until the software was corrected, tested and separately validated by two human subjects.

4.3.3 Sense of Agency. The physical setting of the experimentation was designed so that to insure that the subject's participation was voluntary during all three phases of the process: the subjects were invited to come to the training area (but could, and sometime did, refuse); at any time the subjects could fly from the training area back to their aviary, to a transportation pack with a large amount of seeds suspended above the training area, or to an alternate training area on the side, presenting an alternate choice of training exercises. Concerning the psychological aspects, the main menu of the application was designed so that each subject can choose in which visualisation mode they wish to play (see Figures 6 and 7), and a large orange "exit" button is present on the playing screen allowing the subject to signal that they do not wish to play this game any more, and to return to the main menu.

4.3.4 Approval of the experimental protocol by IACUC. All interactions with animals were governed by a protocol reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) of the researchers' institution, through a form of Experimentation Protocol of Management and Care of Animals.

4.4 Statistical Analysis Process

4.4.1 Statistical tools used. The statistical analysis was performed in a python notebook, executed and shared via the collaborative website https://colab.research.google.com. Such python notebook was developed and tested on the logs generated during the (masked and unmasked) training sessions, to be used later without major modification on the logs generated during the masked experimental sessions of the testing phase.

4.4.2 Binomial Tests. The average accuracy of each subject for each display mode and each size of the set of values presented to the subject is then the average of the `Correction` entry in the log (replacing `True` by `1` and `False` by `0`) over all data points matching the criteria. For each such accuracy, we performed a `binomial test` in order to decide if such accuracy was substantially better than that achieved by selecting a value uniformly at random. We performed such statistical analysis on the data of each particular session and on their union, on each particular visualization mode and on the type of visualisation mode (discrete or continuous) and on all visualisation modes (see Tables 2, 3, 5 and 6).

Subject	Set Size	Dice	Heap	Discrete	Disc	Rectangle	Continuous	Total
1	2	449	400	849	103	262	448	1214
1	3	249	120	0	126	120	0	588
1	4	154	51	205	13	0	13	218
2	2	190	0	190	193	26	219	409
1	total	852	571	1054	242	382	461	2020
2	total	190	0	190	193	26	219	409
total	total	1042	644	1244	435	468	680	2429

Table 1. Number of data points collected separated by display modes ("Dice", "Heap", "Disc" and "Rectangle"), accumulated by the type of display mode ("Discrete" or "Continuous") and accumulated over all display modes ("Total"). The imbalance between the frequencies of the display modes and between the amounts of test results for each subjects is explained by the care to support the agency of the subjects: they could interrupt the session at any time, and had the option to choose the display mode at any time (which they seldom did).

4.4.3 Pearson Correlation Analysis. In order to compare our results with that of Al Aïn et al. [1]'s experiments, we performed a Pearson correlation analysis of the relation between the accuracy of the subjects' answers when asked to select the maximal out of two values on one hand, and the three variables they considered on the other hand:

- the *sum* of the values for each test (e.g. from $1 + 2 = 3$ to $4 + 5 = 9$),
- the *difference* between the two extreme values presented within a trial (e.g. from 1 to $5 - 1 = 4$) and
- the *ratio* of continuous quantities presented, by dividing the smallest presented value by the largest one (e.g. from $\frac{1}{5} = 0.2$ to $\frac{4}{5} = 0.8$).

5 RESULTS

After relatively long phases of development and training (15 months) using various domains of values (from $\{0, 1\}$ to $\{0, 1, \ldots, 9\}$), the experimental phase was quite short (one week), with all experiments performed using a masked setup and a domain of values restricted to the set $\{1, 2, 3, 4, 5\}$. A testing session typically lasted some 5 to 10 games of 20 questions each, resulting into a log of 100 to 200 data points: see Figures 14 and 15 for a shortened example of log. The testing phase occurred between the 19th of May 2022 and the 26th of May 2022. The experiments used four different display modes ("Dice", "Heap", "Disc" and "Rectangle"), requesting the subject to select the maximal value out of a set of 2, 3 or 4 values, randomly chosen among a set of five values $\{1, 2, 3, 4, 5\}$, in order to produce a setup relatively similar to that of Al Aïn et al. [1], with the vast majority of experiments selecting the maximal out of two values, and only a few out of three or four values. Each log corresponds to a separate training session and device, containing between 80 and 400 entries (each entry being a separate question and answer). While the subjects were free to choose the display mode of their choice (which they seem to choose at random) during the training phase, the experimenter chose the initial display mode during the testing phase, and very rarely did any subject use the exit button to choose a distinct display mode. In total, 14 logs were collected for the first subject, and 5 logs were collected for the second subject: the first subject was requested to select the maximal value out of 2,3 or 4 values, while the second subject was requested to select the maximal value only out of 2 values. See Table 1 for a summary of the number of data points collected separated by display modes ("Dice", "Heap", "Disc" and "Rectangle"), accumulated by the type of display mode ("Discrete" or "Continuous") and accumulated over all display modes ("Total").

