Laura Hennefield and Hyesung G. Hwang
with Daniel J. Povinelli

Going Meta: Retelling the Scientific Retelling of Aesop's the Crow and the Pitcher

Abstract: The Crow and the Pitcher, a classic Aesop's fable, has surprisingly (re)captured the interest of comparative cognition scientists in the past decade. These researchers examine whether corvids (e.g., rooks, crows, and jays) can complete a laboratory analog of the fable by training the corvids to drop stones and other similar objects into tubes of water to retrieve floating worms. This Aesop's Fable Paradigm is argued to be an experimental method that can prove corvids have the ability to engage in complex causal reasoning—implying that they understand something fairly rich about the ideas of volume and water displacement. However, critiques—including our own meta-analysis—suggest that corvids' behaviors in this paradigm could be explained by trial-and-error learning combined with an instinctive, initial preference for functional objects rather than complex causal reasoning. With this line of research as the case example, we explore historical and sociocultural factors in the field of psychology that incentivizes scientific research that tells a "good story."

As we sit down to write, we are both postdoctoral research fellows in psychology. More colloquially, we are "postdocs"—members of that swelling army of young PhDs competing for a seemingly shrinking number of tenure-track faculty positions in the sciences. Specifically, we are both developmental psychologists who are building our careers studying the social and cognitive abilities of infants, toddlers, and preschool-aged children.

Journal of Folklore Research, Vol. 56, Nos. 2–3, 2019 • doi:10.2979/jfolkrese.56.2_3.04

In this essay, we are not writing about children *per se*. Instead, we want to provide some insight into our experience with a puzzling development in the closely allied field of animal cognition: the widely celebrated experimental research with crows based on the classic Aesop's fable of the Crow and the Pitcher, in which a thirsty bird uses pebbles to raise the level of water in a vase to get a drink. Let us say at the outset that our experience with these studies does not, as one might expect, concern the nature of children's psychology in the contexts of narrative, or of fables. We are not going to consider questions of human development and narrative comprehension; we are not going to discuss children's understanding of water displacement. Rather, this is our (unexpected) retelling of the scientific retelling of an ancient fictional story.

An Experimental Paradigm Based on a Fable

For more than one hundred years, psychologists who study the cognitive abilities of human children have been intrigued by similar studies involving animals.[1] Even undergraduate students of developmental science cannot escape reading about cognitive studies involving chimpanzees or dolphins or birds. Early and often, developmental psychologists are reminded that the animal-cognition literature is replete with discoveries of cognitive capabilities once thought to be solely present in humans [*Editors' Note:* See Appendix, "Doctor Fomomindo's Preliminary Notes for a Future Index of Anthropomorphized Animal Behaviors."]

As young students (and technically as outsiders to the animal science disciplines), we had always thought that the various claims about animal cognition seemed rather muddled and tricky to interpret. On the surface, the studies seemed to show that other animals are very similar to humans. We learned that tool use and tool making—once considered uniquely human—has been observed in the behavior of many animal species in the wild. This list includes chimpanzees, capuchin monkeys, gorillas, dolphins, sea otters, woodpecker finches, and yes, even some species of crows.[2] But, it was never clear to us whether or not the ethological evidence of tool use proves that when chimpanzees or crows, for example, use sticks to probe for insects or larvae, they understand what they are doing in the ways that human children—not to mention human adults—do. And although we were confident that nonhuman animals communicate (clearly, they do),

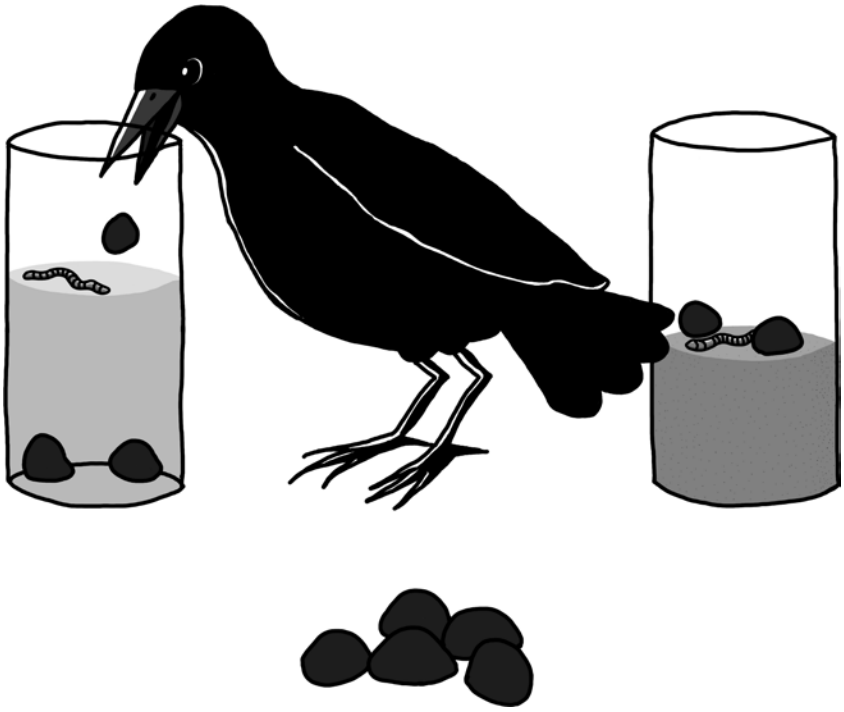

FIGURE 1
Dropping stones into the water-filled tube on the left raises the level of the water and brings the worm closer to the crow; dropping stones into the sand-filled tube on the right does not. Cartoon by Gavin Rackoff.

we were not completely convinced that the waggle-dance of bees has anything to do with the human language's abstract properties, such as recursion or complex hierarchical syntax. Then there were the claims of empathy in rats and numerical reasoning in monkeys, the abilities of orangutans to play games, self-awareness in elephants, and even autobiographical narratives in chimpanzees [*Editors' Note*: see Appendix]. It all seemed simultaneously convincing ("There are so many studies and everyone else seems to be buying into it!") and unconvincing ("There are so many gaps in the experimental logic; how can we look past them?").

Later, when we were both graduate students at Washington University in St. Louis during the Fall of 2014, an expert in animal cognition, Daniel Povinelli, showed up in our department as a visiting professor. We decided to take a class with him to learn more about the field of animal cognition—straight from the horse's mouth, as

they say. The surface structure of the course was familiar enough. Each week, we had to read a gathering of empirical research papers on a particular topic, and we needed to be prepared to discuss and to critique the papers in seminar. Other aspects were less familiar. For one, we were encouraged to perform our own literature searches and to bring to class the best and most compelling research in support of each given topic that we could find. For another, in every class someone was in charge of commenting on how the popular press had reported on the studies we were covering that week. We quickly began to detect certain patterns. Not surprisingly, what we read in the popular press (and in some textbook summaries) did not always match up very well with the details in the actual papers and studies themselves. More interestingly, the press seems to be *inexhaustibly* interested in studies about smart—especially "human-level" smart—animals.

