# Anecdotal Experiments: evaluating evidence with few animals

**Mike Dacey**

Comparative psychology came into its own as a science of animal minds, so a standard story goes, when it abandoned anecdotes in favor of experimental methods. However, pragmatic constraints significantly limit the number of individual animals included in laboratories experiments. Studies are often published with sample sizes in the single digits, and sometimes samples of one animal. With such small samples, comparative psychology has arguably not actually moved on from its anecdotal roots. Replication failures in other branches of psychology have received substantial attention, but have only recently been addressed in comparative psychology, and have not received serious attention in the attending philosophical literature. I focus on the question of how to interpret findings from experiments with small samples, and whether they can be generalized to other members of the tested species. As a first step, I argue that we should view studies with extreme small sample sizes as *anecdotal experiments*, lying somewhere between traditional experiments and traditional anecdotes in evidential weight and generalizability.

## 1. Animal Anecdotes and the Founding of Comparative Psychology

Darwin's views on evolution suggest that continuity across species is the rule. Evolution occurs when small changes build up slowly over long periods of time, so we should expect to see cross-species continuity in most traits. Nowhere was this result more significant than when it came to the mind. The fiercely-held conventional wisdom at the time was that human minds were entirely unlike animal minds. To challenge this conventional wisdom, Darwin reports anecdotes about various clever and heroic animals. For instance:

> "I will give only one other instance of sympathetic and heroic conduct in a little American monkey. Several years ago a keeper at the Zoological Gardens, showed me some deep and scarcely healed wounds on the nape of his neck, inflicted on him while kneeling on the floor by a fierce baboon. The little American monkey, who was a warm friend of this keeper, lived in the same large compartment, and was dreadfully afraid of the big baboon. Nevertheless, as soon as he saw his friend the keeper in peril, he rushed to the rescue . . ." (1871 pg. 75)

This anecdotal approach continued in the work of George Romanes, Darwin's appointed successor on psychological topics. Describing similar animal heroism, Romanes says (also reporting the story secondhand) that a column of ants "rushed to the rescue" of an individual pinned with a rock, and "This

observation seems unequivocal as proving fellow-feeling and sympathy, so far as we can trace any analogy between the emotions of the higher animals and those of insects" (1888 pp. 48-49).

Near the turn of the 20th century, authors such as C. Lloyd Morgan (1894) and Edward Thorndike (1911) vocally disproved of the reliance on anecdotes. To be a science on firm founding, they felt, the field would need to shift to rigorous experimental methods. The resulting shift, so a common story goes, brought comparative psychology into its own as a rigorous science (e.g. Shettleworth 2012).

It is easy to see what is objectionable about the way Darwin and Romanes use anecdotes. They relay the stories secondhand without scrutiny, and leap to a heroic interpretation without considering other explanations. There is also a particular worry that work on animal minds will be systematically biased by the unconscious human tendency to anthropomorphize; to interpret animal actions in the same ways they would interpret human actions (e.g. Dacey 2017). Narrative anecdotes seem particularly ripe for such a bias. They often presume intentions behind the action (as when we describe a reach *for* an object, or a glance *towards* a person), and often elicit emotional reactions and bonds with characters that may threaten impartial scientific analysis.

To put it simply, rejecting anecdotes makes comparative psychology look more like other successful sciences (e.g. Thorndike 1911). Scientists across fields shun anecdotes. There are many reasons to do so. I attempt to summarize the key concerns about anecdotes below, listed to aid later discussion. These concerns overlap, and are not exhaustive:

1. Anecdotes can be cherry-picked to make a predetermined point.
2. We lack control over and knowledge of background conditions of anecdotes.
3. Anecdotes are narrative in structure, rather than providing analyzable data.
4. Anecdotes are non-repeatable (non-replicable), and so can't be confirmed independently.
5. Anecdotes don't support generalization.

Performing controlled experiments can alleviate these concerns. One cannot pick and choose which individual responses in any given experiment to report (though one can choose which experiments to report, as discussed below). A good experiment is defined by control over the variables that might influence behavior. Experiments produce evidence in the form of data, which is cold, dispassionate, and suited for statistical analysis. As a result, when done well, experiments are replicable (worries noted in section 3), and they can support generalization.