Session	Dice	Heap	Discrete	Disc	Rectangle	Continuous	Total
19,17h	$65(1e^{-1})$	$60(2e^{-1})$	$62(7e^{-2})$	$90(2e^{-4})$	$75(2e^{-2})$	$82(2e^{-5})$	$72(3e^{-5})$
21,17h	$80(5e^{-17})$	$93(1e^{-15})$	$84(1e^{-29})$	$91(3e^{-5})$	$95(3e^{-14})$	$94(3e^{-18})$	$86(6e^{-45})$
23,08h	$80(1e^{-5})$	$84(2e^{-3})$	$81(8e^{-8})$	(no data)	$90(8e^{-15})$	$90(8e^{-15})$	$86(1e^{-20})$
23,15h	$70(5e^{-3})$	$86(1e^{-20})$	$82(8e^{-21})$	$88(1e^{-8})$	(no data)	$88(1e^{-8})$	$83(1e^{-27})$
24,10h	$66(1e^{-4})$	(no data)	$66(1e^{-4})$	(no data)	(no data)	(no data)	$66(1e^{-4})$
24,17h	(no data)	$83(3e^{-16})$	$83(3e^{-16})$	(no data)	(no data)	(no data)	$83(3e^{-16})$
25,08h	(no data)	(no data)	(no data)	$60(3e^{-1})$	$86(4e^{-14})$	$84(1e^{-13})$	$84(1e^{-13})$
25,13h	$71(3e^{-2})$	(no data)	$71(3e^{-2})$	(no data)	(no data)	(no data)	$71(3e^{-2})$
Total	$74(3e^{-25})$	$85(78e^{-49})$	$79(6e^{-69})$	$86(81e^{-15})$	$89(3e^{-40})$	$88(2e^{-53})$	$82(1e^{-117})$

Table 2. Finer analysis of the first subject's performance on selecting the maximal value out of two, separated by display modes ("Dice", "Heap", "Disc" and "Rectangle"), accumulated by the type of display mode ("Discrete" or "Continuous") and accumulated over all display modes ("Total"). The sessions occurred during the month of May 2022 and are identified by the date d and hour h (e.g. the session which occurred at 17:02 on the 19th of May 2022 is identified by the tag "19,17h"). Each entry is in the format $a(p)$ where a is the accuracy reported, and p is the probability of achieving such accuracy or better by selecting answers uniformly at random. Note how the accuracy percentages are mostly above 80%, and that the probability of such accuracy or a better one to be attained by selecting answers uniformly at random is smaller than 0.001 in almost all the cases.

Session	Dice	Heap	Discrete	Disc	Rect	Continuous	Total
21,10h	$84(5e^{-15})$	(no data)	$84(5e^{-15})$	(no data)	$73(1e^{-2})$	$73(1e^{-2})$	$82(6e^{-16})$
23,15h	$64(5e^{-2})$	(no data)	$64(5e^{-2})$	(no data)	(no data)	(no data)	$64(5e^{-2})$
24,08h	(no data)	(no data)	(no data)	$79(5e^{-13})$	(no data)	$79(5e^{-13})$	$79(5e^{-13})$
24,17h	$51(5e^{-1})$	(no data)	$51(5e^{-1})$	$57(2e^{-1})$	(no data)	$57(2e^{-1})$	$55(2e^{-1})$
Total	$74(3e^{-12})$	(no data)	$74(3e^{-12})$	$73(2e^{-11})$	$73(1e^{-2})$	$73(1e^{-12})$	$74(2e^{-23})$

Table 3. Finer analysis of the second subject's performance on selecting the maximal value out of two, separated by display mode and combined. Note how the accuracy percentages are in between 51% (not much better than random, on the last session) and 84% with an average of 74% (both much better than random), and that the probability of such accuracy to be attained by selecting answers uniformly at random is $p < 0.001$ in almost all the cases.