The research directly inspired by the ancient fable of the Crow and the Pitcher immediately raised our suspicions. In these studies, researchers had taught some crows to drop stones into test tubes of water in order to raise the water level high enough to retrieve a floating worm.[3] Some of the crows became so adept that they even learned to avoid dropping stones in test tubes filled with sand (see Figure 1).[4] The researchers claimed that these results show that crows are capable of "complex cognition"—implying that the crows understood something fairly rich about the ideas of volume and water displacement.[5] And it was not just one study. To our surprise, we discovered that over a period of about eight years, five peer-reviewed research articles containing over thirty-two experiments had been inspired by the fable! Each paper focused on a small number of birds and a growing list of slight variants of the task. Time and time again, the researchers concluded that the fable-inspired tasks were somehow special—uniquely suited to reveal the higher-order mental abilities of animals.[6] One research group even claimed that their work showed that crows understand the physics behind the test even better than seven-year-old children.

We were puzzled. How could such a uniquely productive experimental design have been buried in an ancient folkloric narrative? How could crows be outsmarting seven-year-olds? Upon closer reading of the original research, our suspicion and puzzlement quickly turned to doubt: No matter how intelligent crows are, we began to find reasons to think that this fable-induced test was not a good way of measuring it. How could training birds to drop stones into a test

tube (using an Aesopian fable as inspiration) necessarily show complex cognition? What, exactly, do we mean by "complex cognition" in this case? Surprisingly, none of the researchers seemed to tackle these issues head-on. Moreover, when we saw how popular these studies had become in the science news media, we found ourselves asking, "Why does no one else seem to be skeptical?"

Committed to acting as our own skeptics, several aspects of the experimental designs struck us right away. First, the birds that participated in the original study, rooks (part of the crow family), do not naturally use tools. In addition, in the initial "pre-test" phase (before they had to decide whether to drop stones in a test tube filled with water versus one filled with sand), the birds were taught to drop stones into a single, water-filled test tube. In other words, the birds did not encounter the pile of stones an experimenter conspicuously set next to the test tube and spontaneously start dropping them into the test tube. Instead, the crows had to be cajoled to do so: the experimenters had to balance a stone on the lip of the test tube, whereupon the birds would accidentally knock it into the tube and fortuitously see the worm rise a little. Only then did the birds start manipulating stones on their own. This pretraining was a necessary precedent for each of the dozens of variants of the same basic paradigm—having crows drop objects of all sorts into tubes while attempting to systematically vary key aspects of the objects, such as heavier vs. lighter or sinking vs. floating. But again, the amount of training required for the birds to perform even the most basic variant of the task (just dropping stones in a single tube filled with water) made us pause—if crows need extensive training to perform the stone dropping action, how could any subsequent learning "prove" higher-order cognition?

In fact, everything about the test appeared to scream "associative (trial-and-error) learning." Each time a crow drops a stone in the water-filled tube, the worm rises and gets a little closer to the surface where the waiting bird can snatch it. All that the experiments could demonstrate was that the birds could learn to keep repeating the same action over and over until they got their reward.

So, even at first glance, it seemed to us that the birds could just be learning to drop stones the same way a rat might learn to press a blue lever several times instead of a red one—analogous, for example, to a hungry rat placed in one of B. F. Skinner's classic "Skinner boxes." The rat initially wanders around, exploring the box until it bumps into a lever, which releases a food reward. But after several instances

of this accidental behavior, every time the rat is subsequently placed in the box, it heads straight toward the lever and paws at it until the food is released. From there, the rat can learn any contingency the experimenter decides to impose on the situation (e.g., that the lever will only release food after it is pressed three times, or that pressing a blue lever releases food, but pressing a red lever does not). In fact, the predictable ways in which reward and punishment shape this kind of learning is so well established—it dominated American experimental animal psychology for half a century or more—that any assertion of a "new" type of learning or reasoning needs to first account for the roles of these already well-known processes. With these basic learning principles in mind, we became increasingly dubious of what the Aesop's fable-inspired studies could tell us about higher-order cognition. The crows' behavior in these tasks did not seem to be capturing anything like human insight: We were not hearing Archimedes shout "Eureka!" as he leapt from his bathtub and raced naked through the streets of Syracuse. What about these studies would make researchers jump to the conclusion that crows understand that the volume of one set of objects (the stones) "displaces" a comparable (or even any) volume of water?

A second component of these experiments that struck us even more was that most of the time the data from the main tests (for example, the choice between a water-filled versus a sand-filled tube) was judged as an all-or-nothing, either/or set of possibilities, and a given bird either "passed" or "failed" each trial. That is, after a crow had dropped all of the stones, it either succeeded in getting the worm or it did not. Thus, depending on the final outcome of twenty trials, the original researchers concluded that a crow had either "understood" the test or had "not understood" the test. But even when later researchers discussed the results in terms of learning, they focused on how many trials it took the birds to become regularly "successful" in getting the worm. But to us, an obvious fact about each trial was being swept under the rug. After all, each trial consisted of many individual stone drops. And just like the rat pressing levers, each individual stone drop was a learning opportunity: the worm either rises a little (water tube) or it does not (sand tube).

Thus, it was the treatment of the data in the Aesop's fable-inspired experiments that became central to our decision to investigate the data in these experiments using a more fine-grained approach. By analyzing the data at the level of each trial (or group of twenty trials)

and not at the level of each stone drop, the researchers are essentially masking valuable information provided by each discrete data point (i.e., each stone drop). It was curious though. Every article did visually depict the data for the individual stone drops. For example, Figure 2 depicts sample data from a bird named Oliver. The way to understand this table is to see that in this experiment Oliver was given five stones per trial for twenty total trials—twenty different opportunities to try to use a pile of stones to get a worm when presented with the water-filled versus sand-filled tubes. Each trial began when the bird dropped the first stone into a tube, and each trial ended when the bird either 1) was able to retrieve the food from the water-filled tube, 2) exhausted all available objects, or 3) gave up and stopped dropping stones. Reading Figure 2 horizontally, however, you can see that each trial consisted of multiple, discrete acts of stone dropping. Sometimes Oliver dropped the stones into the water-filled tube and sometimes into the sand-filled tube (dark-gray squares for the former and light-gray squares for the latter). In fact, if you have the patience to count them up, you can see that although Oliver was given only twenty trials, he was given one hundred opportunities to learn about the different consequences of dropping stones in the two tubes (he seemed to catch on after about fifty and only dropped seventy-three stones across the twenty trials). Analyses of the results by trial, instead of by individual stone-drop, obscure important clues about how crows initially approached the task and if or how their behavior changed as the task progressed.

