Animal Cognition (2018) 21:735–748
https://doi.org/10.1007/s10071-018-1206-y

ORIGINAL PAPER

Meta-analytic techniques reveal that corvid causal reasoning in the Aesop's Fable paradigm is driven by trial-and-error learning

Laura Hennefield[1] · Hyesung G. Hwang[1] · Sara J. Weston[2] · Daniel J. Povinelli[3]

Received: 7 October 2017 / Revised: 7 June 2018 / Accepted: 4 August 2018 / Published online: 21 August 2018

Abstract

The classic Aesop's fable, *Crow and the Pitcher*, has inspired a major line of research in comparative cognition. Over the past several years, five articles (over 32 experiments) have examined the ability of corvids (e.g., rooks, crows, and jays) to complete lab-based analogs of this fable, by requiring them to drop stones and other objects into tubes of water to retrieve a floating worm (Bird and Emery in Curr Biol 19:1–5, 2009b; Cheke et al. in Anim Cogn 14:441–455, 2011; Jelbert et al. in PLoS One 3:e92895, 2014; Logan et al. in PLoS One 7:e103049, 2014; Taylor et al. in Gray R D 12:e26887, 2011). These researchers have stressed the unique potential of this paradigm for understanding causal reasoning in corvids. Ghirlanda and Lind (Anim Behav 123:239–247, 2017) re-evaluated trial-level data from these studies and concluded that initial preferences for functional objects, combined with trial-and-error learning, may account for subjects' performance on key variants of the paradigm. In the present paper, we use meta-analytic techniques to provide more precise information about the rate and mode of learning that occurs within and across tasks. Within tasks, subjects learned from successful (but not unsuccessful) actions, indicating that higher-order reasoning about phenomena such as mass, volume, and displacement is unlikely to be involved. Furthermore, subjects did not transfer information learned in one task to subsequent tasks, suggesting that corvids do not engage with these tasks as variants of the same problem (i.e., how to generate water displacement to retrieve a floating worm). Our methodological analysis and empirical findings raise the question: Can Aesop's fable studies distinguish between trial-and-error learning and/or higher-order causal reasoning? We conclude they cannot.

Keywords Causal reasoning · Causal understanding · Comparative cognition · Aesop's fable · Corvid · New Caledonian crows · Object bias · Perceptual-motor feedback

Introduction

Ghirlanda and Lind (2017) recently re-evaluated evidence from a series of research studies in comparative cognition that were inspired by the Aesop's fable, *Crow and the Pitcher*. The studies demonstrate that corvids (e.g., rooks, jays, and crows) can learn to drop stones into water-filled tubes to raise the water level to reach a floating worm or piece of meat (Bird and Emery 2009b; Cheke et al. 2011; Jelbert et al. 2014; Logan et al. 2014; Taylor et al. 2011), and have been used as evidence of "complex cognition" in these birds. The new analyses by Ghirlanda and Lind challenge this conclusion. They combined trial-level data across subjects within each article and conclude that the initial preferences for functional objects, combined with trial-and-error learning, may account for subjects' behaviors. Furthermore, they pointed out that the initial preferences for functional

Laura Hennefield and Hyesung G. Hwang contributed equally to this manuscript.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10071-018-1206-y) contains supplementary material, which is available to authorized users.

✉ Laura Hennefield
lhennefield@wustl.edu

✉ Daniel J. Povinelli
povinelli@louisiana.edu

[1] Washington University, St Louis, MO 63130, USA

[2] Northwestern University, Chicago, IL, USA

[3] Department of Biology, University of Louisiana, Lafayette, LA 70504, USA

objects, when found, could be expected from stimulus generalization or other associative learning processes.

In the present paper, we report an analysis of the Aesop's fable tasks that was independently inspired and allows for a more fine-grained investigation of the data. The starting point for our analysis is the observation that, within every experimental variant of these tasks, each single "trial" is, in fact, comprised of a variable number of object drops. The culmination of these drops results in the subject either retrieving a food reward or failing to do so. Traditionally, the final outcome of the trial has been assumed to be the fundamental unit of learning. However, we note that each instance of a subject dropping an object (within a trial) offers an opportunity for learning (i.e., the food reward either moves closer to the subject or remains stationary). This insight allows us to (among other things) precisely estimate the relevant rate of learning demonstrated by subjects in these studies. Specifically, in what follows, we: (1) unpack the rationale of these studies, (2) provide new analyses of both rate and mode of learning in several versions of the tasks, (3) examine whether there is transfer of learning between tasks, and (4) assess whether this paradigm has unique value in understanding corvid cognition, in particular, and/or the field of comparative cognition more broadly. Our work thus complements that of Ghirlanda and Lind (2017). We intend this as a constructive exercise to help identify general methodological and theoretical pitfalls in comparative research that can help to shape future research strategies.

An experimental paradigm inspired by a fable

Bird and Emery (2009b) were the first to employ the Aesop's fable paradigm. In their first experiment, the researchers introduced rooks to a clear tube partially filled with water and baited with an out-of-reach worm. A pile of ten stones sat nearby. To solve the task, the rooks needed to drop between one and seven stones into the tube to raise the water level and reach the worm. All four rooks dropped stones until the worm was within reach and then retrieved the worm. In a second experiment, the rooks learned to drop large stones over small stones. In a third experiment, the rooks were presented with two clear tubes: one partially filled with sawdust and the other with water. Each tube had a worm placed on the surface of the substrate. Over a number of trials, the rooks learned to drop stones into the water tube to retrieve the worm. In subsequent studies, other researchers have replicated these tasks and further manipulated the properties of the objects which the birds were given to drop (e.g., hollow vs. solid, sink vs. float; Cheke et al. 2011; Taylor et al. 2011; Jelbert et al. 2014; Logan et al. 2014). Across many (but not all) variations of the basic procedure, at least some birds have been successful in retrieving the food—thus, "solving" the task.

Formulations of what abilities that the Aesop's fable paradigm measures have been somewhat obscure, but researchers have consistently stressed its unique potential for understanding animal cognition. Bird and Emery (2009b) suggest that the rapid learning and efficient solutions demonstrated by rooks provide evidence that rooks could solve "complex physical problems via causal and analogical reasoning" (p. 1410). Taylor et al. (2011) suggest the paradigm measures whether subjects "can process causal information" (p. 1). Likewise, Jelbert et al. (2014) state that the paradigm can be used to investigate whether the subjects understand "causal regularities" (p. 2). Such descriptions are of limited use, however, because phrases such as "process causal information" and "understanding causal regularities" do not define the underlying processes in question. Given that the history of comparative cognition is replete with demonstrations of animals' understanding of causal regularities (e.g., rats learning to press a lever multiple times to obtain a food reward), this definitional ambiguity is worrisome. Indeed, as noted by the early theorists (e.g., Tolman 1932), linking a cause/action to an effect is the bedrock of goal-directed behavior in animals (for a review of the evidence that animals treat causal relations differently from non-causal ones, see Penn and Povinelli 2007). Presumably, the Aesop's fable researchers are not trying to provide yet more evidence for such a well-established phenomenon. Instead, they seem to pit a generic "associative learning" model against "complex cognition". This approach fails to account for the rich empirical and theoretical literature aimed at addressing the causal aspects of animal cognition within and across species (see Cheng 1997; Penn and Povinelli 2007, for a review).