5.1 Selecting the maximal value out of two

5.1.1 First Subject. The results show a clear ability from the first subject to discriminate the maximal value out of two quantities. Over all experimentations requesting to select the maximal value out of two, the first subject responded correctly 993 times out of a total of 1214 trials, corresponding to an average accuracy of 81.79%. A simple binomial tests indicates that the probability to achieve such an accuracy by answering uniformly at random 1214 such binary questions is $p = 1.95 \cdot 10^{-117}$ (see Table 2 for a more detailed description of the results, by session and by display mode).

5.1.2 Second Subject. The second subject was more reluctant to participate, often staying in the housing space or choosing among other activities available, but showed a similar ability when they did. Overall experimentations requesting to select the maximal value out of two during the testing phase, the second subject responded correctly 303 times out of a total of 409 trials, corresponding to an average accuracy of 74%. A simple binomial tests indicates that the probability of answering correctly 303 or more such binary questions out of 409 by answering uniformly at random is $p = 2.24 \cdot 10^{-23}$.

5.1.3 Relation between accuracy and variables. When selecting the maximal value out of two, both subjects showed a lower accuracy when the two values were close (difference or ratio close to 1): see Table 4 for the percentages of correct

Value Set	Total	Difference	Ratio	Accuracy	
(x, y)	$x + y$	$y - x$	y/x	1st Subject	2nd Subject
{1,2}	3	1	0.5	81%	69%
{1,3}	4	2	0.33	90%	70%
{1,4}	5	3	0.25	93%	78%
{1,5}	6	4	0.2	94%	94%
{2,3}	5	1	0.66	82%	57%
{2,4}	6	2	0.5	81%	68%
{2,5}	7	3	0.4	96%	76%
{3,4}	7	1	0.75	67%	45%
{3,5}	8	2	0.6	73%	70%
{4,5}	9	1	0.8	55%	71%

Table 4. Both subject's *Accuracy* for each pairs of values from the domain $\{1, 2, 3, 4, 5\}$. *Total* is the total value of the representation shown to the test subject, *Difference* is the difference between the two values presented, *Ratio* is equal to the smallest quantity divided by the largest quantity. For the first subject, note how the lowest *Accuracy* (55%) corresponds to the highest *ratio* (0.8), while for the second subject the lowest *Accuracy* (45%) corresponds to the second highest *ratio* (0.75), suggesting a trend confirmed by the Pearson's correlation tests.

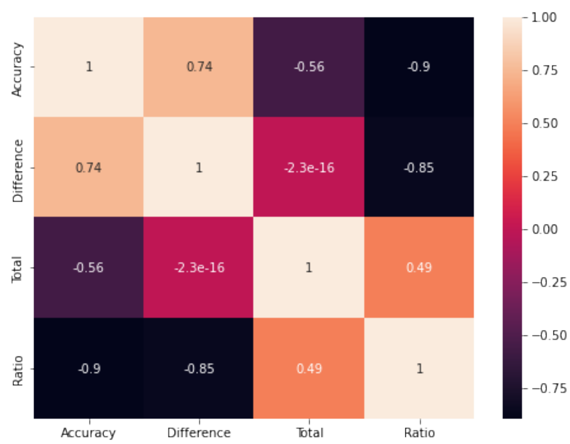

Fig. 18. Heat map correlation plot between the variables described in Table 4 for the first subject. Notice the strong negative correlation (-0.9) between *Accuracy* and *Ratio* on one hand, and the strong positive correlation (0.74) between *Accuracy* and *Difference* on the other hand.