Summing up, anecdotes are usually opposed to experiments. A common foundation story for comparative psychology tells that it came into its own as a science when it chose experiments over anecdotes. However, it is not clear whether this foundation story holds up when we look at current practice.

## 2. Sample Sizes in Animal Labs

When running laboratory experiments on animals, practical constraints significantly restrict sample sizes. Animals must be kept and cared for, and labs can only afford and fit a certain number. Ethical concerns often dictate that the number of animals involved should be as low as possible.[1] Individual experiments usually require time-consuming training, so some subset of the overall groups is chosen.[2] There are also often basic tasks that an animal must successfully perform to even participate in the experiment, and those of the original group chosen who fail will be excluded. I take these to be challenges intrinsic to the subject of study, and do not intend to criticize the researchers who face them. Nonetheless, the implications are stark. Experiments frequently include samples of individual animals in the single digits, and sometimes only 1 or 2. Figure 1 shows the number of individual animals included in every individual experiment published in four top journals in the field in 2019. Out of 151 experiments in 90 papers, 50 experiments include data from 10 or fewer animals (nearly 1/3 of the total), and 98 include

[1] Both of these issues are especially difficult with primates, and even more so with chimpanzees, as in my example below.

[2] Additionally, having been trained on one task may influence later performance on other experiments, so sometimes animals are excluded so that they remain 'naïve' to the tasks at hand.

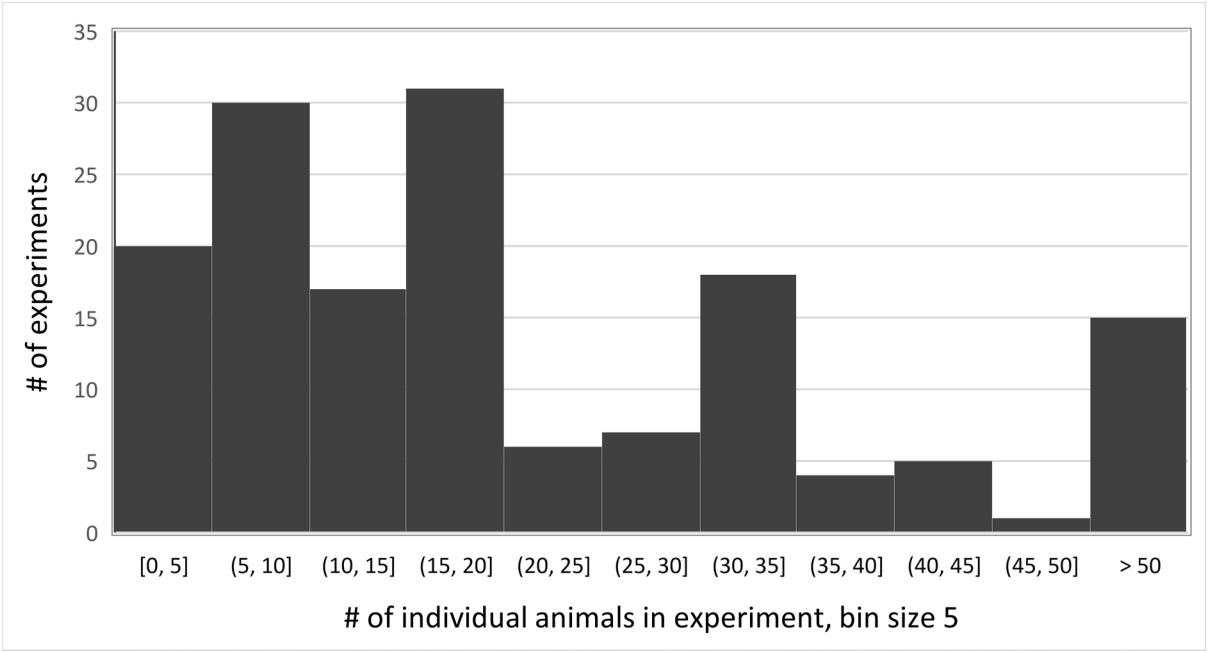


*Fig. 1*: A histogram of all experiments published in the journals *Animal Behavior and Cognition*, *Animal Cognition, Journal of Comparative Psychology,* and *Journal of Experimental Psychology: Animal Learning and Cognition* in the year 2019, sorted by number of individual animals in reported data. This includes 151 experiments described in 90 papers; any experiment that made an intervention, laboratory or field. 54 papers were excluded, as they did not present new behavioral data (29 papers), were unable to report the number of animals involved (seven papers), were purely observational field studies (six papers), or used only human subjects (12 papers).

data from fewer than 20 (nearly 2/3).[3] To put it bluntly, these sample sizes would be unacceptable in other branches of psychology.