The Aesop's fable-inspired researchers claimed that crows demonstrated "complex cognition" in the water versus sand task because crows "rapidly" learned to drop the stones into the water tube. To that end we realized that at least three specific questions could be addressed by a meta-analysis:

1) Did the crows show any preference for the water tube (over the sand tube) at the very beginning of the tests?
2) How quickly did the crows learn to select the water tube over the sand tube (i.e., what exactly does "rapidly" mean)?
3) What was the source of the bird's learning?

The third question was especially intriguing to us: Did the birds learn anything when they dropped stones into the sand-filled test tubes, or did the learning only occur when they dropped stones in

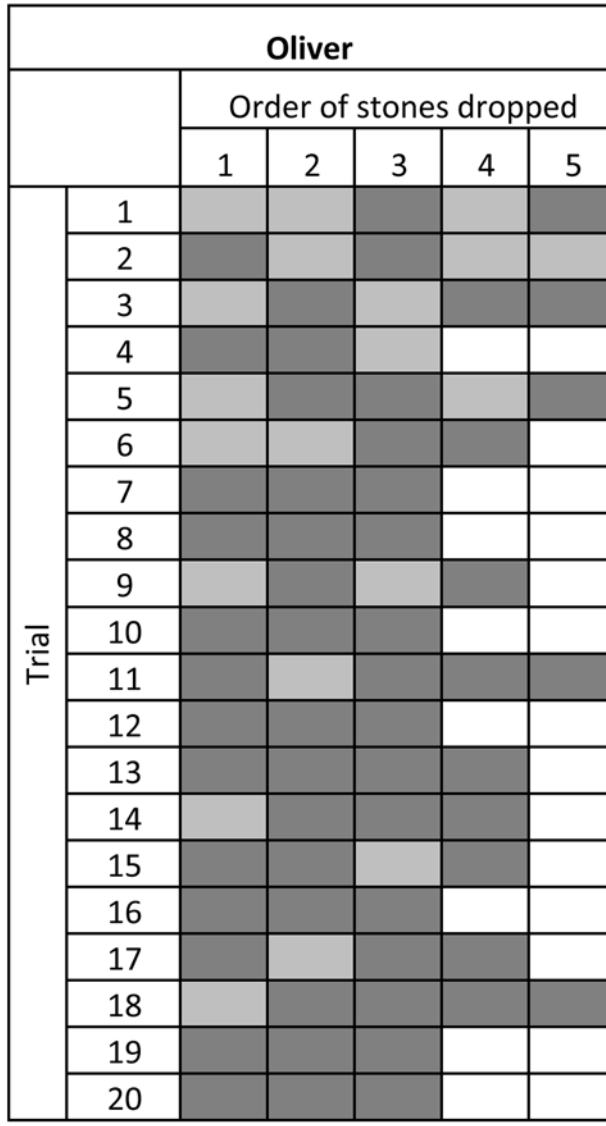

FIGURE 2
An example of how the data was depicted in the published articles. Light gray squares indicate that Oliver dropped a stone into the sand tube; dark gray squares indicate his choice of the water tube. The white squares indicate he did not use the remaining stones. Researchers provided this stone-drop level of data for each bird in each task but did not use it in their analysis. This bird (Oliver), for example, would likely have been described as "successful" despite the fact that he exclusively dropped stones into the incorrect (sand) tube on the first trial and his behavior was essentially random across the first twenty individual stone drops!

the water-filled tubes? The question made us realize that it would be possible to reanalyze the data from each test (e.g., the water vs. sand task) on a drop-by-drop basis within each article and then to combine the data from across multiple articles in the form of a meta-analysis—an analysis in which all the birds could be included to increase the power of the analyses. Because many of the research reports conducted multiple variants of the Aesop's fable task, we were also able to analyze how well the birds transferred what they learned in earlier tasks to later tasks. Below we discuss this further.

The Work of the Meta-Analysis

The first Aesop's fable-inspired study was published in a journal called *Current Biology*—a prominent and well-respected, peer-reviewed journal with a reasonably high impact factor.[7] The majority of the subsequent replications and variants of the paradigm, conducted by researchers across several well-established laboratories, were published in journals with lower impact factors, journals that were nonetheless well-respected and peer-reviewed (e.g., *Animal Cognition* and *PLoS ONE*). In other words, these studies were quite prominent, not something dredged up from some dark repository of questionable repute.

The first concrete step in any meta-analysis is to define the criteria for what articles to include in the larger data pool. We settled on three criteria that a given article had to meet in order to be included in our analyses:

1) The research had to be published in a peer-reviewed journal.
2) The subjects (birds) in the studies had to belong to the same taxonomic group (the Corvidae family, see note 3).
3) At least some birds in the articles had to take part in at least the original water vs. sand test, plus at least one other variant.[8]

We then launched a broad search of the literature, which included combing databases with multiple variants of our search terms (e.g., corvid or crow; Aesop fable; water displacement) and consulting review articles and other articles that cited the original *Current Biology* paper. After searching through nearly one hundred abstracts, and examining several dozen papers in detail, five articles made our final cut. Two additional peer-reviewed articles were considered but

ultimately rejected from inclusion—one because subjects were western scrub jays and thus not members of Corvidae and one because the subjects were grackles—who *are* members of Corvidae—but only one grackle took part in the key water vs. sand task and that grackle refused to continue past the second trial.[9]

In the end, we were able to compile the data from twenty-eight birds from five separate peer-reviewed research articles: nineteen New Caledonian crows, five Eurasian jays, and four rooks. Of particular importance to our project was the fact that the majority of these birds (22 out of 28) participated in the original water vs. sand task. This enabled us to combine the data from these birds to investigate patterns of learning using a statistical technique called "multilevel modeling." Multilevel modeling essentially estimates or "models" underlying patterns in a dataset, and thus requires a larger amount of data than was available in any individual article.[10] In addition, across all of these articles, the subjects took part in over a dozen variations of the task.[11]

On a more practical note, each of the articles depicted the results from each bird (the "raw" data) in grids similar to that depicted in Figure 2, with one grid representing each bird's performance on a particular task. Each row represented one trial, and each column represented which object or tube the birds chose. This format allowed us to compile the data from across the studies to enter into our meta-analysis, but to do so, we had to enlist several undergraduate students to transpose the data for each choice, for every bird, and for every task variant into a giant excel spreadsheet organized by task. And we had them do this twice! To give some perspective, for the water vs. sand task we entered 1,528 data points. Across ten of the key task variants, we entered and kept track of 6,724 choices. After the data had been entered, one of us had to cross-check each data point to be sure it had been entered correctly.[12] By way of comparison, because they ignored the individual stone drops and only analyzed the results of each trial, the combined group of original researchers (spread out across the five separate publications), only had to keep track of 408 data points across the variants we analyzed. Having summarized the data in this way, we could "model" the data to get some answers to our three main questions (see above), as well as several others.