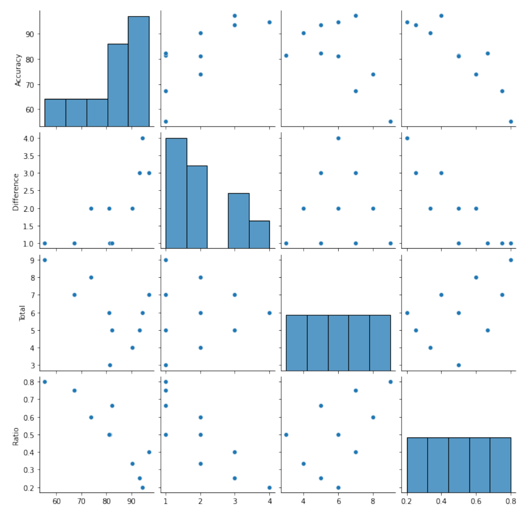

Fig. 19. Scatter-plot of the variables described in Table 4 for the first subject. The diagonal plots show the distribution of the values of each variable. Note the uniform distribution of the *Total* and *Ratio*.

answers for each subject and each of the 10 sets of values presented (ignoring the order). Such results corroborate those of the three African Grey parrots in Al Aïn et al. [1]'s study.

Pearson's correlation tests for the first subject (see Figure 18 for the corresponding heat map and Figure 19 for the corresponding scatter plots) suggest an inverse correlation between the accuracy of the subject's selection and the ratio of the smallest value to the largest one: for example, for a combination with small ratio $\frac{1}{5} = 0.2$, the subject is more likely to correctly select the maximal value. There is a strong negative correlation ratio of $r = -0.9$ between the

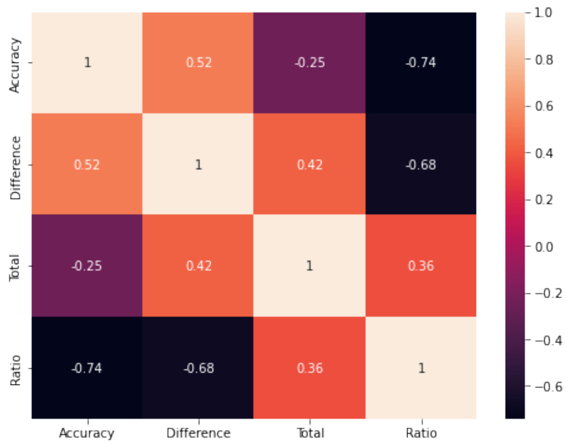

Fig. 20. Heat map correlation plot between the variables described in Table 4 for the second subject. Notice the negative correlation (−0.74) between *Accuracy* and *Ratio*.

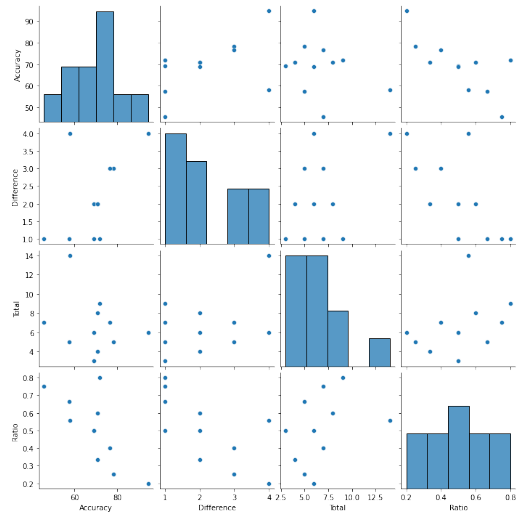

Fig. 21. Scatter plots for the variables described in Table 4 for the second subject.

accuracy and the ratio, and a positive correlation ratio of $r = 0.74$ between the accuracy and the difference (see the heat map in Figure 18). The scatter plots (in Figure 19) show a decreasing relationship between the accuracy and the ratio, and an increasing relationship between the accuracy and the difference.

There is a similar correlation between accuracy and ratio in the results of the second subject (see the heat-map in Figure 20 and the scatter plots in Figure 21). There is a strong negative correlation ratio of $r = -0.74$ between the ratio and the accuracy. The correlation ratio of $r = 0.52$ between the difference and the accuracy is much weaker than for the first subject's results.