As an illustrative example of the interpretive challenges raised by sample sizes like these, I will focus on Inoue & Matsuzawa's 2007 paper, "Working memory of numerals in chimpanzees." This paper compares human and chimpanzee performance on a short-term memory task. The authors state their conclusions unequivocally: "Our study shows that young chimpanzees have an extraordinary working memory capability for numerical recollection better than that of human adults" (pg. 1005). The paper has

[3] Thanks to Abraham Brownell for performing this analysis. This data is not meant to present a statistically rigorous picture of the field at large, but simply to provide a reasonably representative snapshot. This illustrates the issue to those unfamiliar with the norms of the field. These journals are among the top that focus on animal cognition, and were chosen in large part to limit potentially subjective inclusion criteria. However, they are not the only such journals, and animal cognition studies are often published in more generalist journals as well (for instance, the example discussed below was published in *Current Biology*). Several of these experiments also divided participants into different conditions, further limiting the number of individuals observed making specific responses, though we did not analyse these divisions.

been cited extensively, and in the media, this conclusion was accepted uncritically ("Chimps Exhibit Superior Memory, Outshining Humans," *New York Times* 12/4/2007).

The task was as follows. Participants (human and chimpanzee alike) sit in front of a computer screen. The computer quickly flashes several digits in random locations on the screen (all shown simultaneously). After a presentation of a few hundred milliseconds (650, 430, and 210 *ms* in different trials), each digit is masked with a small white square. Participants were asked to then tap each masking square in order of the digits previously at each location. The researchers measured both response times and accuracy. The task is meant to test the ability to rapidly store working memories for the visual scene (210 *ms* is too fast to saccade through the sequence).

Inoue and Matsuzawa begin the study with 6 chimpanzees (three mother-child pairs; there were 14 total on-site). While all six were able to learn the basic masking task, only four performed at the level of five numerals, which was the number used in the key test (Supplemental materials Table S1). So, the experiments include these four animals. The actual data presented, however, only compares one chimpanzee at a time against a human average (human n=9 in one experiment, n=12 in another). So for each actual comparison, chimpanzee *n*=1. In fact, the assertion that chimpanzees perform better than humans seems to be based on a single chimpanzee, Ayumu, the best chimpanzee performer (see figure 2). Based on the data presented in supplemental material (see figure 3), Ayumu matched the human average accuracy rate with 650 ms presentation times, but still had a lower accuracy rate than the majority of the individual humans.[4] So they key claim here seems to be based on a simple size of one.

Given this reliance on extremely small sample sizes, we must question whether the field has really moved on from its anecdotal roots. I suggest that performance of animals like Ayumu is just another kind of anecdote; it's a single animal (or very small number) displaying an interesting behavior. It can be hard

[4] All three chimpanzees shown did show faster response times than all humans (response time was measured as the latency before the first number was touched).

to know exactly what conclusions we can draw from a study like this. At the very least, though. findings like this cannot ground general claims like "chimpanzees outperform humans."[5]

This study is a particularly salient example, both in the sense that it reaches the limit case of *n*=1, and in its strong conclusion and broad uptake. But the core concerns here generalize, given the number of experiments published with extremely small sample sizes. To be clear, these restrictions result from practical issues intrinsic to the field. I do not criticize researchers for this, as I see no reasonable way around it (absent massive funding increases and means to address ethical concerns).

## 3. The replication crisis and comparative psychology

In recent years, other branches of psychology have instituted reforms to address prominent and repeated replication failures (Romero 2019). Despite the obvious worry that small sample sizes leave comparative psychology vulnerable to these same problems, the field has only just begun to respond (Beran 2018, Farrar, Boekle, & Clayton 2020). Stevens (2017) notes that comparative psychology makes frequent use of within-subjects methods[6] that might protect the field compared to social psychology.