Given the ambiguity of what the paradigm is measuring, it seems important to ask why it has been so strongly embraced. Some researchers have described the value of the Aesop's fable paradigm as demonstrating that the subjects can learn to solve a "novel" problem "rapidly". Bird and Emery (2009b) intimate that, because their birds had never dropped stones into a water-filled tube before the test trials (although they had participated in an earlier study in which they dropped stones into tubes to collapse a platform to retrieve food; Bird and Emery 2009a), the relative contribution of prior task-related conditioning and learning can be screened off from "causal knowledge". Indeed, the speed with which subjects solve the tasks has been noted by all research teams using the paradigm (Bird and Emery 2009b; Cheke et al. 2011; Jelbert et al. 2014; Logan et al. 2014; Taylor et al. 2011). We surmise that this is guided by the assumption that "rapid" learning is indicative of more complex causal reasoning, whereas "slow" learning is more indicative of association learning (see Bird and Emery 2009b; Jelbert et al. 2014). This is troubling for at least several reasons. First, what is meant by rapid vs. slow learning is currently undefined and unquantified. Second, it implies that "novel behaviors" cannot be produced using classical learning techniques (i.e., operant and instrumental learning)—a decidedly incorrect proposition. Third, it implies that if subjects were familiar with the task, or were natural stone-dropping tool users, it would be impossible to determine whether their performance should be attributed to causal reasoning abilities vs. prior learned associations. This third assumption is especially problematic, because it fails to specify which "novel" actions (e.g., lifting an object, lifting and dropping an object, lifting and dropping an object through a gap, lifting and dropping an object through a gap into water to obtain a reward, etc.) would warrant the assumption that the task is novel and therefore measuring "causal knowledge". It also fails to specify *what* rate of learning would support which specific model of causal understanding, or the unique role played by higher-order, role-based constructs such as mass, volume, or displacement (see Penn et al. 2008; Povinelli 2011).

Additional caution is needed as the ease of learning varies widely depending on the biological preparedness of the organism's sensory systems to detect certain regularities, and their biological preparedness to respond (see Garcia and Koelling 1966; Domjan 1983; Shettleworth 1998; Timberlake 1993; Dunlap and Stephens 2014). Indeed, some corvid species have been observed using simple tools in the wild (New Caledonian crows: Taylor et al. 2011; Jelbert et al. 2014; Logan et al. 2014), whereas others have not (rooks: Bird and Emery 2009a; Eurasian jays: Cheke et al. 2011). Bird and Emery suggest that the ability to solve the tasks does not depend upon the ecological factor of tool use. However, any task has many demands. Some of these (e.g., attending to and orienting to food location and distance) will articulate well with the bird's evolutionarily prepared behavior, whereas others will not (e.g., dropping stones).

In the present context, Rutz et al. (2016) recently reexamined the ecological basis for the widely cited example of a New Caledonian crow 'spontaneously' bending straight pieces of wire into hooked tools to retrieve rewards from an experimental apparatus similar to the one employed in the Aesop's fable tasks (see Weir et al. 2002). Rutz et al. (2016) reported that wild New Caledonian crows routinely bend the shaft of stick tools during their manufacture, using techniques that are indistinguishable from those reported in captivity.

Using individual action data for insight into the rate and mode of learning

In their recent meta-analysis, Ghirlanda and Lind (2017) used trial-level data that were derived by pooling all of the choices made by all subjects in each trial (within each article). Their primary goal was to assess whether corvids' success in the Aesop's fable tasks could be attributed to an initial preference for the functional option and/or learning across trials. Consistent with these predictions, in the tasks that contrasted *large vs. small, sinking vs. floating*, and *hollow vs. solid* objects, New Caledonian crows, jays, and grackles (though not rooks), all showed significant first-trial preferences for functional objects (i.e., large over small stones, sinking over floating objects, and solid over hollow objects). Many subjects also selected the functional objects more frequently within trials as the task progressed. The researchers also tested subjects' first-trial preferences for functional substrates (i.e., water) in the *water vs. sand* contrast. Here, the results were mixed, with New Caledonian crows showing a first-trial preference for the water tube in two of three experiments. Ghirlanda and Lind also found that performance increased substantially across trials for most subjects, but did not find evidence for meaningful individual differences between birds. Together, these findings suggest that successful performance in the Aesop's fable tasks can be accounted for by a combination of an initial preference for the functional option and the learning that occurs across trials within each task (see Ghirlanda and Lind 2017 for additional discussion of stimulus generalization and associative learning effects).

Our meta-analytic approach differs from Ghirlanda and Lind (2017) in two important ways. First, our analyses utilize a finer-grained unit of learning (i.e., each discrete object drop), instead of considering learning at the level of a trial (i.e., successful or unsuccessful retrieval of the food). Each object that a subject inserted had the potential to provide the subject with task-relevant information and a learning opportunity. Thus, analyzing the data by each object insertion, instead of combining multiple insertions into a single trial, provides a more fine-grained measure of learning. To that end, we analyzed subjects' learning rates as a function of each object insertion, which we argue is a more accurate unit of learning. Second, we combined subjects across studies for equivalent tasks, allowing for much larger Ns in our analyses. This strategy allowed us to quantify the actual rates and modes of learning. Specifically, we pooled and analyzed the data from five published research articles to assess learning within and between several Aesop's fable tasks. Using these data structures, we (1) conducted multilevel analyses to estimate subjects' initial preference and rate of learning in the original Aesop's fable task involving the choice between *water vs. sand*, (2) replicated and extended these multilevel analyses to two additional tasks: *sink vs. float* and *hollow vs. solid*, (3) tested whether rate of learning changes as a function of each prior action taken by the subject (i.e., within-task transfer), and (4) estimated whether there was any transfer of learning across tasks (i.e., between-task transfer).

Methods

Literature search

An initial search was conducted through the electronic Web of Science database. We searched for all available records using the following combinations of keywords: (*corvid OR crows*) AND (*Aesop*); (*corvid OR crows*) AND (*water displacement*). The search yielded 11 hits (with removal of duplicates). We also searched for articles citing Bird and Emery (2009b), which resulted in 76 articles. Finally, we consulted review articles for additional relevant studies (Ghirlanda and Lind 2016; Jelbert et al. 2015; Shettleworth 2009, 2012; Taylor 2014; Taylor and Gray 2009).

Inclusion criteria

The following three criteria were used to select research articles for this meta-analysis:

1. The article had to be published in a peer-reviewed journal; no unpublished data were considered.
2. Subjects in the article had to belong to the monophyletic group, including rooks, Eurasian jays, and New Caledonian crows. This monophyletic group is in turn a nested subset of the larger passerine bird family, the Corvidae (or "corvids").
3. Subjects in the article had to take part in at least one variant of the *water vs. sand* displacement task first published by Bird and Emery (2009b).

A total of five research articles, describing 33 separate tasks, were identified for inclusion in this meta-analysis. These articles are Bird and Emery (2009b), Cheke et al. (2011), Jelbert et al. (2014), Logan et al. (2014), and Taylor et al. (2011). Application of the second criterion excluded an article by Logan et al. (2015) that tested two western scrub jays (in addition, these birds did not demonstrate reliable learning in the water displacement task). Application of the third criteria excluded an article by Logan (2016) that tested six grackles. Only one grackle began the *water vs. sand* task and refused to continue past the second trial; thus, the task was eliminated from the study.

