“How Helpful is this Observation?”: Children’s Evaluations of Scientific Evidence

Judith Danovitch (j.danovitch@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
347 Life Sciences, Louisville, KY 40205 USA

Candice Mills (candice.mills@utdallas.edu)
School of Behavioral and Brain Sciences, The University of Texas at Dallas
800 West Campbell Rd., Richardson, TX 75080 USA

Ravit Golan Duncan (ravit.duncan@gse.rutgers.edu)
Graduate School of Education, Rutgers University
10 Seminary Place, New Brunswick, NJ 08901 USA

Allison Williams (allison.williams.1@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
317 Life Sciences, Louisville, KY 40205 USA

Lauren Girouard (lauren.girouard@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
317 Life Sciences, Louisville, KY 40205 USA

Abstract

Being able to identify causally relevant evidence is essential in order to evaluate scientific claims, yet doing so can be challenging, especially for children. In some cases, identifying causally relevant evidence can involve recognizing similarities in context and causal mechanisms underlying seemingly different observations. Two studies explore how children ages 7-10 (n = 98) judge the relevance of different observations for evaluating the accuracy of a scientific explanation. Observations varied based on topic (i.e., the same animal as the explanation or a different species) and the presence or absence of the same underlying causal mechanism as the target explanation. All children recognized that observations involving the same process in the same animal would be helpful. However, children ages 7-8 held a more fragile understanding than children ages 9-10 that observations involving a different animal but the same causal mechanism would be more helpful than observations involving the same animal but a different causal mechanism. Implications for conceptual development and scientific reasoning are discussed.

Keywords: conceptual development; evidence; explanation; scientific reasoning

Introduction

Collecting evidence and using it to evaluate hypotheses or explanations is at the heart of the scientific process. For example, the explanation that bears hibernate in the winter because food is scarce could be supported by the observation that bears in the wild hibernate, while bears in captivity (where food is always available) do not. Being able to evaluate explanations in light of available evidence is important not only for scientists constructing theories and models (e.g., Haack, 2007), but also for children who are forming new concepts and learning about the world around them (Duncan, Chinn, & Barzilai, 2018; McNeill & Berland, 2017; Rinehart, Duncan, Chinn, Atkins, & DiBenedetti, 2016).

Causal explanations are a powerful means of supporting children’s understanding of a phenomenon (Keil, 2006; Legare, 2014), but simply receiving an explanation is not enough; children must also be able to evaluate whether an explanation is likely to be correct. One way they can do so is by judging whether the source of the explanation is reliable. As early as the preschool years, children take into account a source’s personal characteristics (e.g., age, area of expertise) and their prior history of accuracy when determining whether the information the source provides is likely to be valid (see Harris, 2012 for a review). Young children also consider how a claim fits with their existing knowledge (see Lane, 2018 for a review), to the point that they sometimes reject accurate claims that do not align with what they already know (e.g., Shtulman & Carey, 2007). By early elementary school, children can sometimes recognize when an explanation is weak or illogical (e.g., Mills, Danovitch, Rowles, & Campbell, 2017; Ruffman, 1999). However, as children gain exposure to more complex phenomena, they are likely to encounter explanations from reliable sources that seem reasonable in terms of their causal structure and fit with prior knowledge, but that may still be inadequate or incorrect, as is sometimes the case in the sciences.

Although children are capable of evaluating information sources and recognizing weak explanations, linking evidence with explanatory claims in different contexts may be particularly challenging (Sandoval, Sodian, Koerber, & Wong, 2014). Part of the challenge is that evidence can be complex and “messy” – it does not always perfectly align with the explanation at hand (Haack, 2007; Duncan et al., 2018). For example, eastern box turtles hibernate when food is scarce – does this support the explanation that bears' hibernation is related to food scarcity? Or is this observation irrelevant because turtles are a different species than bears? There are many instances in which animals of different species engage in similar behaviors, yet identifying the instances that are informative for building a strong explanation of a behavior requires looking past the animal involved and focusing on the underlying causal mechanisms instead. The current study explores how effectively elementary school age children identify whether evidence is relevant to an explanation. Deciding whether evidence is relevant as support for a particular explanation entails looking past superficial aspects of the context (e.g., animal species) and identifying the underlying causal process in order to determine its potential relevance.