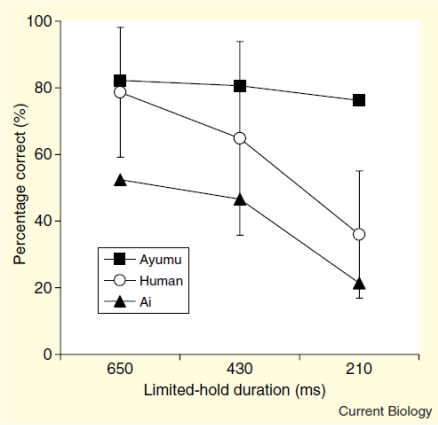


*Fig. 2:* Data from Inoue & Matsuzawa 2007 (pg. R1005). Two chimpanzees, Ayumu and Ai are compared to a human average.

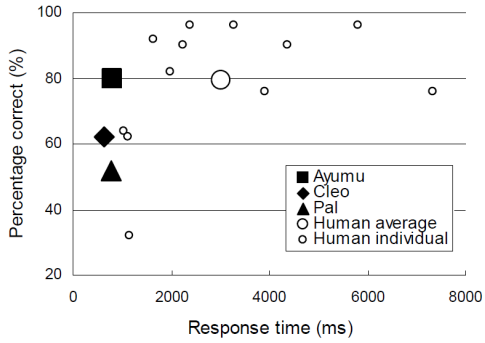


*Figure 3:* Data from Inoue & Matsuzawa 2007, with 650 *ms* presentation of stimuli before mask (supplemental materials, figure S2). This is the same condition as the leftmost data-points of figure 2, this page.

[5] This is compounded by the fact that Ayumu here is an outlier among even the top performers: only those individuals able to perform the basic task were included, and Ayumu's performance was an outlier among them. There are also concerns that the life-history of laboratory animals makes them unrepresentative.

[6] I note that within-subjects statistical analyses may be more likely replicate even with few individuals, but those methods do not help the problem of generalizing findings to other members of the species.

However, he says, there are several reasons to think that comparative psychology is vulnerable to replication failures. He makes several recommendations for the field to address these concerns. Some of these recommendations have also begun to be implemented. I will focus here on recommendations that inform the current discussion.

One such recommendation is for researchers to pre-register their methods before the test, or for journals to adopt the practice of registered reports, in which a journal accepts or rejects a paper based on methods alone, before experiments are run. This practice has grown in fields like social psychology. The purpose is to prevent fishing-expedition approaches to studies and statistical analyses: These can lead to cherry-picking which studies are reported, and P-hacking by, for instance, simply trying various statistical analyses until one gets a significant result. In 2018, the journal *Animal Behavior and Cognition* began accepting registered reports (Vonk & Kraus 2018), though the editors report that uptake by researchers has been slow (Beran 2020).

Worries about sample size are more complicated. For instance, social psychology has massively increased sample sizes in their studies simply by making greater use of online platforms like Mechanical Turk and Qualtrics. Comparative psychology has no such option. And indeed, for reasons noted above, it seems impossible to completely avoid small sample sizes. Nonetheless, Stevens does make some recommendations that can help. First, different labs can collaborate and combine their subject pool. In fact, the ManyPrimates Project was launched in 2019 to facilitate collaboration across labs spanning the globe, allowing for larger and more diverse samples in studies of primate cognition (Many Primates et al. 2019). Secondly, he suggests that researchers can take advantage of facilities like zoos that may have larger numbers of animals available. Thirdly, researchers can reconsider their choice of species, either by running studies pooling multiple species, or by switching to species that are easily available in the community, such as dogs.

I have little to add on recommendations regarding species choice, but I will take on-board the rest of the recommendations I've mentioned. While the recommendations aimed at increasing sample size are

unlikely to completely address the problem (they simply cannot have an impact like we've seen in social psychology), they certainly help. Registered reports are also valuable; if papers are evaluated based on methods rather than results, it will significantly impact our interpretation of studies will small sample sizes in ways I discuss below (section 6).

Even large-scale changes are not likely to completely address sample size worries in comparative psychology. But even if they do in the future, we should still consider how to interpret existing small sample studies. Either way, interpretive challenges remain. To face these challenges, we can start by look to other research programs that employ very small samples, or even samples of one. To the extent these programs are analogous to comparative psychology, they might provide concrete suggestions.