The unit of learning: each object insertion versus each trial

In all articles, the data were presented as individual trials, each of which consisted of multiple object drops. All studies within the articles implemented 20 trials for each subject, except Cheke et al. (2011), which used 15 trials for each subject. Despite the availability of the data for individual drops, all five articles analyzed learning rate at the level of a trial. Each trial began when subjects inserted their first object into a tube, and ended when the subject retrieved the food in the tube, exhausted all available objects, or ceased participation. A successful trial was defined as all object insertions until subjects ultimately retrieved the food. An unsuccessful trial was defined as subjects' failure to retrieve the food during a set period of time or when all available objects were inserted by subjects and they were still unable to retrieve the food. However, each trial consisted of multiple, discrete acts of object insertion, each of which either brought the reward closer to the subject or did not. The number of insertions per trial varied from 1 to 17. Thus, the length of each trial, and the subsequent amount of information that a subject could learn during each trial, varied across trials and across individual subjects.

Preliminary data transposition

All five articles primarily presented their raw data in grids, with one grid representing each individual subject's performance on a particular task. In four of the articles, within each grid, each row represented one trial, and each column represented an object insertion choice made by the subject; Cheke et al. (2011) presented the rows and columns in reverse. The squares within the grid were color-coded to indicate the subject's specific choice (e.g., blue if a stone was dropped into a water tube; green if it was dropped into a sand tube). We transposed each color-coded data point into a binary numerical data point (e.g., 1 for water tube insertion; 0 for sand tube insertion) for multilevel logistic modeling analyses. Two research assistants who were blind to the hypotheses of this study transposed the data. Agreement was extremely high (Cohen's $\kappa = 0.985$). The second author resolved any discrepancies in the data transposition.

Analysis plan

Despite the small number of birds studied in each article ($Mdn = 4$, $M = 4.2$ per task; see Table 1), there is now a larger sample on which to base multilevel modeling analyses. We used multilevel modeling to better account for the dependencies that are present within the experimental designs in these studies. Neither object insertions nor trials are independent measures, as the same subjects repeatedly perform each behavior; therefore, these data do not meet the standard assumptions of independence necessary for the conventional statistical approaches such as *t*-tests and ANOVAs. In contrast, multilevel modeling analyses allow us to statistically account for the dependent nature of the observations at each nested level (e.g., insertions nested within subjects and tasks, and subjects and tasks nested within articles), and

Table 1 Tasks and subjects included in the transfer effect analyses

Article	Subjects, $N=28$	Training	Basic water[a]	Large vs. small stones	Water vs. sand	Air vs. water	Sink vs. float[b]	Baited vs. unbaited	U-Tube	Hollow vs. solid	Narrow vs. wide equal	Narrow vs. wide unequal	Uncovered U-Tube	Other tasks[c]
Bird and Emery (2009b)	4 captive rooks	Trained in Bird and Emery (2009a)	3/3 (100)[d]	3/3 (100)	3/3 (100)	–	–	–	–	–	–	–	–	
Taylor et al. 2011	5 wild-caught New Caledonian crows	✓	4/4 (100)	4/4 (100)	4/4 (100)	4/4 (100)	4/4 (100)	–	–	–	–	–	–	2 Tube search tasks 2 tool-use tasks
Cheke et al. (2011)	5 hand-raised Eurasian jays	✓	–	–	2/4 (50)	1/2 (50)	2/2 (100)	2/2 (100)	0/3 (0)[e]	–	–	–	–	Arbitrary reward movement cues
Jelbert et al. (2014)	6 wild-caught New Caledonian crows	✓	–	–	6/6 (100)	–	6/6 (100)	–	0/4 (0)	5/5 (100)	0/5 (0)	3/4 (75)	–	
Logan et al. (2014)	8 wild-caught New Caledonian crows	✓	–	–	3/5 (60)	–	6/6 (100)	–	1/5 (20)	6/6 (100)	4/6 (67)	3/4 (75)	0/5 (0)	Multistone platform
		Totals	7/7 (100)	7/7 (100)	18/22 (82)	5/6 (83)	18/18 (100)	2/2 (100)	1/12 (8)[f]	11/11 (100)	4/11 (36)[f]	6/8 (75)	0/5 (0)[f]	

[a] The basic water task did not involve a binary choice and was thus not included in the transfer analyses

[b] Sink vs. float is called heavy vs. light in Taylor et al. (2011)

[c] The column lists additional tasks presented in each article that did not involve water (displacement), and were not included in the transfer analyses

[d] x/n = number of Ss that the authors described as successful/number of Ss participating in the task. Percent of successful birds are in parentheses. Each task involved a binary choice and water (displacement) except basic water

[e] Following the initial U-Tube task, two corvids completed a second U-Tube task in which the color and shape that denoted the functional tube was reversed. These additional data were included in the task transfer analyses in "Characterization of subjects' transfer of learning between tasks"

[f] The majority of corvids tested did not successfully complete the U-Tube, narrow vs. wide equal, and uncovered U-Tube tasks

to more accurately characterize subjects' behavior within and between tasks (see Raudenbush and Bryk 2002 for an introduction to multilevel modeling).

Multilevel logistic model

Due to the variation in number of object insertions that each subject completed—ranging from 1 to 150 insertions—we selected the median number of insertions, 63 (per subject, per task), as our cut-off point for our primary analyses. By doing so, we limit the degree to which a few subjects may over-influence the results, because they have more data points. This approach is more conservative than one based on the maximum number of object insertions, because the subjects with the most object drops also tended to make the most inefficient choices throughout the tasks. Primary analyses using the median number of object insertions are presented in ESM Appendix (1), and parallel analyses using the maximum number of object insertions are presented in ESM Appendix (2). All qualitative statistically significant effects remained the same except the one noted below in "Initial preference at task onset (intercept)".

In each task, subjects had to choose between a tube or object that would yield the reward or yield the reward at a faster rate (which we call the "efficient choice") versus a tube or object that would not (the "inefficient choice"). Because these choices were binary (e.g., either efficient tube or object was chosen [1] or not chosen [0]), we used a multilevel logistic model, modeled with the lme4 package version 1.1–12 (Bates et al. 2015) with insertion order nested within subject, and subject nested within each article. We also included species as a predictor of performance in the models. Species was not included as a predictor when modeling the *hollow vs. solid* task, because all subjects were of the same species (New Caledonian crows). We used the likelihood ratio test for model selection. For each model, we started with the maximal structure and then removed terms one at a time, starting with the study level and then proceeding to the bird level. We tested for variance of random slopes as well as the covariance between terms (see ESM Appendix 1 for model selection details). The output of the model is given in logit units (e.g., a one unit increase in insertion order leads to a predicted b_i logit increase in selecting the efficient choice). To interpret these effects, we converted the logits to odds ratios (OR) with the equation $\mathrm{OR} = e^{bi}$.