The easiest way to think about our statistical approach is to realize that on each trial the bird is confronted with either one pile of objects and two test tubes (sand vs. water), or one water test tube and

two kinds of objects (floating vs. sinking blocks). Either way, the bird has two options. If the birds were just picking test tubes (or objects) at random, they should pick each one about 50 percent of the time. We can thus ask: When the birds initially began each task, was their performance random? If so, how many choices did it take for their performance to improve? Were some tasks learned faster than others? When they made a "good" choice, was their next choice more likely to also be a "good" choice? What about "bad" choices—did they learn anything from those? And finally, did they get any better as they encountered new variants of the task, or did they have to learn each one from scratch?

We have since completed and published our meta-analysis (see Hennefield, Hwang, et al. 2018).[13] In Figure 3, we have graphically depicted the choices that the crows made in three of the most important variants of the Aesop's Fable Paradigm. Each thin line represents one bird and relates their preference for one option over another as a function of increasing number of individual stone drops. The thick lines represent the overall relationship. Thus, it is possible to see how each bird's behavior changed (or did not change) as they progressed through each task.

Two things are immediately striking about these results. The first is that for the water vs. sand and float vs. sink tasks, the birds' choices started out near the 50 percent mark (statistically their choices did not initially differ from chance). As the task progressed, however, nearly all birds began to choose the "good" choice with increased regularity. This pattern is exactly what we would expect to see if the birds are learning how to more quickly retrieve the worm as they gain experience with the God's-eye, immutable facts about what happens when a stone is dropped into a test tube of water with a worm floating on top (the worm moves closer) versus when a stone is dropped into a test tube filled with sand (the worm remains just as far away).

The second aspect of Figure 3 worth noting is that in the solid vs. hollow task, the birds' choices were essentially at ceiling throughout the entire task. That is, they started out by initially choosing the solid "good" option and kept choosing that option as time went on. This result is compatible with several hypotheses. First, the birds' may have begun the task with an understanding of volume and water displacement. Second, the birds may have learned something general from their prior testing (to pick up and drop objects that require this much effort). Or third, as some of the authors themselves argue, the

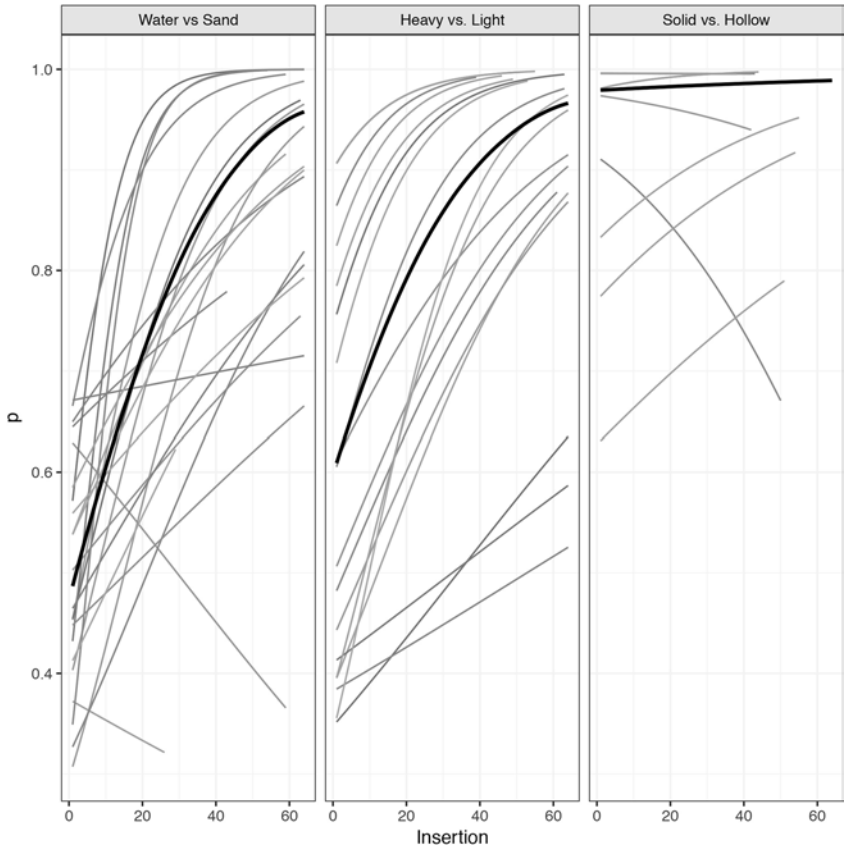

FIGURE 3
This figure depicts the choices each crow made in three key variants of the Aesop's fable tasks. Each thin gray line represents one bird and depicts their preference for one option over another as a function of increasing number of individual stone drops. The thick dark line represents the overall relationship. For example, in the water vs. sand task crows' initial choices tend begin near chance (the 0.5 mark on the y axis) and the upwardly sloped lines indicate the crows that increasingly chose to drop stones into the water tube as the task progressed.

birds either have an *a priori* preference for the solid objects, dislike the feel of the hollow objects in their beaks, or any number of possible alternative reasons. We should note that although there may have been some exceptions (see above), our modeling revealed that the birds did not, in general, transfer information learned in one task to the many subsequent tasks they were given. That is, the birds did not perform better on later as opposed to earlier tasks. This suggests to us that the birds did not "frame" these tasks as, for example, any good

folklorist would: variants of the same underlying motif, such as water displacement. (Just to give a flavor for the diversity of those variants, here are some of their names: large vs. small stones, air vs. water, sinking objects vs. floating objects, baited test tube vs. unbaited test tube, hollow object vs. solid object, narrow test tube vs. wide unequal test tube, etc. The complete list of tasks that we analyzed can be found in our published report (Hennefield, Hwang, et al 2018).

We have saved our most important finding for last: our models revealed a curious fact about the source of the bird's learning, hidden in the flurry about stones drops across the many trials they were given on each task—a fact that is difficult to reconcile with the idea that the bird's either started with or learned something about water displacement. Specifically, in the tasks where they performed better across time (the water vs. sand and float vs. sink tasks), the source of their learning was restricted to their successful stone drops! This is rather remarkable. Let us use the water vs. sand task to illustrate. When a bird made a good choice (i.e., dropped a stone in the water) their very next choice was about 5 percent more likely to be a good choice as well. This small but steady bias (presumably the result of the worm moving closer to their beaks), incrementally led them to home in on the correct choice more and more frequently. Startlingly, however, when the birds dropped stones into the sand tube (the bad choice), they were just as likely to repeat that bad choice on the next stone drop. In other words, they learned nothing from dropping stones into the sand tube. Our modeling revealed the same pattern in the float vs. sink task.