## 4. Candidate Analogue One: Cognitive Neuroscience

Lesion studies in cognitive neuroscience present the first candidate analogue. In many of these studies, researchers test a single patient with known brain damage on a battery of tasks aimed at delimiting a certain cognitive capacity.[7] Studies like this generally focus on two kinds of question. The first are questions about the neural underpinnings of a particular cognitive capacity. Here, the goal is locating damage, and correlating it with deficits. The second are the so-called dissociations of capacities that might otherwise be thought to be expressions of a single system. For instance, if a deficit in experiential memory does not also bring with it a deficit in memories for facts, then we have reason to believe that the two are separate capacities subserved be separate systems (episodic and semantic memory), and moreover, the intact capacity does not require the damaged capacity.

The evidential value of lesion studies has long been controversial. As a result, there is a substantial literature aimed at uncovering the methodological assumptions behind the research (e.g. Caramazza 1986, Bub & Bub 1988, McClosky & Caramazza 1988, Glymour 1994, Shallice 2015). The actual damage and

[7] As in the Matsuzawa study, these individuals are also outliers; they are chosen precisely because their performance is abnormal.

deficits observed in individuals vary substantially, and the 'clean' cases of a particular deficit are rare. As a result, it can be difficult to know what aspects of any study can be generalized. Arguably, these concerns, along with improvements in other methods, have driven a reduction in reliance on lesion studies in recent decades. However, if one *is* dealing with lesion studies, the focus on a specific individual is arguably (but controversially) an advantage. The very fact that individual deficits vary so much means that effects would likely wash out in any cohort study, leaving them impossible to interpret (Caramazza 1986).

Even so, there is at least one kind of general claim that these studies do seem to license. These are claims about the *necessity* of one capacity for another, as made in dissociation studies. If Task A can be performed by an individual who cannot perform Task B, then it cannot be the case that the capacity responsible for performance of Task A is necessary for performance on Task B.[8] This inference can be transferred. For instance, the fact that Ayumu was able to do so well on the memory task without using language suggests that language is not required. Necessity claims are strong claims though, especially for a field like psychology, where pretty much everything can vary across individuals. So, the denial of a necessity claim may not always be hugely informative. Nonetheless, even if this is a limited result, it's something.

## 5. Candidate Analogue Two: Anecdotes in Cognitive Ethology

Researchers in cognitive ethology will also sometimes report anecdotes, or "incident reports" of particular observed behaviors. As with lesion studies, this practice is controversial (Mitchell, Thompson, & Miles 1997). In general, data based on repeated observation is preferred, if possible. Even so, incident reports reports may describe low-frequency behaviors, that would be difficult to observe frequently or to elicit in a laboratory setting. They can also introduce behaviors that researchers had been wholly unaware of. Field anecdotes can also arguably provide some evidence about cognitive processes on their own: field

[8] This basic inference structure is also employed in developmental psychology, though with larger sample sizes (Perner & Lang 1999).

observations don't face any concerns about ecological validity, and anecdotes can often supply richer context about the individual behaving and its context than experiment (Mitchell 1997).[9]

Nonetheless, incident reports do suffer from the limitations described above, with concerns about anthropomorphism and generalizability at the fore. Indeed, the use of anecdotes has been declining in primatology (Ramsay & Teichroeb 2019), suggesting that the downside of anecdotes is winning out in the minds of researchers. Even if these anecdotes do not provide much evidential value, they have heuristic value in generating hypotheses, guiding future observation or experimentation, and identifying behaviors worthy of more systematic study (Silverman 1997¸ Andrews 2020).

## 6. **Anecdotal Experiments**

As a start towards coming to grips with the sample size problem in comparative psychology, I argue that we should view studies with extreme small samples sizes as *anecdotal experiments*. Anecdotal experiments have some of the strengths that are usually ascribed to well-designed experiments (they are controlled and meticulously recorded), and some of the weaknesses ascribed to standard anecdotes (they may not be reliably repeatable, and they do not support straightforward generalization to other individuals). They occupy a middle-ground, providing stronger evidence than that provided by a one-off observation, but not as strong as that provided by experiments with larger sample sizes.

To illustrate more specifically, I return to the concerns lodged against anecdotes in section 1. Anecdotal experiments avoid the most significant concerns, while the rest could be lodged against these studies anyway. I'll work through each in turn.