Within-task analysis plan

We characterized overall learning in each task by first assessing subjects' initial choice of tube or object at the onset of a task, and then statistically modeling the rate at which subjects chose the most efficient tube/object as that task progressed. We also assessed whether subjects learned from their choices at each insertion—that is, Did making either an efficient or inefficient choice affect whether their subsequent choice was efficient or inefficient? We restricted our within-task analyses to the three tasks with the largest samples of subjects that were claimed to demonstrate successful learning: (1) *water vs. sand*, (2) *float vs. sink*, and (3) *hollow vs. solid*. We selected the *water vs. sand* contrast (with *sand* used to refer to all non-functional substrates including sand, sawdust, and woodchips) for our analyses because it was reported in all five articles and included the largest sample of subjects ($N = 22$). In addition, in three of the articles, *water vs. sand* was the first task that subjects participated in after the initial training. We selected the *float vs. sink* and *hollow vs. solid* contrasts, because subjects were reported as having successfully completed these tasks in four articles, and these tasks had the second and third largest sample of subjects, respectively ($N = 18$ for *float vs. sink;* $N = 11$ for *hollow vs. solid;* see Table 1 for more details). Although the U-tube task included 12 subjects across three articles, and the *narrow vs. wide* task included 11 subjects across two articles, few subjects successfully learned to retrieve the food in either task (U-Tube success rate = 8.3%; *narrow vs. wide* success rate = 36.4%; see Table 1). As the purpose of these analyses was to model learning across the tasks, we excluded tasks in which learning did not appear to occur.

Between-task analysis plan

We also tested whether birds demonstrated transfer of learning across tasks—that is, whether subjects' rate of learning increased in subsequent tasks as they gained more experience with the Aesop's fable tasks. Specifically, we tested whether subjects more quickly learned to choose the efficient option in later (relative to earlier) tasks (across articles, the tasks were not given in any consistent order). This pattern of results should be expected if subjects either used task-specific information learned in earlier tasks to complete subsequent tasks, or developed a higher-order, role-based representation of water displacement. Across the five articles included in this meta-analysis, subjects took part in 16 distinct tasks. To maximize the number of tasks included in our between-task analyses, our minimal inclusion criteria consisted of all tasks that involved both water (displacement) and a binary choice. These criteria yielded a total of ten tasks: *large vs. small stones, water vs. sand, air vs. water, sink vs. float, baited vs. unbaited, U-tube, hollow vs. solid, narrow vs. wide equal, narrow vs. wide unequal, uncovered u-tube* (see Table 2 for a brief description of each task). In addition, we also conducted between-task analyses (including all subjects who participated in those tasks) including only those tasks in which learning was reported to occur: *large vs. small stones, water vs. sand, air vs. water, sink vs.*

Table 2 Brief description of tasks

Task name(s)	Objects	Apparatus	Description of measure
Basic water	1 type: similar sized stones	1 tube: water	Total number of stones dropped into the tube to retrieve food
Large vs. small stones	2 types: smaller stones and larger stones	1 tube: water	Tests which size stones Ss drop into the tube
Water vs. sand	1 type: similar sized stones	2 tubes: water and sand/woodchip/sawdust	Tests which tube Ss drop stones into
Air vs. water	1 type: similar sized stones	2 tubes: water and air (empty tube w/ bait suspended inside)	Tests which tube Ss drop stones into
Sink vs. float (heavy vs. light)	2 types: rubber and foam/polystyrene	1 tube: water	Tests which object Ss drop into tube
Baited vs. unbaited	1 type: similar sized stones	2 tubes: water baited with worm and water not baited	Tests which tube Ss drop stones into
U-tube	1 type: similar sized stones	2 tubes: "functional" tube connected to a water-baited third tube and "non-functional" tube not connected to water-baited third tube. Connections were hidden	Tests which tube Ss drop stones into
Hollow vs. solid	2 types: wire/metal frames and clay/metal cubes	1 tube: water	Tests which object Ss drop into tube
Narrow vs. wide equal	1 type: similar sized rubber/clay blocks*	2 tubes: narrow tube and wide tube; water level/bait was placed at the same height in both tubes	Tests which tube Ss drop stones into
Narrow vs. wide unequal	1 type: similar sized rubber/clay blocks*	2 tubes: narrow tube with low water level/bait and wide tube with high water level/bait	Tests which tube Ss drop stones into
Uncovered U-tube	1 type: similar sized stones	2 tubes: "functional" tube connected to a water-baited third tube and "non-functional" tube not connected to water-baited third tube. Connections were not hidden	Tests which tube Ss drop stones into

*Logan et al. (2014) used both clay and rubber objects in this task; Jelbert et al. (2014) used rubber blocks

Table 3 Intercept results of multilevel binary logistic regressions predicting probability of efficient option by insertion within each task

Tasks	Number of subjects	Intercept	Odds ratio (OR)	Confidence interval (CI)	*p*
Water vs. sand	22	−0.63	0.53	[0.27, 1.07]	.08
Float vs. sink	18	0.21	1.23	[0.17, 9.18]	.84
Hollow vs. solid	11	4.04	57.02	[10.18, 319.33]	< .001

float, baited vs. unbaited, hollow vs. solid, and *narrow vs. wide unequal*.

Table 4 Slope results of multilevel binary logistic regressions predicting probability of efficient choice by insertion within each task

Tasks	Number of subjects	Slope	Odds ratio (OR)	Confidence interval (CI)	*p*
Water vs. sand	22	0.04	1.04	[1.02, 1.06]	<.001
Float vs. sink	18	0.05	1.05	[1.03, 1.07]	<.001
Hollow vs. solid	11	0.01	1.01	[0.98, 1.03]	.67

Results

Characterization of subjects' overall learning

We used multilevel logistic modeling to characterize subjects' overall learning within each task. Analyses of the intercepts of these models (i.e., subjects' initial preference for the efficient or inefficient options at the onset of the task)