5.2 Selecting the maximal value out of three and four values

Only the first subject was tested on selecting the maximal value out of three and four values. The subject performed with a lower accuracy in these contexts than in the previous one: on average they achieved an accuracy of 70% for selecting the maximal out of three values and 62% for selecting the maximal out of four values, but still much better than what would be expected (33% and 25% respectively) if the subject chose uniformly randomly among the values proposed (see Tables 5 and 6 for the detailed performances separated by display mode and sessions). Two simple binomial tests give a more formal measure of how much better the subject performed compared to one choosing uniformly at random: the probabilities of obtaining an accuracy equivalent or superior by randomly choosing the same number of answers is $p = 3.479 \cdot 10^{-77}$ with probability 0.33 of success (for selecting the maximal out of 3 values 588 times) and $p = 2.549 \cdot 10^{-31}$ with probability 0.25 of success (for selecting the maximal out of 4 values 136 times): with very high probability, the subject showed their ability to discriminate between three and four values.

Session	Dice	Heap	Discrete	Disc	Rect	Continuous	Total
19,17h	$55(4e^{-3})$	$75(2e^{-4})$	$61(7e^{-6})$	$6(1e^{-2})$	$85(3e^{-6})$	$72(5e^{-7})$	$66(3e^{-11})$
22,09h	$51(4e^{-3})$	$65(4e^{-4})$	$56(1e^{-5})$	$78(1e^{-10})$	(no data)	$78(1e^{-10})$	$64(2e^{-13})$
22,11h	$57(1e^{-4})$	$64(9e^{-6})$	$60(6e^{-9})$	$90(1e^{-16})$	$89(6e^{-31})$	$89(3e^{-46})$	$77(2e^{-47})$
25,16h	$67(6e^{-12})$	(no data)	$67(6e^{-12})$	(no data)	(no data)	(no data)	$67(6e^{-12})$
Total	$59(1e^{-17})$	$66(1e^{-11})$	$61(2e^{-27})$	$80(1e^{-25})$	$88(7e^{-36})$	$84(2e^{-59})$	$70(3e^{-77})$

Table 5. Finer analysis of the first subject's performance on selecting the maximal value out of three, separated by display mode and combined. Note how the average accuracy of random position selection in this case is 33%, so an accuracy between 51% and 84% is a reasonable measure, as well as the probability to achieve such accuracy or above when choosing one of the value between 3 options uniformly at random

Session	Dice	Heap	Discrete	Disc	Rect	Continuous	Total
25,16h	$59(7e^{-20})$	(no data)	$59(7e^{-20})$	(no data)	(no data)	(no data)	$59(7e^{-20})$
26,09h	(no data)	$72(1e^{-12})$	$72(1e^{-12})$	$53(2e^{-2})$	(no data)	$53(2e^{-2})$	$68(2e^{-13})$
Total	$59(7e^{-20})$	$72(1e^{-12})$	$62(2e^{-30})$	$53(2e^{-2})$	(no data)	$53(2e^{-2})$	$62(3e^{-31})$

Table 6. Finer analysis of the first subject's performance on selecting the maximal value out of four, separated by display mode and combined. Note how the average accuracy of random position selection in this case is 25%, so an accuracy between 53% and 72% is a reasonable measure, as well as the probability to achieve such accuracy or above when choosing one of the value between 4 options uniformly at random

6 CONCLUSION

6.1 Achievements

Whereas Al Aïn et al. [1]'s protocol requested the subject to choose between two pieces of cardboard holding distinct amount of food, for discrete and continuous types of food material; we proposed a protocol which requests the subject to choose the largest among a set of values (of parameterized size) on a visual display, using discrete and continuous representations of values, by touching a touchscreen on the representation of the largest value. By developing a simple but extensively parameterized Digital Life Enrichment web application requesting the user to select the largest among two to four values chosen at random, using discrete and continuous representations of values and providing visual and audio feedback about the correctness of the answer, we demonstrated that an experiment using a digital application can reproduce results from purely physical experiments with better guarantees against subjects reading potential cues from the experimenter, with an increased agency of the subjects of the experiments, increased settings (choosing the maximal value out of tuples rather than just pairs), with more diverse representations (four instead of two), at a lower cost allowing to increase the number of experiments ten fold, allowing us to use truly random instances, while automatically generating logs with more information and with better precision than in traditional settings. As a side result, we gathered good arguments that Monk Parakeet parrots are at least as able to discriminate between small discrete and continuous quantities as African Grey parrots.