Relevance is a key component of effective communication (Grice, 1975; Sperber & Wilson, 1996), and even young children seem to be sensitive to it when evaluating other people. Eskritt, Whalen, and Lee (2008) found that children ages 3-5 preferred to consult informants who had previously provided relevant information over those who had provided irrelevant information for help solving a simple problem. In fact, Eskritt et al. found that violations of the Gricean maxim of relevance seemed more salient to young children than violations of other Gricean maxims such as quality. More recently, Johnston, Sheskin, and Keil (2019) explored the development of children's ability to identify whether a piece of evidence (independent of an informant) was relevant for generating a causal explanation. In a series of experiments, children ages 4-8 and adults rated the helpfulness of observations about cars for understanding a mechanical process (i.e., how cars go). Children ages 4-6 had a rudimentary sense that some kinds of evidence (e.g., cars have engines that turn gasoline into power) were more helpful than other kinds of evidence (e.g., cars have radios that play music), yet not until ages 7-8 did children systematically recognize that causally relevant information was more helpful than causally irrelevant information for building a conceptual understanding. In Johnston et al.'s experiments, all of the evidence that participants evaluated involved the same target topic (e.g., cars). In reality, though, information about similar processes in other entities (e.g., trains have engines that turn fuel into power) can also be a powerful means of forming a causal understanding. Children tend to gravitate towards clustering information according to topics (e.g., Danovitch & Keil, 2004), but when it comes to understanding complex scientific processes, this strategy is not always effective: an off-topic piece of evidence involving a similar process or characteristic can sometimes be more causally relevant than an on-topic piece of evidence involving a different process or characteristic.

Study Design

Evaluating evidence is important not only for building new explanations or models, but also for evaluating the quality of a proposed or existing model. Here, we examine what kinds of evidence children consider helpful for judging the accuracy of an explanation of an unfamiliar animal behavior. We focus on explanations involving animals because animals are inherently interesting to children (Lobue, Bloom Pickard, Sherman, Axford, & Deloache, 2013), yet the biological mechanisms underlying their behavior are often complex and nonobvious. Because children may not be familiar with the terms "relevance" or "evidence," we frame the experimental task as a rating of the "helpfulness" of "observations" for determining whether a speaker's claim is correct. To ensure that children's ratings are not influenced by their own opinions about whether the evidence is accurate, children are informed up front that all of the evidence is true. In addition, the statements of evidence are described as observations made by groups of scientists to signal the presence of expertise and consensus, which should reinforce children's belief that the information is accurate (see Corriveau & Harris, 2010).

Our studies focus on children ages 7 to 10 for three reasons. First, children in this age range are capable of evaluating the strength of different kinds of single explanations, although their ability to do so is still improving (e.g., Mills et al., 2017). Second, children in this age range have experienced formal science instruction and are likely to be familiar with fundamental biological concepts such as survival tactics or reproductive mechanisms, yet their conceptual understanding may still be relatively weak (see Kelemen, 2019). Third, as their reading ability improves and they develop other information-seeking skills, children gain access to a wider variety of evidence in multiple formats. For example, elementary school age children can read about evidence in books or on websites, hear about evidence from other people, or gather evidence on their own through observation and experimentation, yet even children older than age 10 need substantial training and support in order to sort through different pieces of evidence and determine which ones are causally relevant to a proposed model of a biological process (Rinehart, et al., 2016; Duncan et al., 2018).