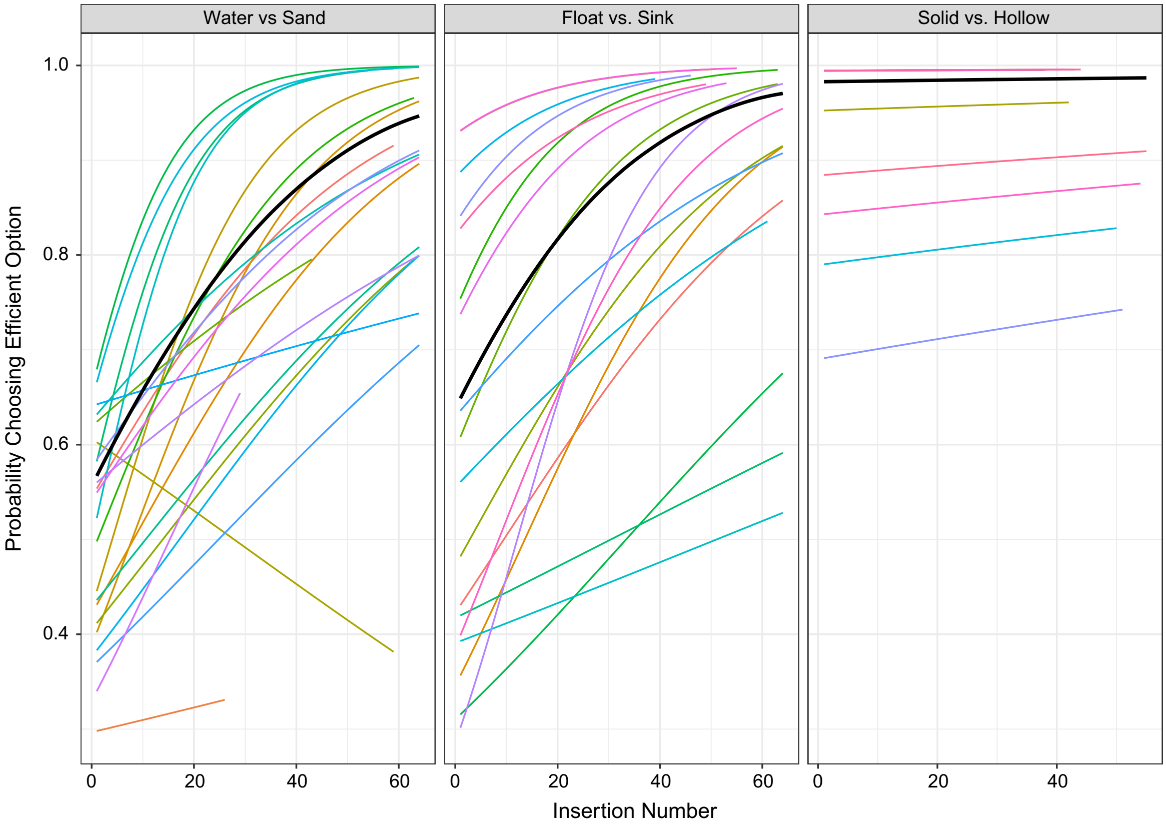

Fig. 1 Probability that subjects choose the more efficient option as a function of tube/object insertion order. Each color represents a different subject, and the solid black line depicts the overall relationship. The overall relationship is a weighted average of the different subjects and articles, with subjects and articles that have more data being weighed more heavily

are presented in Table 3. Analyses of the slopes of these models (i.e., the rate of learning) are presented in Table 4. Results are also depicted graphically in Fig. 1. Figures were created using the ggplot2 package version 2.2.1 (Wickham and Chang 2015). All analyses were conducted considering water as the efficient choice in *water vs. sand*, sink as the efficient choice in *float vs. sink*, and solid as the efficient choice in *hollow vs. solid*.

In the *water vs. sand* task, the best-fitting model allowed intercepts to vary at the article level and both intercepts and slopes to vary at the bird level. There was no covariance between intercepts and slopes at the bird level. In the *float vs. sink* task, the best-fitting model allowed intercepts to vary at the article level and both intercepts and slopes to vary at the bird level. There was no covariance between intercepts and slopes at the bird level. In the *hollow vs. solid* task, the best-fitting model allowed intercepts to vary at the article level and at the bird level.

Initial preference at task onset (intercept)

The intercepts indicate the initial preference subjects had for each tube or object at the onset of each task. If subjects were equally likely to choose the efficient or inefficient option, the odds ratio would be 1. In the *water vs. sand* task, the intercept was not statistically significant, suggesting that subjects did not prefer either the water or sand tube at the onset of the task ($p = .08$; the odds of subjects choosing the water tube were 0.53 times greater than the odds of subjects choosing the sand tube at task onset, 95% Wald Confidence Intervals [CI] = [0.27, 1.07]). Similarly, in the *float vs. sink* task, the intercept was not statistically significant, suggesting that subjects did not prefer either the sinking or floating objects at the onset of the task ($p = .84$; the odds of subjects choosing the sinking object were 1.23 times greater than the odds of subjects choosing the floating object at task onset, 95% CI = [0.17, 9.18]). However, in the *hollow vs. solid* task, the intercept was statistically significant, suggesting that subjects were more likely to choose the solid object over the hollow object at the onset of the task ($p < .001$; the odds of subjects choosing the solid object were 57.02 times greater than the odds of subjects choosing the hollow object at task onset, 95% CI = [10.18, 319.33]). This finding suggests that subjects had a robust preference for solid objects over hollow objects at the onset of the task, which likely influenced their performance on this task (see “General Discussion”).

In our parallel analyses that used the maximum number of insertions, the intercept for the *float vs. sink* task was statistically significant, suggesting that subjects had a preference for sinking objects over hollow objects at the onset of the task ($p = .03$). Although we made an a priori decision to base our interpretations on the analyses that use the median number of insertions, this maximum insertion analysis raises the possibility that subjects did have an a priori preference for the sinking object. If so, this finding would be in line with the *hollow vs. solid* task finding that subjects had a preference for solid objects at the onset of the task. This was the only significant difference between our primary analyses that used the median number of insertions and the parallel analyses that used the maximum number of insertions.

Rate of learning (slope)

The slope of the models represents the rate at which birds chose the efficient choice—in other words, the rate of learning within each task (see Table 4). In the *water vs. sand* task, the slope was statistically significant, suggesting that, with each stone insertion, subjects became more likely to choose the water tube over the sand tube ($p < .001$; the odds of subjects choosing the water tube were 1.04 times greater than the odds of subjects choosing the sand tube over the course of the task, 95% CI = [1.02, 1.06]). This finding indicates that, over the course of the task, subjects were learning to prefer dropping stones into water tubes over sand tubes.

Similarly, in the *float vs. sink* task, the slope was statistically significant, suggesting that, with each insertion, subjects became more likely to choose a sinking object over a floating object ($p < .001$; the odds of subjects choosing the sinking object were 1.05 times greater than the odds of subjects choosing the floating object over the course of the task, 95% CI = [1.03, 1.07]). This finding indicates that subjects learned to select sinking object over floating object over the course of the task.

In contrast, in the *hollow vs. solid* task, the slope was not statistically significant, suggesting that, with each insertion, subjects were not more likely to choose a solid object over a hollow object ($p = .67$; the odds of subjects choosing the solid object were 1.01 times greater than the odds of choosing the hollow object over the course of the task, 95% CI = [0.98, 1.03]). This finding indicates that subjects did not learn to select solid object over hollow object over the course of the task. This is likely because on every insertion—including the initial insertion—subjects had a very high (ceiling) rate of selecting the solid object (see Fig. 1).

Rate of learning (as a function of the previous choice)

In addition to examining the learning rate across insertions to determine whether birds showed an increase in choosing the efficient option as object insertions increased, we also analyzed whether birds in the *water vs. sand* and *float vs. sink* tasks learned from their choices at each insertion (see Table 5). Given that subjects in the *hollow vs. solid* task did not demonstrate learning over the course of that task, we excluded that task from these analyses. We used multilevel modeling to determine subjects' probability of selecting the efficient option (i.e., water tube/sinking object) when the inefficient option (i.e., sand tube/floating object) was chosen in the previous insertion compared to when the efficient option was chosen in the previous insertion. Again, we nested insertion order within subject and subject within article to account for dependencies and used likelihood ratio tests to find the best-fitting model. To ease interpretability, we calculated the expected odds when the inefficient option was previously chosen and when the efficient option was previously chosen, instead of reporting an intercept and slope. When presenting the results, we converted the logits into odds ratios.