6.2 Discussion

6.2.1 Non proportional rewards and reward withdrawal. The protocol defined by Al Aïn et al. [1] instructs to reward the subject with the content of the container they chose: the importance of the reward is proportional to the value being selected. The protocol we defined instructs to reward the subject with a single type of reward each time it does select the maximal value of the set, and to withdraw such reward when the subject fails to do so. Such a difference might alter the results of the experiment in at least two distinct ways:

- The proportionality of rewards could result in a larger incentive to select the maximal value when the difference between the two values is the largest, and a reduced incentive when the difference is small, and Al Aïn et al. [1] indeed noticed a correlation between the gap between the two values and the accuracy of the answer from the subjects of their experiment. The absence of such proportionality in our experiments might have reduced such an incentive, but we observed the same correlation than they did (described in Section 5.1.3).
- The withdrawal of rewards when the subject fails to select the largest value of the set is likely to affect the motivation of the subject to continue to participate in the exercise on the short term, and in the experiment in the long term. To palliate the frustration caused by such withdrawal, extensive care was taken to progressively increase the difficulty of the exercises (first through the size of the domain from which the values were taken, then through the size of the set of values from which to select the maximal one). No frustration was observed, with both subjects often choosing to continue playing at the end of a game.

Implementing the proportionality of rewards is not incompatible with the use of a digital application. For instance, it would be relatively easy to extend the web application to vocalize the value selected by the subject, so that the experimenter could reward the subject with the corresponding amount of food. Such an extension was not implemented mostly because it would slow down the experimentation, for relatively meagre benefits.

6.2.2 Irregular pairs and tuples. The web application generates the sets of value presented to the subject uniformly at random (without repetitions) from the domain of values set in the parameter page. While such a random generation yields various advantages, it has a major drawback concerning the statistical analysis of the results, as some sets of value might be under-represented. An unbalanced representation of each possible set of values is guaranteed only on average and for a large number of exercises; whereas Al Aïn et al. [1]'s protocol, using a systematic enumeration of the possible sets of values (presented in a random order to the subject), does not yield such issues. Such issue was deliberately ignored in order to develop a solution able to measure discrimination abilities on values taken from large domains, and presenting the subject with a systematic enumeration of the possible sets of values is practical only for small domains (e.g. values from 1 to 5). For a domain of size 5 (as that of Al Aïn et al. [1]), enough data points were generated that no pair was under represented (see Table 4).

6.2.3 Extension to sensory diverse species. The colors displayed by digital displays and the sound frequencies played by devices are optimized for the majority of humans. It is not always clear how much and which colours and sound can be seen and heard by individual of each species, an issue absent from purely physical experimental set-ups such as that of [1]. The web application presents extensive parameters to vary the colours displayed and the sounds played to the subject. Even less intuitively, species can differ in their Critical Flicker Fusion Frequency (CFFF) [13], the frequency at which they perceive the world and can react to it (in some species, such frequency even vary depending on the time of the day or of the season [9, 18]). For instance, dogs have higher CFFF while cats have lower ones, and the CFFF of reptiles vary with the ambient temperature. Such variation might affect not only their ability to comprehend the visual display and sound play from devices, but might also affect how they comprehend some application designs over others. The web application presents extensive parameters to vary the time between each exercise and which game, so that part of the rhythm of the application can be adjusted by the experimenter to the CFFF of the subject, but more research is required in order to automatically adapt the rhythm of such applications to the CFFF of individuals from a variety of species.

6.3 Perspective on future work

6.3.1 Random Dice and Heap representations. The discrete representation modes `Dice` and `Heap` associate each value with a fixed representation of a number of points corresponding to the value being represented. This differs from what happens in Al Aïn et al. [1]'s experimental protocol, where the seeds are in no arranged configuration on the cardboard. This might affect the results of the experience in that a subject could learn to select a particular symbol (e.g. the one corresponding to the largest value of the domain) anytime it is present, without any need for any comparison between the presented values. The development and evaluation of their impact on the discrimination abilities of human and nonhuman subjects will be the topic of a future study, once the corresponding randomized representations have been added to the web application.