Our primary conclusion from our meta-analysis is that these studies simply do not tell us anything new or interesting about animal cognition. Our results are highly consistent with a model suggesting the birds were learning through trial and error, not higher-order ideas like "volume" or "mass" or "displacement." In sum, we find no evidence of these birds having their Archimedes-like "Eureka!" moments.

In a strange bonus of sorts, after we had completed our work and submitted it for publication, we discovered that one other team had suspected a similar explanation of the Aesop's Fable Paradigm and conducted their own meta-analysis.[14] Although they raised several of the issues that we have discussed, they still chose to base their analyses on the trial-level data as reported in the original articles, rather than the drop-by-drop data. Equally puzzling to us, they did not challenge the ability of the paradigm to provide new insight into crow cognition.

How Crows Helped Us Become More "Compleat" Academics

At first, our investigation of the Aesop's Fable Paradigm was just an interesting intellectual exercise that closely mirrored the challenges we were facing as young experimental psychologists designing and conducting our own studies with children. Animals have a lot in common with children. Animals do not use language and young preschoolers' grasp of language is limited, so the problem of developing experimental tasks that can assess their respective cognitive abilities is similar.

Because we are not comparative psychologists, it felt somehow easier to be objective as we started digging into the research literature on animal cognition. We had nothing directly at stake in the questions, and we did not really know who the "key players" were in that field. It is an inescapable fact that seeing a "famous" psychologist's name before reading an article definitely colors one's assessments of the work. Here, there was a lot less pressure and background noise as we began to assess the premise, methods, and interpretations of the Aesop's Fable Paradigm. We could deploy our passion in understanding the experiments, ranging from seemingly minor details of methods (e.g., how many training trials did the crows need before they could even do the test trials?), to deeper conceptual questions (e.g., does dropping stones into a tube to retrieve a worm indicate that crows have an idea of water displacement?), without worrying about how it might affect our careers.

We entered experimental psychology with a strong passion for and trust in experiments. We thought, "Experimental research is the real key to science. Experiments provide us with the means to objectively test hypotheses via systematic manipulation of variables, and to make subsequent causal claims about objective truths. Experimentation is the tool to getting us closer to the real truth." We have come to realize that experiments are not always objective. The experimenters—the scientists themselves—have subjective biases that influence how they set up experiments and how they interpret the results and then present the findings to the public. Experimenter bias runs on a continuum from biases as benign as only looking for evidence that supports one's theory and not evidence that disproves it to as malicious as altering the data itself. Simply put, subjectivity is an inescapable issue in all experimental fields.

Then, too, there was the pall of the replication crisis that was hanging over psychology.[15] Just as we were starting graduate school, a distributed effort of scientists around the globe had discovered that a sizeable collection of very famous psychological findings were not replicable. Despite the fact that these findings were in textbooks and widely heralded in both the scientific and popular media, the results seemed to be illusions—statistical and sociological artifacts. The replication crisis has been attributed to numerous factors, but one of those factors felt all-too-real to us: the pressure to present nice, tidy findings and to ignore null findings (i.e., when experiments do not show a statistically significant difference between two conditions). Our increasing awareness of the threat to experimental psychology from nonreplicable or exaggerated research claims also played some role in our decision to dive into the meta-analysis of the Aesop's fable-inspired research.

Let us be absolutely clear: we are not seeking to lead a crusade against crow intelligence. We have nothing against the idea of crows having a concept of water displacement. We even admit that earlier in our careers we would have reacted to any refutation of the Aesop studies with boredom and distaste. Let's face it: there is nothing flashy and exciting about a couple of graduate students trying to undermine research that produces headlines such as "The Rook and the Test Tube: Fable Made Fact" (*Science Magazine*), "Much to Crow About" (*The Economist*), "Clever Crows Prove Aesop's Fable is More than Fiction" (*Wired*), "Crows Understand Water Displacement Better Than Your Kid" (*Smithsonian*), "Aesop's Fable? This One Turns Out to Be True" (*The Independent*), and "The Moral: Aesop Knew Something About Crows" (*The New York Times*).

Throughout our career as graduate students, we had heard that it was difficult to publish experiments that do not show directionality in their findings ("under condition Q, outcome X is far more likely than outcome Y"). Even if the design is well done, we were told, no journal wants to hear a story that is not exciting or definitive. We had been told over and over that we had to be able to tell a good story about our research in order to get noticed. We even took a career development seminar taught by a prominent psychologist and based on his coauthored book, *The Compleat Academic*. In hindsight, one of the quotes on the back jacket of that book seems especially revealing: "You may think science is somehow the opposite of storytelling, but this is not the case. Good science tells a story."[16] From this vantage point, it made

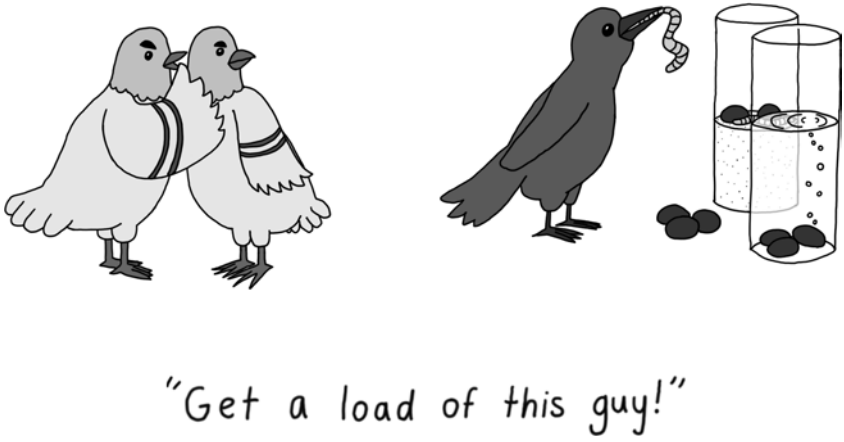

FIGURE 4
Tidy stories drive scientific conversation . . . so we are told. Cartoon by Gavin Rackoff.

perfect sense to us why the Aesop's Fable Paradigm had become so popular. It was a tidy story with a catchy interpretation.