In the *water vs. sand* task, the best-fitting model allowed intercepts and slopes to vary at both the article level and bird level. Intercepts and slopes covaried at the article level and the bird level. When the sand tube was the previous choice, subjects were not more likely to choose the water tube over the sand tube on the subsequent insertion ($p = .08$; the odds of subjects choosing the water tube were 0.67 times greater than the odds of subjects choosing the sand tube on the subsequent insertion, 95% CI = [0.43, 1.05]). However, when subjects previously chose the water tube, they were more likely to choose the water tube again on the subsequent insertion ($p < .001$; the odds of subjects choosing the water tube again were 5.0 times greater than the odds of subjects choosing the sand tube on the subsequent insertion, 95% CI = [2.66, 9.38]). The rates of learning according to whether the previous insertion was the water tube or the sand tube were also significantly different ($p < .001$). These findings indicate that subjects used information gained by their insertions into the water tube—but not the sand tube—to inform their subsequent insertions.

In the *float vs. sink* task, the best-fitting model allowed intercepts to vary at both the article level and bird level. Slopes also varied at the bird level, but did not covary with intercepts. When the floating object was the previous choice, subjects were not more likely to choose the sinking object

Table 5 Differences in odds ratio of making the efficient choice based on whether the previous choice was inefficient or efficient

Tasks	Odds ratio if previous inefficient	Odds ratio if previous efficient	Test of difference *p*
Water vs. sand	0.67	5.00	< .001
Float vs. sink	1.57	2.43	.01

over the floating object on the subsequent insertion ($p = .40$; the odds of subjects choosing the sinking object were 1.57 times, 95% CI = [0.55, 4.44] greater than the odds of subjects choosing the floating object on the subsequent insertion). In contrast, when the sinking object was the previous choice, subjects were significantly more likely to choose the sinking object again on the subsequent insertion ($p = .01$; the odds of subjects choosing the sinking object again were 2.43 times greater than the odds of subjects choosing the floating object on the subsequent insertion, 95% CI = [1.23, 4.80]). The rates of learning according to whether the previous insertion was the sinking or floating object were also significantly different ($p = .01$). Together, these findings indicate that subjects used information gained by their efficient choices (i.e., food moving closer when sinking objects were dropped into the water tube) to inform their subsequent insertions, but did not learn from inefficient choices (i.e., food remaining stationary when floating objects were dropped into the water tube).

Characterization of subjects' transfer of learning between tasks

The tasks and subjects included in the transfer effect analyses are found in Table 1. Within each article, subjects participated in multiple tasks. It is possible that learning to select the most efficient option in one task transferred to subsequent tasks, such that subjects learned more quickly to select the efficient option in latter tasks compared to former tasks. To investigate the possible transfer of learning across tasks, we tested whether the order in which subjects completed each task served as a significant predictor of rate of learning across tasks. The best-fitting model allowed intercepts, object insertion order, and task order to vary and covary at the article level. At the bird level, only intercepts and insertion order varied, and they did not covary.

Table 6 shows the fixed effects of the model predicting the effects of learning by task order. Fixed effects are the weighted average across all the subjects in all the articles—that is, they represent the overall effect of task order on efficient option choice for the average subject in all the articles. The slope was not statistically significant ($p = .22$), indicating that task order did not affect the rate at which the subjects learned to choose the efficient option for latter tasks (see Appendix 1 for coefficients of species level).

Table 6 Fixed effects of logistic model predicting odds of making the efficient choice across object insertion by task order

Coefficient	Odds ratio (OR)	Confidence interval (CI)	*z*	*p*
Intercept	1.53	[0.85, 2.76]	1.41	.16
Object insertion order	1.02	[1.01, 1.04]	4.20	< .001
Task order	0.81	[0.57, 1.14]	– 1.23	.22

Table 7 shows the random effects or variability of the coefficients across subjects. Random effects, assigned to each subject, represent how each subject deviates from the average subject. None of the variances are significant (*p*s > 0.94), indicating that there were no significant individual differences in subjects' rate of learning as a function of task order. We also conducted these analyses with task order as a binary variable, with 0 = first experiment and 1 = not the first experiment. Again, the order of task did not significantly predict the slope (rate of learning across the tasks).

As most birds did not successfully learn to retrieve the reward in the *narrow vs. wide equal, U-tube*, and *uncovered U-tube tasks*, we conducted between-task transfer of learning analyses that excluded these tasks. We posited that restricting our analyses to the tasks in which subjects successfully retrieved the food (*large vs. small stones, water vs. sand, air vs. water, sink vs. float, baited vs. unbaited, hollow vs. solid, and narrow vs. wide unequal)* would provide the strongest (and most conservative) test of transfer of learning across tasks. The best-fitting model allowed intercepts, object insertion order, and task order to vary but not covary at the article level. Intercept and experiment order varied and covaried at the bird level. Table 8 shows the fixed effects of the model predicting the effects of learning by task order. The slope was not significant ($p = .06$), indicating that task order did not affect subjects' learning. Table 9 shows the random effects across subjects. Variances were not significant (*p*s > 0.84), indicating that there were no significant individual differences in subjects' learning as a function of task

Table 7 Random effects across subjects (individual differences) predicting odds of making the efficient choice across object insertion by task order

Coefficient	Variance	Chi-square	*df*	*p*
Intercept	0.21	22.45	34	.94
Object insertion order	0.0002	15.76	34	1.00

Table 8 Fixed effects of multilevel model predicting odds of making the efficient choice across stone insertion by task order, excluding tasks in which subjects demonstrated no learning

Coefficient	Odds ratio (OR)	Confidence interval (CI)	*z*	*p*
Intercept	0.78	[0.42, 1.44]	− 0.80	.43
Object insertion order	1.03	[1.02, 1.04]	6.13	< .001
Task order	1.43	[0.99, 2.08]	1.90	.06

Table 9 Random effects across subjects (individual differences) predicting odds of making the efficient choice across stone insertion by task order, excluding tasks in which subjects demonstrated no learning

Coefficient	Variance	Chi-square	*df*	*p*
Intercept	0.58	19.73	27	0.84
Task order	0.78	0	27	1.00

order. Thus, even when restricting the analyses to the tasks in which subjects successfully retrieved the food, there is no evidence that subjects transferred what they learned in one task to subsequent tasks (see Appendix 1 for coefficients of species level).

General discussion

The results of our analyses of the main variant of the Aesop's fable paradigm (*water vs. sand,* $N=22$ subjects) allow for some clear initial conclusions. First, although the birds did not have an initial preference for one tube over the other, each successive stone drop was associated with an approximately 5% increase in the likelihood of choosing the water tube. Furthermore, after each stone drop into the water tube, birds were significantly more likely to choose that tube again on the very next drop. We detected no such learning effect when they had previously chosen the sand tube. These results suggest that the birds' learning was driven by perceptual feedback of the food moving incrementally closer to their beak each time that they dropped a stone in the water tube. The fact that the birds learned nothing from dropping stones in the sand is difficult to reconcile with the idea that they were reasoning about higher-order, role-based constructs such as mass, volume, or displacement. In any event, we conclude that their learning on the *water vs. sand* task can be completely explained by associative learning and/or perceptually based (first-order) relational reasoning.

As with the *sand vs. water* task, in the *float vs. sink* task, birds did not initially prefer either floating or sinking objects when they were required to choose one to drop into a single tube. However, after dropping a sinking object, birds were significantly more likely to select a sinking object on their next drop. These findings indicate that the birds used information gained by their insertions of sinking objects—but not floating objects—to inform their subsequent insertions. Together, these findings support the assertion that the birds' learned behavior in this variant of the task is again driven by the perceptual feedback of the food moving closer to their beak.