6.3.2 Systematic logs. The easiness with which logs are generated tends to make one forget about it, to the point that the bottleneck could become the transfer of the logs from the device used to perform the experience to a central repository. As one guardian might get more excited to transfer the logs of sessions where the subjects excelled at the activities than that of less positive sessions, this might create a bias toward positive results in their report. While not an issue while implemented by personal with a scientific training, such risk of a bias might become more problematic in the context of a citizen science project [2]. The development of a website serving as a central repository of experimental data sent by web applications such as the one presented in this work will be the topic of a future study. The roles of such a central "back-end" website could include the automatizing of the most frequent statistical tests on the data received; a greater ease of separation between the roles of experimenter and researcher, which will be an important step toward a true citizen science generalisation of this project (see Section 6.3.4 for a short discussion about the challenges of such projects); and the aggregation of sensory and cognitive data from distinct applications, individuals and species.

6.3.3 Adaptive Difficulty. The great amount of parameters available in the settings page of the web application makes it possible to adapt the difficulty of the activities to the level of abilities of the subject. Such abilities evolve with time, most often advancing and only rarely receding (such as after a long period without using the web application). Choosing which values of the parameters is the most adequate to the current level of abilities of the subject requires an extensive understanding of the mechanisms of the application. An extension of the web application presenting the subject with a sequence of parametrization of increasing difficulty, along with a mechanism raising or lowering the difficulty of the activities presented to the subject would greatly simplify the task of the experimenter, and will be the topic of a future study.

6.3.4 Citizen Science Extensions. The term "*Citizen Science*" refers to scientific projects conducted, in whole or in part, by amateur (or nonprofessional) scientists [8]. It is sometimes described as "public participation in scientific research", with the dual objectives to improve the scientific community's capacity, as well as improving the public's understanding of science and conscience about the research's themes [2]. Citizen Science has become a means of encouraging curiosity and greater understanding of science whilst providing an unprecedented engagement between professional scientists and the general public.

Such methodology must be used with care, in particular about the validity of volunteer generated data. Projects using complex research methods or requiring a lot of repetitive work may not be suitable for volunteers, and the lack of proper training in research and monitoring protocols in participants might introduce bias into the data [20]. Nevertheless, in many cases the low cost per observation can compensate for the lack of accuracy of the resulting data [7], especially if using proper data processing methods [11].

Scientific researchers in comparative psychology could definitely benefit from some help, with many cognitive aspects to explores for so many species. In the process of defining the *anecdotal method* of investigation for creative and cognitive processes, Bates and Byrne [3] mentioned that "*collation of records of rare events into data-sets can illustrate much about animal behaviour and cognition*". Now that the technology is ready to analyze extremely large data-sets, what is lacking in comparative psychology are the means to gather such large data-sets.

Delegating part of the experimental process to citizens without proper scientific training is not without risk. Given the conflicted history of Comparative Psychology [15] in general and Animal Language Studies [14] in particular, the challenge of avoiding "Clever Hans" biases and related ones will be of tremendous importance. We hope that applications and experimental protocols such as the one described in this work could help to design citizen science projects for the study of sensory and cognitive abilities in nonhumans species living in close contact with humans.

REFERENCES

[1] Syrina Al Aïn, Nicolas Giret, Marion Grand, Michel Kreutzer, and Dalila Bovet. 2008. The discrimination of discrete and continuous amounts in African grey parrots (Psittacus erithacus). *Animal cognition* 12 (09 2008), 145–54. https://doi.org/10.1007/s10071-008-0178-8

[2] Citizen Science Association. 2021. CitizenScience.org. Website https://citizenscience.org/. Last accessed on [2022-05-27 Fri].

[3] Lucy Bates and Richard Byrne. 2007. Creative or created: Using anecdotes to investigate animal cognition. *Methods (San Diego, Calif.)* 42 (06 2007), 12–21. https://doi.org/10.1016/j.ymeth.2006.11.006

[4] Simon Coghlan, Sarah Webber, and Marcus Carter. 2021. Improving ethical attitudes to animals with digital technologies: the case of apes and zoos. *Ethics and Information Technology* 23 (12 2021), 1–15. https://doi.org/10.1007/s10676-021-09618-7

[5] Jennifer Cunha and Susan Clubb. 2018. Advancing Communaction with Birds: Can They Learn to Read? https://www.academia.edu/45183882/Advancing_Communication_with_Birds_Can_They_Learn_to_Read