Digging deeper into the Aesop's fable, though, also made us recognize the tendency in the study of both developmental and animal/comparative cognition to approach cognitive questions by putting forth a theory (which is a great first step) but not trying to actively disprove it. This is a fundamental philosophical problem in trying to establish the viability of an idea. Whereas you cannot hope to find all of the necessary evidence to prove a theory, you only need one contrary piece of evidence to disprove it. This is the classic, "All swans are white" idea.[17]

On the one hand, you can spend your time and money trying to gather all the swans in the world (or as experimental psychologists have come to tackle this problem, trying to get a representative sample of all the swans in the world). On the other hand, you can design your approach and your resources to do everything possible to find that one black swan. Of course, it is difficult to come up with a disprovable, falsifiable theory and to present that theory in ways that can be actively tested. In the search for mental continuity between humans and animals, it is a much more common practice to gather evidence *supporting* a theory rather than it is to work toward disproving it. We have come to wonder if this has something to do with the practical impact of finding that "black swan"—the fear that you will

not have a good story to tell, that you will fail to produce the kinds of novel and exciting research that will allow your work—and with it your scientific career—to rise to the top.

In our minds, the Aesop's Fable Paradigm came to exemplify this problem. The inherent goal in these studies, it seems, is to find evidence that proves crows have complex cognition rather than to find evidence that disproves this statement. The fact that these articles got picked up by the popular press and widely disseminated also contributed to our decision to devote our time to reanalyzing and writing up the findings in our meta-analysis—not that we had any expectations that our work would receive any popular acclaim. Even our initial inspection of the data (see, for example, our discussion of Figure 2 above) strongly suggested that our results might be quite deflationary. That is, we did not feel that our meta-analysis would make a good story. Nonetheless, we felt compelled to proceed.

We were taught early on how difficult it is to be objective in one's own research—that there is a psychological bias to give more weight to evidence that fits one's current framework than to evidence that contradicts it. Psychologists should know these biases exist—these tendencies that distort our thinking—but knowing this does not mean we are not susceptible to the biases just the same. It is a bit like St. Louis' iconic Gateway Arch. The Arch looks much taller than wide, but actually its height and width at the base are exactly the same. It is an optical illusion that is hard to unsee—even when you know the measurements (630 feet in both directions, as a matter of fact). We thought that, as outsiders, tackling a meta-analysis of the Aesop's fable tasks would be an opportunity to provide a different perspective on the broad theoretical and methodological assumptions employed in the Aesop's Fable Paradigm. This perspective, we fully recognize, is much easier to proffer when one is less familiar with the players and the conventions of a given field.

How Many Stone Drops Does It Take to Be Human?

Data is the gold standard of scientific research. For scientists, new data has the potential to provide new knowledge about the world. Small sample sizes and the painstaking work that goes into collecting each data point is something that connects researchers studying cognitive development in animals and humans. Trying to elicit a meaningful response from a two-year-old child in a word-learning paradigm

("Which one is the *blicket*?"), or to elicit a valid verbal response from a four-year-old in an event-expectation task ("Do you think the kite will or will not get stuck in the tree?") is no small feat. It cannot be any easier to figure out how to ask a crow if it understands why dropping a stone in a waterfilled test tube makes the floating worm get closer.

The theoretical and methodological shortcomings encountered in the Aesop's fable-inspired studies are not unique to the paradigm. Instead, they exemplify many of the pitfalls that appear with surprising regularity in comparative and animal cognition. As we diligently sifted through recent research in comparative cognition in that seminar back in 2014, time and time again we found researchers presenting clever experimental designs purporting to demonstrate some new cognitive ability—complete with a bevy of sometimes very odd (i.e., irrelevant) control conditions. The researchers would proceed to employ a rich interpretation of the behaviors that went far beyond those warranted by the experiment. The fact that there were so many poorly designed studies and lapses in critical thinking was disheartening—so much so, that we began to keep a running list of "Fundamental Obstacles in a Valid Science of Comparative Cognition" that detailed some of the same pitfalls we encountered repeatedly in our readings.

When we first read the Aesop's fable experiment, we were just starting to learn about the rules of the game of publishing and surviving in academia. In the years that have passed since then, we worked on our meta-analysis while working hard to finish our own doctoral research with preschoolers and to secure grant funding for our postdoctoral experiences. We have had a lot of time to think about how individual researchers (including ourselves) struggle to shape this game. And now, as we both start to establish new lines of developmental research and navigate our increasing scientific independence, we see how this project has sharpened our focus on things that have concerned us all along: How do we manage the need for objectivity in our science with the need to be a complete academic—to tell a good story about our research, to raise money to run our labs? How do we make our work stand out from the background? Why do some findings rise to the top? Faced with these challenges, will we have the courage to see limitations in our own research? We hope our work on the Aesop's Fable Paradigm constitutes a first step in the right direction.

That being said, we also see a larger, cautionary tale unfolding—namely, the dangers of humans' folk narratives becoming embedded into scientific storytelling. Aesop's fable is the most obvious example,

but the fact that an alternative explanation for the results of such a widely heralded set of studies has been largely overlooked, leads us to wonder just how much of science is being driven by the need to tell good stories. We wonder whether or not the story used to frame the findings is more culturally important in codifying findings into the scientific canon than the quality of the methods used to obtain those findings. We wonder whether turning to folktales and fables for inspiration is a reasonable way to advance science. And finally, we still wonder exactly how the fable-turned-science has risen to the top.

No fable ends without its moral, and as with many fables, the moral of our meta-fable is variable. Variants might include: Crows are as smart as lever-pressing rats. Twenty-eight crows as smart as lever-pressing rats, does not a good story make. Even if a chimpanzee sitting at a typewriter might eventually hack out a line of Shakespeare, crows will never drop enough stones to produce *The Tempest*.

Washington University
St. Louis

University of Chicago
Chicago

University of Louisiana
Lafayette

Acknowledgments

It is an important tradition among scientists to acknowledge any applicable funding sources that support our research and our writing. Laura Hennefield's work was supported by NIH training grants T32 MH100019 and F32 HD093273.

Notes

1. This interest has come from both directions. From the standpoint of those studying animals, consider the closing paragraph of Wolfgang Köhler's 1917 landmark monograph regarding chimpanzee intelligence:

> One would like to have a standard for the achievements of intelligence described here by comparing with our experiments the performances of human beings (sick and well) and, above all, human children of different ages. As the results in this book have special reference to a particular method

> of testing and the special test-material of optically-given situations, the psychological facts established in human beings (especially children), under the same conditions, would have to be used. But such comparisons cannot be instituted, as, very much to the disadvantage of psychology, not even the most necessary facts of this sort have been ascertained. Preliminary experiments—some have been mentioned—have given me the impression that we are inclined to over-estimate the capabilities of children of all ages up to maturity, and even adults, who have had no special technical training in this type of performance. We are in a region of *terra incognita*. . . . As experiments of this kind can be performed at the very tenderest age, and are certainly as scientifically valuable as the intelligence tests usually employed, it does not matter so much if they do not become immediately practicable for school and other uses. M. Wertheimer has been expressing this view for some years in his lectures; in this place, "where the lack of human standards makes itself so much felt, I should like to emphasize particularly the importance and—if the anthropoids do not deceive us—the fruitfulness of further work in this direction." ([1917] 1925, 268)