In contrast to the findings that birds did not prefer water tubes or solid objects at task outset, our meta-analysis revealed that the birds demonstrated a significant preference for solid objects over hollow objects at the outset of the *hollow vs. solid* task. Furthermore, the birds did not become more likely to choose solid objects as the task progressed (likely due to the fact that their initial preference for solid objects was near ceiling). Although no higher-order reasoning is needed to explain this result, it is unclear how the birds' behavior in this task connects to overarching aims of the Aesop's fable paradigm. That is, if corvids do have higher-order relational reasoning that allows them to represent constructs such as water displacement, they should also represent mass and volume, and thus have an a priori preference to choose solid over hollow objects (possibly related to their natural behavior of dropping heavy nuts on anvils in the wild; see Hunt 2014). Success on the first trial of all of these tasks would seem to be a plausible prediction if corvids did, indeed, utilize higher-order relational reasoning to complete the tasks. However, the Aesop's fable tasks are predicated on the assumption that the task is novel and subjects do not begin the tasks with a priori preferences for the functional options (see Bird and Emery 2009b; Jelbert et al. 2014; Logan et al. 2014; Taylor et al. 2011). This methodological catch-22 highlights a key problem inherent in these tasks that future researchers need to resolve if these types of empirical methods are to be used to advance the field of comparative cognition. We discuss this issue further below.

Our analyses also reveal no between-task transfer effects—that is, the order in which birds completed the tasks did not affect the rate at which they learned to select the efficient options as they gained experience with the Aesop's fable tasks. This lack of transfer across tasks is striking given that the birds demonstrated learning within at least some of the tasks (e.g., *water vs. sand; sink vs. float*), and that the perceptual properties, configurations, and goals appeared to be relatively similar across tasks (i.e., a binary choice between tubes or objects, where one choice more efficiently yields a reward). The lack of transfer across tasks raises the distinct possibility that the birds were not deploying a core strategy—such as reasoning about water displacement—to solve the tasks. Instead, the birds appeared to approach each task as a new and distinct problem to solve rather than slight variations on the same problem. (Of course, even if the subjects had shown some transfer between tasks, the difficulty of ascribing this to higher-order causal reasoning vs. learning about general adaption would be problematic.)

Our work is not the first to consider the types of information that subjects use to solve the Aesop's fable tasks. Jelbert et al. (2015) discussed two alternative models that could challenge the idea that higher-order causal reasoning is required for successful performance: perceptual-motor learning (i.e., first-order relational reasoning; Penn et al. 2008) and object biases (i.e., a priori preferences for objects with some specific perceptual characteristics vs. others). Ghirlanda and Lind (2017) also consider these possibilities, yet they suggest that small changes to the stimuli and methods can address them. Indeed, they state that, "every time a confound is suggested, an experiment can be designed to address it" (p. 247). However, that claim does not address the larger issue of underspecification inherent in the Aesop's fable tasks. That is, although success on the first trial is not a necessary condition of causal understanding, the question of what it means for subjects to have (or not have) an initial preference, and how task-based learning (including trial-and-error learning) connects to specific models of causal understanding, are of central importance to the methods and interpretations of these studies. Critically, these questions cannot be addressed via a series of control conditions and follow-up tasks.

Perceptual-motor learning (i.e., first-order relational reasoning; Penn et al. 2008) provides a model in which birds solve the task by choosing the tube in which a successful action (e.g., a stone drop) has previously served to bring the food closer to them, and then repeating that rewarded action until they have retrieved the food (Jelbert et al. 2015). In fact, it is hard to imagine how the subjects would not use the powerful information contained in such perceptual feedback. Jelbert et al. contend that the question of import is not whether perceptual feedback is used, but the extent to which subjects use it to solve the task. We agree. Unfortunately, the present analysis raises the thorny question of whether the Aesop's fable tasks could ever dissociate between perceptual-motor learning and higher-order, role-based reasoning. Several ways to test the perceptual feedback hypothesis have been suggested (e.g., Jelbert et al. 2015; Logan et al. 2014), but, because the causal power of these two functions is unknown, attempts to discriminate between them are likely to prove elusive.

For example, Jelbert et al. (2015) suggest blocking subjects' visual access to the movement of the food reward. However, as they note, "Subjects did not … typically succeed from the very first trial." (p. 2). Our meta-analysis provides further evidence that subjects did not choose the more efficient option at the onset of either the *water vs. sand* or *float vs. sink* task. Thus, the subjects appear to need some perceptual feedback to succeed at these tasks, and such feedback may be completely sufficient. Without a formal specification of the relative causal power of alternative learning functions, it is not possible to determine what function an organism may be using.

The second alternative explanation put forth by Jelbert and colleagues (2015) is that a priori preferences for objects with specific perceptual characteristics could lead subjects to succeed at the Aesop's fable tasks. When subjects were initially trained to drop stones into tubes, they might have developed a preference for those stones. When later presented with sinking and floating objects, subjects chose the sinking objects, as those were most similar to the sinking stones from their training sessions. Jelbert et al. argue that an a priori object bias can, therefore, account for birds' success in the *float vs. sink* object tasks, but *not* in tasks that involve functional and non-functional tubes (e.g., *water vs. sand*). We agree with Jelbert et al.'s former assessment, but disagree with the latter. If subjects have an a priori preference for the water tube—for example, a predisposition to prefer the visual characteristics of water over sand or prior negative experiences with sand—that a priori preference could drive them to interact more frequently with the water tube, biasing them towards the more efficient substrate. Our analyses suggest that birds did not have an initial preference for the water tube, but they did have a preference for the solid stones, and this preference for solid stones likely accounted for subjects' near-ceiling performance in the *hollow vs. solid* task.

To rule out an object-bias explanation, Jelbert et al. (2015) suggest that these biases could be ameliorated before the experimental tasks begin (e.g., the birds could be differentially reinforced for interacting with the less preferred option; Logan et al. 2014). We question whether attempts to induce unbiased neutral states through training can be a productive starting point. For example, imagine a bird with an a priori preference for heavy objects. If that bird is trained to use a light object on a specific task, does this imply that the bird has lost their preference for the heavy objects, or that the original preference will not bias the bird toward heavy objects in the subsequent experimental tasks?

Ghirlanda and Lind (2017) also contend that the previous reinforcement for stones similar to the solid objects in the *solid vs. hollow* contrast could have created a preference for solid objects that led to subjects' success in that task. They suggest that painting the solid objects to look less like stones and ascertaining subjects' predispositions to interact with solid over hollow objects are both solutions to this problem. These suggestions allow for a more nuanced understanding of the subjects' performance within the tasks, such as what perceptual cues are most salient and how predispositions interact with task-based learning. However, it is difficult to understand how they will allow us to discriminate between alternative learning functions (i.e., perceptual feedback vs. higher-order relational reasoning). Thus, although we agree with Ghirlanda and Lind about the importance of considering how predispositions, previous experiences, and trial-and-error learning interact when investigating causal reasoning, we do not see how the proposed task modifications could "yield better tests of causal cognition" (p. 239).