[6] Crystal Egelkamp and Stephen Ross. 2018. A review of zoo-based cognitive research using touchscreen interfaces. *Zoo Biology* 38 (11 2018), 220–235. https://doi.org/10.1002/zoo.21458

[7] Mary M Gardiner, Leslie L Allee, Peter MJ Brown, John E Losey, Helen E Roy, and Rebecca Rice Smyth. 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. *Frontiers in Ecology and the Environment* 10, 9 (2012), 471–476. https://doi.org/10.1890/110185 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/110185

[8] Trisha Gura. 2013. Citizen science: Amateur experts. *Nature volume* 496 (2013), 259–261.

[9] Kevin Healy, Luke McNally, Graeme D. Ruxton, Natalie Cooper, and Andrew L. Jackson. 2013. Metabolic rate and body size are linked with perception of temporal information. *Animal Behaviour* 86, 4 (2013), 685–696. DOI: 10.1016/j.anbehav.2013.06.018.

[10] B. Kohn. 1994. Zoo animal Welfaire. *Rev Sci Tech* 13, 1 (1994), 233–45. doi: 10.20506/rst.13.1.764.

[11] Eva C. McClure, Michael Sievers, Christopher J. Brown, Christina A. Buelow, Ellen M. Ditria, Matthew A. Hayes, Ryan M. Pearson, Vivitskaia J.D. Tulloch, Richard K.F. Unsworth, and Rod M. Connolly. 2020. Artificial Intelligence Meets Citizen Science to Supercharge Ecological Monitoring. *Patterns* 1, 7 (09 Oct 2020). https://doi.org/10.1016/j.patter.2020.100109

[12] Wormald R. Morris D, Fraser S. 2007. Masking is better than blinding. *BMJ : British Medical Journal* 334, 7597 (Apr 2007).

[13] Mankowska ND, Marcinkowska AB, Waskow M, Sharma RI, Kot J, and Winklewski PJ. 2021. Critical Flicker Fusion Frequency: A Narrative Review. *Medicina* 57, 10 (Oct 2021), 1096.

[14] Irene Pepperberg. 2016. Animal language studies: What happened? *Psychonomic Bulletin & Review* 24 (07 2016). https://doi.org/10.3758/s13423-016-1101-y

[15] Irene Pepperberg. 2020. The Comparative Psychology of Intelligence: Some Thirty Years Later. *Frontiers in Psychology* 11 (05 2020). https://doi.org/10.3389/fpsyg.2020.00973

[16] Irene Maxine Pepperberg. 1999. *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots.* Harvard University Press, Cambridge, Massachusetts and London, England.

[17] Bonnie M. Perdue, Andrea W. Clay, Diann E. Gaalema, Terry L. Maple, and Tara S. Stoinski. 2012. Technology at the Zoo: The Influence of a Touchscreen Computer on Orangutans and Zoo Visitors. *Zoo Biology* 31, 1 (2012), 27–39. https://doi.org/10.1002/zoo.20378

[18] E. Reas. 2014. Small Animals Live in a Slow-Motion World. *Scientific American Mind* 25, 4 (2014).

[19] W. K. Richardson, D. A. Washburn, W. D. Hopkins, E. S. Savage-Rumbaugh, and D. M. Rumbaugh. 1990. The NASA/LRC computerized test system. *Behavior Research Methods, Instruments, and Computers* 22 (1990), 127–131. https://doi.org/10.3758/BF03203132.

[20] Brett Thelen and Rachel Thiet. 2008. Cultivating connection: Incorporating meaningful citizen science into Cape Cod National Seashore's estuarine research and monitoring programs. *Park Science* 25 (06 2008).

[21] Michael Trestman. 2015. Clever Hans, Alex the Parrot, and Kanzi: What can Exceptional Animal Learning Teach us About Human Cognitive Evolution? *Biological Theory* 10 (03 2015). https://doi.org/10.1007/s13752-014-0199-2

[22] David Washburn. 2015. The Four Cs of Psychological Wellbeing: Lessons from Three Decades of Computer-based Environmental Enrichment. *Animal Behavior and Cognition* 2 (08 2015), 218–232. https://doi.org/10.12966/abc.08.02.2015