**Concern 1:** Anecdotes can be cherry-picked to make a predetermined point.

This worry can be avoided by making use of registered reports, such that papers are accepted based on methods, before experiments are done. It remains a worry that existing studies report cherry-picked

[9] Mitchell advocates specifically for anthropomorphic anecdotes as a way to conceptualize behavior. I set the issue of anthropomorphism aside for now, as I see it as less of a concern here (see next section).

experiments, though perhaps not to the degree of full anecdotes: the number of individual behaviors one might observe and dismiss in reporting an anecdote is much less than the number of experiments one might perform and dismiss.

**Concern 2:** We lack control over and knowledge of background conditions of anecdotes.

This worry does not apply here to any greater degree than it does in psychology generally. A well-designed experiment controls immediate background conditions, such that we can have a reasonable idea of what features of the task the animal is responding to.

**Concern 3:** Anecdotes are non-repeatable (non-replicable), and so can't be confirmed independently.

Anecdotal experiments have records of methods, which make replication possible. However, replication problems in other areas suggest that comparative psychology should be concerned about replicability (Farrar, Boekle, & Clayton 2020). Perhaps the focus on within-subject tests puts comparative psychology in somewhat better position than it might be otherwise (Stevens 2017), but the extremely small sample sizes suggest that replicability cannot be assumed. This is a worry either way, and framing these as the anecdotal experiments can make it more explicit.

**Concern 4:** Anecdotes are narrative in structure, rather than providing analyzable data.

Anecdotal experiments do rely on data, so seem to pass this test. Nonetheless, we should be careful in what we take that data to show. If, as just suggested, we should question the replicability of these studies, statistics can mislead. A careful reevaluation of statistical measures can help here (as in social psychology). However, absent that, statistics can present a false sense of generalizability. For instance, we can statistically show that Ayumu himself reliably outperforms the human sample average in this study. What that means about chimpanzees more generally is a different question.

**Concern 5:** They don't support generalization.

As with concern 3, this is just to recognize limits already present. It is common to restate an experimental finding by simply plugging generics into a literal description of the study. For instance: "Ayumu outperformed the average performance of twelve humans in our study" becomes "chimpanzees outperform humans." This move is clearly too quick. If we have good reason to believe that a study includes a representative sample of a larger population, we generalize to that population. These generalizations should become more tentative as confidence in the representativeness of the sample increases. With very few animals, we can't generalize this way. This is compounded by the fact that the animals performing in these experiments, like Ayumu, are often outliers.

Treating experiments with extremely small sample sizes as anecdotal experiments marks their limitations, and helps guide their proper use. There are many important unanswered questions here. We would want to know how to determine which experiments are anecdotal and which are not; where is the cut-off? Moreover, in light of the interpretive limitations of anecdotal experiments, I have said little about what, concretely, we can learn from them. I will offer some brief comments on that topic here.

The fact that one member of a species is able to perform a task to a certain criterion shows that it is *possible* for some members of that species to do so. However, this doesn't guarantee any particular cognitive mechanism. Though, we can follow work in cognitive neuroscience and conclude that successful performance shows that some capacity believed to be absent (say, language) is not *necessary* for performance on the task. They may also provide some evidence for one hypothesized mechanism over another if that level of performance is impossible or highly implausible according to the devalued hypothesis. Absent such strong claims, one competing hypothesis may still predict better performance on a task (this is not the aim of the Inoue & Matsuzawa study). If so, a convincing finding of strong performance might provide a small (minute, even) amount of evidence for that hypothesis. Additionally, following cognitive ethology, the fact that at least one individual succeeds in a task might motivate new hypotheses about the cognitive capacities involved, or identify new areas worthy of further study. These are useful conclusions, but they are not often deeply helpful in evaluating models of the actual cognitive

processes involved. Psychological models rarely make claims of possibility or impossibility, and one cannot conclude a capacity is not necessary for the task unless one is confident the animal does not possess that capacity (most of the interesting options are still up in the air).

Even with these limitations in scope and strength, any generalization from an extremely small sample to a species at large must be significantly hedged: these individuals might just be doing something completely different than other members of the species.[10] But even so limited, there is still value to that evidence. Often when it comes to nonhuman minds, strong evidence is very hard to come by, so any amount of evidence is worth considering.

## 7. Implications and Conclusion

The basic point of framing extreme small sample studies as anecdotal experiments is to reduce their weight in general claims about the nature of nonhuman cognitive capacities. Indeed, I argue that the field ought to reduce the evidential weight of individual experiments in general, to help move away from a pernicious 'critical experiment' framing that still too often pervades. The actual evidential value of individual experiments must be assessed on a case by case basis, depending on the kind of model being evaluated, and the nature of the anecdotal experiment. This is tough work of course, but it always has been.