Although Jelbert et al. (2015) acknowledge that alternative explanations have not been ruled out, they conclude that, "across all these tasks, corvids were able to rapidly learn the most functional option, indicating that they appear to understand aspects of the causal nature of water displacement" (p. 2). They also note: "To understand the cognitive mechanisms that seemingly enable corvids to learn causal rules more effectively than arbitrary rules, future studies controlling for the object-bias hypothesis, and the perceptual-motor feedback hypothesis, will be highly informative." (p. 6). Similarly, Logan et al. (2014) describe a number of methodological problems associated with the paradigm, but maintain that minor methodological improvements will allow for "more powerful comparisons between humans and other animal species and thus help us to determine which aspects of causal cognition are distinct to humans" (p. 1). In contrast, we propose that, without a formal specification of the causal power of alternative learning models, these tasks are unsuited to discriminating between alternative meanings of "causal reasoning". Moreover, our analyses detect a pattern of learning that is consistent with associative learning and/or first-order relational reasoning.

We raise the question of whether many of the methodological practices that appear throughout these projects (e.g., reporting learning rates at the level of the trial instead of the individual object drop, attempting to control for subjects' preferences via differential rewarding, discounting the influence of prior experience, etc.) may result from underspecification of the constructs to be tested, and the alternative models against which the constructs in question are being tested. If "causal reasoning" is intended to be isomorphic with "goal-directed behavior", then the birds in the Aesop's fable studies can properly be regarded as engaging in "causal reasoning". A caution, however, is needed. Dropping stones into a water-filled tube with the intention of eliciting an observable effect (e.g., the water rising and the bait moving closer) may be structurally analogous to rats learning to press levers to dispense a food reward or primates using a stick to retrieve an out-of-reach food reward. On the other hand, if "causal reasoning" is restricted to entail higher-order, role-based reasoning about constructs such as mass, volume, and water displacement (see Penn et al. 2008; Povinelli 2011), the question arises as to whether the present tasks provide any evidence of such reasoning in corvids, or if they are in principle capable of doing so. Future research could begin with detailed specifications of the alternative constructs and/or learning processes under consideration, combined with the specification of why specific patterns of results are inconsistent with specific alternatives.

Compliance with ethical standards

Conflict of interest Laura Hennefield, Hyesung G. Hwang, Sara J. Weston, and Daniel J. Povinelli declare that they have no conflicts of interest.

Human and animal rights statement This article does not contain any studies with animals performed by any of the authors.

References

Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1–48. https://doi.org/10.18637/jss.v067.i01

Bird CD, Emery NJ (2009a) Insightful problem solving and creative tool modification by captive nontool-using rooks. Proc Natl Acad Sci 106:10370–10375. https://doi.org/10.1073/pnas.0901008106

Bird CD, Emery NJ (2009b) Rooks use stones to raise the water level to reach a floating worm. Curr Biol 19:1–5. https://doi.org/10.1016/j.cub.2009.07.033

Cheke LG, Bird CD, Clayton NS (2011) Tool-use and instrumental learning in the Eurasian jay (Garrulus glandarius). Anim Cogn 14:441–455. https://doi.org/10.1371/journal.pone.0040574

Cheng P (1997) From covariation to causation: a causal power theory. Psychol Rev 104:367–405

Domjan M (1983) Biological constraints on instrumental and classical conditioning: Implications for general process theory. Psychol Learn Motivat 17:215–277

Dunlap AS, Stephens DW (2014) Experimental evolution of prepared learning. Proc Natl Acad Sci 111:11750–11755. https://doi.org/10.1073/pnas.1404176111

Garcia J, Koelling RA (1966) Relation of cue to consequence in avoidance learning. Psychonom Sci 4:123–124

Ghirlanda S, Lind J (2017) 'Aesop's fable' experiments demonstrate trial-and-error leaning in birds, but no causal understanding. Anim Behav 123:239–247. https://doi.org/10.1016/j.anbehav.2016.10.029

Hunt GR (2014) Vice-anvil use in nut processing by two Corvus species. N Z Journal of Zool 41(1):68–76. https://doi.org/10.1080/03014223.2013.809368

Jelbert SA, Taylor AH, Cheke LG, Clayton NS, Gray RD (2014) Using the Aesop's Fable paradigm to investigate causal understanding of water displacement by New Caledonian crows. PLoS One 3:e92895. https://doi.org/10.1371/journal.pone.0092895

Jelbert SA, Taylor AH, Gray RD (2015) Investigating animal cognition with the Aesop's Fable paradigm: Current understanding and future directions. Communicat Integr Biol 8:e1035846. https://doi.org/10.1080/19420889.2015.1035846

Logan CJ (2016) Behavioral flexibility and problem solving in an invasive bird. Peer J 4:e1975. https://doi.org/10.7717/peerj.1975

Logan CJ, Jelbert SA, Breen AJ, Gray RD, Taylor AH (2014) Modifications to the Aesop's Fable paradigm change New Caledonian crow performances. PLoS One 7:e103049. https://doi.org/10.1371/journal.pone.0103049

Logan CJ, Harvey BD, Schlinger BA, Rensel M (2015) Western scrub-jays do not appear to attend to functionality in Aesop's Fable experiments. Peer J 4:e1707. https://doi.org/10.7717/peerj.1707

Penn DC, Povinelli DJ (2007) On the lack of evidence that chimpanzees possess anything remotely resembling a 'theory of mind'. Philosophical Trans R Soc B 362:731–744

Penn DC, Holyoak KJ, Povinelli DJ (2008) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behav Brain Sci 31:109–178

Povinelli DJ (2011) World without weight: Perspectives on an alien mind. Oxford University Press, Oxford

Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, vol 1. Sage Publications, Thousand Oaks

Rutz C, Sugasawa S, van der Wal JEM, Klump BC, St. Clair JJH (2016) Tool bending in New Caledonian crows. R Soc Open Sci. https://doi.org/10.1098/rsos.160439

Shettleworth SJ (1998) Cognition, evolution and behavior. Oxford University Press, Oxford

Shettleworth SJ (2009) The evolution of comparative cognition: is the snark still a boojum? Behav Process 80:210–217. https://doi.org/10.1016/j.beproc.2008.09.001

Shettleworth SJ (2012) Modularity, comparative cognition, and human uniqueness. Philos Trans R Soc B 367:2794–2802

Taylor AH (2014) Corvid cognition. Cognit Sci 5:361–372. https://doi.org/10.1002/wcs.1286

Taylor AH, Gray RD (2009) Animal cognition: Aesop's Fable flies from fiction to fact. Curr Biol 19:R713–R732. https://doi.org/10.1016/j.cub.2009.07.055

Taylor AH, Elliffe DM, Hunt GR, Emery NJ, Clayton NS, Gray RD (2011) New Caledonian crows learn the functional properties of novel tool types. PLoS ONE 12:e26887. https://doi.org/10.1371/journal.pone.0026887

Timberlake W (1993) Behavior systems and reinforcement: an integrative approach. J Exp Anal Behav 60:105–128

Tolman EC (1932) Purposive behavior in animals and men. The Century co, New York

Weir AAS, Chappell J, Kacelnik A (2002) Shaping of hooks in New Caledonian crows. Science 297:981. https://doi.org/10.1126/science.1073433

Wickham H, Chang W (2015) An implementation of the grammar of graphics. http://ggplot2.org. Accessed 14 Feb 2018