From the perspective of the child psychologist, there is no better early report than Lightner Witmer's report of his investigations of a chimpanzee named Peter, who Witmer was able to examine in his Boston clinic after seeing him perform in a traveling Vaudeville show. Although initially skeptical, Witmer opens his report with great optimism:

> Since that day I have seen Peter in five public performances, have tested him at my Psychological Clinic at the University of Pennsylvania, and privately on three occasions. I now believe that in a very real sense the animal is himself giving the stage performance. He knows what he is doing, he delights in it, he varies it from time to time, he understands the succession of tricks which are being called for, he is guided by word of mouth without any signal open or concealed, and the function of his trainer is exercised mainly to steady and control. (1909, 182)

But Witmer ends his report on a decidedly ambiguous note:

> Peter's activity is not the result of mere animal spirits; he is mentally alert and possessed of unusual power of concentration, not merely for an animal but for a child of his own age. . . . [However] even though we may grant a fair prospect in the direction of intellectual development, we must assume from our present knowledge of men and apes that Peter is and will remain morally imbecile. It would be a nightmare flight of the imagination to suppose that an ape could acquire a will determined consciously by moral motives. [His owners] claim that no one really knows how intelligent Peter is and they appear to believe that Peter excels the human being in quickness of action, thought and comprehension. If they are right, Peter should become the ward of science and be subjected to proper educational influences. He has been trained, he is partly educated, but no effort has

> yet been made to give him what an education really stands for. I venture to predict that within a few years chimpanzees will be taken early in life and subjected for purposes of scientific investigation to a course of procedure more closely resembling that which is accorded the human child. (1909, 203–5)

2. For a short review of tool use in animals, including a pointed discussion of the current controversies surrounding tool use in comparative cognition and citations to original research, we recommend Amanda Seed and Richard Byrne's "Animal Tool-Use" (2010).

3. The studies actually involve a variety of birds from the Corvidae, a taxonomic family that includes rooks, jays, and crows. For simplicity sake, throughout this article we colloquially refer to them all as "crows."

4. Throughout this essay, we use the term "sand" as a general term to cover this variant of the task. In some cases sand was used, in other cases sawdust or wood chips were used.

5. In their original *Current Biology* article, Bird and Emery suggest that the rapid learning and efficient solutions demonstrated by rooks provide evidence that rooks solve "complex physical problems via causal and analogical reasoning" (2009, 1410). A subsequent article by Taylor and colleagues seemed to temper this claim by suggesting that the "crows' performances were not based on associative learning alone" (2011, 1). More recently, Jelbert and colleagues stated in their abstract that "results indicate that New Caledonian crows possess a sophisticated, but incomplete, understanding of the causal properties of displacement, rivaling that of 5–7 year old children" (2014, 1).

6. Specific formulations of the special nature of the tests—that is, what sets them apart from nearly a century's worth of preceding studies on animal learning—are difficult to work out from the articles. However, several of the researchers do briefly touch on this topic. Taylor et al. suggest the paradigm measures whether subjects "can process causal information" (2011, 1). Likewise, Jelbert et al. state that the paradigm can be used to investigate whether the subjects understand "causal regularities" (2014, 2). Unfortunately such descriptions are of limited use because phrases such as "process causal information" and "understanding causal regularities" do not define the underlying processes in question, nor do they elucidate why this test is more suited to measure these abilities than the hundreds (if not thousands) of others that comparative psychologists have devised over the past century.

7. Possibly less important in humanities and social-science disciplines, an impact factor is a score assigned to academic peer-reviewed journals that reflects the number of citations, relative to number of articles, for recent articles published in that journal. Impact factor is often used as an indicator of the relative quality and importance of a journal within a given field. In science, publishing "early and often" in journals with high impact factors is considered a measure of career success, with impact factors often considered in hiring and promotion decisions.

8. Most of the studies included in our meta-analysis followed the rough steps of the first Aesop fable experiment published by Bird and Emery (2009). All subjects first underwent a training procedure in which they learned to drop stones into a tube to retrieve a food reward (either a worm or piece of meat). Then, in

a majority of the articles, subjects participated in the sand vs. water task, followed by several other task variants.

9. Our decision to restrict inclusion to the members of the Corvidae family—and thus exclude Logan, Harvey, Schlinger, and Rensel's (2015) study with four western scrub jays (not members of Corvidae)—was twofold. First, using the established taxonomic grouping of the biological "family" as our cut-off has face validity—that is, on the surface it seems like a reasonable decision. Second, the western scrub jays were not considered "successful" in the tasks by the authors of the study. Two jays did not learn to drop stones into a tube during the training phase. Of the remaining two jays that "passed" the training, one did not complete the water vs. sand task (possibly because his preference for the sand tube resulted in few rewards and decreasing motivation to continue to drop stones) and the other completed the task but did not exhibit a preference for the water tube. This second point is important because our goal was to try and achieve maximum "buy in" from both reviewers and other researchers. Not only do we want our decisions to *appear* objective, but when faced with decisions that others might find questionable, we aimed to be as conservative as possible in our choices. In other words, if we included the jays, it is quite possible that we would have gotten pushback because the birds are not members of Corvidae. After all, including two birds in our analyses who never showed a preference for the water tube could strengthen our conclusions about the role of learning in these tasks (i.e., the jays simply did not learn), but do not serve to advance a story of "complex cognition" throughout the order of Passeriformes (of which corvids and jays both belong).

10. There were two features of the data in the Aesop's fable tasks that governed our choice of analyses. First, although it is possible to count and add and combine data within and across these tasks, each individual data point is *binary*. For example, in the sand versus water test, the subjects either chose the sand tube (which we can assign a score of 0) or the water tube (score of 1). In the other variants, involving choices between two objects (such as light versus heavy), we could also use this binary coding: object A or object B. Binary data is discrete and thus different from measures that are continuous (consider, variables such as income, age, or the amount of time it takes someone to complete a task. Second, the data points are *not independent*. That is, the same subjects repeatedly performed each behavior and each bird contributed multiple data points to each task (up to one-hundred stone drops per task for some birds). Independence is an assumption that must be met in order to use conventional statistical analyses such as t-tests and ANOVAs. Properties of data—in this case binary and not independent—constrain the analyses that are appropriate to use to test the data. These particular constraints led us to multilevel modeling. Multilevel modeling is typically used when the data is "nested" at more than one level; for example, stone drops were nested within subjects, and subjects were nested within articles. Although we were each familiar with this statistical technique, neither of us were experts, so we recruited the assistance of our colleague, Sara Weston, who has expertise in this area. Sara worked closely with us to build code that produced the models, to help us select which models to include in the meta-analysis, and to create the figures for our manuscript that best captured our key findings.