There may be other general impacts on the field. This framing could benefit the field by encouraging more exploratory research and reporting of more varied behaviors. In effect, experimental comparative psychology might look a bit more like field ethology. For example, Stanton et al. (2017) presented raccoons with the Aesop's fable task, in which they can gain access to a treat floating on water by dropping stones in to raise the water level. They report that one of the raccoons managed to get the treat, not by dropping stones, but by ripping the entire apparatus off the floor and dumping it out. A field that

[10] I have ignored worries about ecological validity and differences between captive and wild animals, but they would have to be considered in addressing this possibility.

relies on registered reports, and recognizes the limitations of data from such small sample sizes would likely include substantially more reports of behavior like this. There is value to that, as these behaviors, intended by the experimenter or not, do provide insight into the animals.

Most importantly, though, this framing encourages more honest reporting of the significance of studies. Extreme low sample size studies are limited in evidential value. Reporting them as anecdotal experiments presents them as such.

**References**

Andrews, K. (2020). *How to Study Animal Minds*. Cambridge: Cambridge University Press.

Beran, M. (2018). Replication and Pre-Registration in Comparative Psychology. *International Journal of Comparative Psychology*, *31*.

Beran, M. (2020). Editorial: The Value and Status of Replications in Animal Behavior and Cognition Research. *Animal Behavior and Cognition* 7(1): i-iii.

Bub, J. and Bub, D. (1988): On the Methodology of Single-case Studies in Cognitive Neuropsychology, *Cognitive Neuropsychology*, 5, 563-582.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, 5, 41–66.

Dacey, M. (2017). Anthropomorphism as cognitive bias. *Philosophy of Science*, *84*(5), 1152-1164.

Farrar, B. G., Boeckle, M., & Clayton, N. S. (2020). Replications in comparative cognition: What should we expect and how can we improve? *Animal Behavior and Cognition*, 7(1), 1-22. doi: https://doi.org/10.26451/abc.07.01.02.2020

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Current Biology*, *17*(23), R1004-R1005.

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., ... & Flessert, M. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PloS one*, *14*(10).

McCloskey, M., & Caramazza, A. (1988). Theory and methodology in cognitive neuropsychology: A response to our critics. Cognitive Neuropsychology, 5, 583–623.

Mitchell, R. W. (1997) Anthropomorphic Anecdotalism as Method, in Mitchell, R. W., Thompson, N. S. & Miles, H. L. (Eds.). Anthropomorphism, Anecdotes, and Animals. Albany: State University of New York Press pp 151-169.

Mitchell, R. W., Thompson, N. S. & Miles, H. L. (1997). Anthropomorphism, Anecdotes, and Animals. Albany: State University of New York Press.

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in cognitive sciences*, *3*(9), 337-344.

Ramsay, M. S., & Teichroeb, J. A. (2019). Anecdotes in Primatology: Temporal Trends, Anthropocentrism, and Hierarchies of Knowledge. *American Anthropologist*, *121*(3), 680-693.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, *14*(11), e12633.

Shallice, T. (2015). Cognitive neuropsychology and its vicissitudes: The fate of Caramazza's axioms. *Cognitive neuropsychology*, *32*(7-8), 385-411.

Shettleworth, S. (2012). Fundamentals of Comparative Cognition. Oxford: Oxford University Press.

Silverman, P. S. (1997). A Pragmatic Approach to the Inference of Animal Minds, in Mitchell, R. W., Thompson, N. S. & Miles, H. L. (Eds.). Anthropomorphism, Anecdotes, and Animals. Albany: State University of New York Press pp 170-185.

Stanton, L., Davis, E., Johnson, S., Gilbert, A., & Benson-Amram, S. (2017). Adaptation of the Aesop's Fable paradigm for use with raccoons (Procyon lotor): considerations for future application in non-avian and non-primate species. *Animal cognition*, *20*(6), 1147-1152.

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in psychology*, *8*, 862.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. New York: The MacMillan Company

Vonk, J., & Krause, M. A. (2018). Editorial: Announcing preregistered reports. *Animal Behavior and Cognition*, *5*(2).