11. In much the same way that we developed the inclusion criteria to select the five articles that we used in the meta-analysis we also developed inclusion criteria to determine which tasks within each article to include in the analyses. We used a fairly minimal inclusion criteria here to retain as much data as possible—namely that the task had to involve water (displacement) and a binary choice. These criteria yielded a total of ten tasks across the five articles. Only a handful of tasks were excluded, and these excluded tasks each appeared only once across the articles and did not clearly relate to the broad topic of water displacement (e.g., one involved the use of an arbitrary reward; another was a tube-search task).

12. We had our undergraduate students double-enter the data from the original grids in the published Aesop's fable articles. Each data point was entered twice (by two different students), both blind to the hypotheses of the study, and then the data points were checked for consistency. We found agreement to be extremely high (Cohen's Kappa = 0.985; the score for perfect consistency would be 1), and the few discrepancies were resolved by one of us.

13. Hennefield and Hwang contributed equally to this manuscript.

14. Although it was a bit disheartening to discover they had published a meta-analysis on the same topic as the one we had been working on for several years, we feel a sort of camaraderie with Ghirlanda and Lind (2017) through our mutual skepticism of the claims put forth by the Aesop's fable researchers. In fact, we had not known about their meta-analysis until it was brought to our attention by a journal editor upon the submission of an initial version of our manuscript. It is true, our meta-analysis was a side-project, and perhaps if we had spent more time earlier on with it we could have been the first to publish. It was also mildly frustrating that after carefully preparing our original manuscript (again, not knowing that Ghirlanda and Lind were simultaneously thinking about similar ideas) we had to subsequently revamp large portions of the introduction and discussion to account for their findings and more clearly elucidate what sets our work apart from theirs. However, it is likely that this revision has served both to clarify and strengthen our arguments, and is just one the many types of stumbling blocks that we have learned to handle in our budding careers.

15. For an applicable discussion of the replicability crisis, see Pashler and Harris (2012). They identify three arguments of central importance to the replicability crisis: 1) the prevalence of false-positive findings in the scientific literature, 2) the costs and benefits of direct replications versus conceptual replications, and 3) the notion that the scientific process is self-correcting and erroneous findings will eventually get weeded out. For a discussion of the intersection between replication and falsification, we suggest Earp and Trafimow (2015).

16. Quote by Robert J. Sternberg, Professor of Human Development at Cornell University, on the back cover of *The Compleat Academic: A Career Guide* (2004).

17. Karl Popper (1935) famously argued against the classical approach toward science that seeks to prove theories or hypotheses (such as "all swans are white"). He argued that it is logically impossible to prove a hypothesis from individual cases: "no matter how many instances of white swans we may have observed, this does not justify the conclusion that all swans are white." ([1935] 2002, 4). However, if we can find that one single swan that is not white, deductive logic allows the conclusion that the hypothesis of "all swans are white" is false. Popper

argued that the goal of science should therefore be attempts at falsifying hypotheses and emphasized the importance of reproducibility of experiments and observation. Ultimately, he argued for considering reproducibility necessary for observations to be admitted as sound evidence in science.

References Cited

Bird, Christopher D., and Nathan J. Emery. 2009. "Rooks use Stones to Raise the Water Level to Reach a Floating Worm." *Current Biology* 19 (16): 1410–14.

Earp, Brian D., and David Trafimow. 2015. "Replication, Falsification, and the Crisis of Confidence in Social Psychology." *Frontiers in Psychology* 6: article 621.

Ghirlanda, Stefano, and Johan Lind. 2017. "'Aesop's Fable' Experiments Demonstrate Trial-and-Error Learning in Birds, but no Causal Understanding." *Animal Behaviour* 123: 239–47.

Hennefield, Laura, Hyesung G. Hwang, Sara J. Weston, and Daniel J. Povinelli. 2018. "Meta-Analytic Techniques Reveal That Corvid Causal Reasoning in the Aesop's Fable Paradigm Is Driven by Trial-and-Error Learning." *Animal Cognition*. 21 (6): 735–48. doi.org/10.1007/s10071–018–1206-y.

Jelbert, Sarah A., Alex H. Taylor, Lucy G. Cheke, Nicola S. Clayton, and Russell D. Gray. 2014. "Using the Aesop's Fable Paradigm to Investigate Causal Understanding of Water Displacement by New Caledonian Crows." *PLoS ONE* 9 (3): e92895.

Köhler, Wolfgang. [1917]1925. *The Mentality of Apes*. Translated by Ella Winter. New York: Harcourt.

Logan, Corina J., Brigit D. Harvey, Barney A. Schlinger, and Michelle Rensel. 2015. "Western Scrub-Jays Do Not Appear to Attend to Functionality in Aesop's Fable Experiments." *PeerJ* 4: e1707.

Pashler, Harold, and Christine R. Harris. 2012. "Is the Replicability Crisis Overblown? Three Arguments Examined." *Perspectives on Psychological Science* 7 (6): 531–36.

Popper, Karl. 2002. *The Logic of Scientific Discovery*. 2nd ed. New York: Routledge.

Taylor, Alex H., Douglas M. Elliffe, Gavin R. Hunt, Nathan J. Emery, Nicola S. Clayton, and Russell D. Gray. 2011. "New Caledonian Crows Learn the Functional Properties of Novel Tool Types." *PLoS ONE* 6 (12): e26887.

Seed, Amanda, and Richard Byrne. 2010. "Animal Tool-Use." *Current Biology* 20 (23): R1032–R1039.

Witmer, Lightner. 1909. "A Monkey with a Mind." *The Psychological Clinic* 3 (7): 179205.

Laura Hennefield is a postdoctoral scholar in the Department of Psychological and Brain Sciences and the Department of Psychiatry at Washington University in St Louis. Her current research focuses on the development of optimism in preschoolers, including how optimism affects how children learn from and about the world around them, neural correlates of optimism, and how a lack of optimism may contribute to psychopathology in early childhood. (lhennefield@wustl.edu)

Hyesung Grace Hwang is a postdoctoral fellow in the Department of Psychology at University of Chicago. Her research investigates the development of social discrimination and exclusion by examining how infants and children learn to categorize and view people based on race and language, the impact of racial and linguistic diversity on development, and the neural mechanism behind the preference for one's own social group. (hyesung@uchicago.edu)

Daniel J. Povinelli is Professor of Biology at the University of Louisiana at Lafayette. His primary interests center on the characterizations of the higher-order cognitive functions in great apes and humans. He is the author of *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works* (2000) and *World without Weight: Perspectives on an Alien Mind* (2011). (povinelli@louisiana.edu)