Study 1

Participants

A power analysis using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that, assuming a medium effect size with power = .95 and α = .05, an appropriate sample size would include a minimum of 18 children per age group. Twenty-four children ages 7 and 8 (M_{age} = 7.88; 13 males and 11 females) and 24 children ages 9 and 10 (M_{age} = 9.96; 13 males and 11 females) participated at their school or a university laboratory in the Louisville, Kentucky area. Parents identified 85 percent of the children in the sample as White, 2 percent as Black/African American, 4 percent as Asian American and 6 percent as more than one race (parents of 2 percent of participants chose not to indicate their child's race.) Ninety-four percent of the sample were identified as non-Hispanic, 4 percent were Hispanic, and the remaining 2 percent chose not to identify their ethnicity. Children were tested individually by a female experimenter in a session lasting approximately 15 minutes.

Rating Scale Design

The rating scale for the evidence evaluation task consisted of five sets of four concentric circles arranged in a bullseye pattern. Arrows were placed at regular intervals ranging from the exact center to the outside of each bullseye, and each bullseye had a corresponding label ranging from "not helpful at all" to "extremely helpful" (see Figure 1).

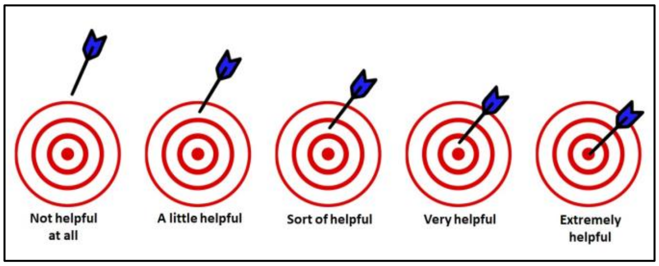

Figure 1: The helpfulness rating scale.

Procedure

Scale introduction Children were instructed that they would be hearing some people explain why animals behave the way they do, and that their job was to rate how helpful different clues would be for figuring out if each person's explanation was right or wrong. The experimenter then placed a laptop computer in front of the child, triggered the appearance of the "extremely helpful" bullseye on the right side of screen, and stated that "some observations are right on the mark, which means that they are extremely helpful for figuring out if an explanation is correct." The experimenter then proceeded to display the other points on the helpfulness rating scale and describe them in terms of helpfulness for determining if an explanation is correct.

Following the scale introduction, children completed three practice items in which they were given the following scenario: "Jane wants to go to the toy store today, but Jane's mom says that the toy store is closed today because it is a holiday." Children were then told about three observations attributed to other groups of people and illustrated with simple line drawings. The first observation involved causally relevant information ("Some people observed a sign last week that said the toy store is closed on holidays"), and the subsequent observations involved irrelevant information ("Some people observed that the toy store sells dolls, trucks, and LEGOs") and somewhat relevant information ("Some people observed that the supermarket next to the toy store is closed today"). Following each observation, children were asked to use the rating scale to indicate how helpful the observation would be for figuring out if Jane's mom was correct. Children received corrective feedback if their response did not correspond to the appropriate area of the scale.

Explanation evaluation task After familiarization with the scale, children were reminded that they would be hearing about animals and different people's explanations about what each animal does, followed by some things that scientists have observed, and that all of the observations were true and accurate. For each trial, the experimenter began by reading a statement about the animal out loud (e.g., "A snapping turtle is a kind of turtle") and presenting a color photo of the animal on the screen, with a brief description of the animal's behavior printed underneath (e.g., "When a snapping turtle is walking on land and a predator comes near, it will stand up on its hind legs"; note that the behavior was not depicted in the photo). The experimenter also read the brief description out loud. The experimenter then presented a new slide that had an image of a stick figure man or woman on the left and a large speech bubble on the right containing the text of the person's explanation. The experimenter narrated this slide by naming the individual and reading their explanation (e.g., "Sue says: The snapping turtle does this to make itself appear larger and prepare to lunge to scare off the predator.")

The presentation of the character's explanation was followed by four statements about scientists' observations, each of which was accompanied by an image intended to represent the group of scientists who made the observation. The image was a cartoon-style outline of four individuals of diverse ethnicities and genders dressed in white lab coats and holding a magnifying glass and a clipboard. To avoid influencing children's judgments, the image showed no facial features. The scientists in each image wore the same color shirts and pants, and their clothing colors varied between trials to illustrate that each observation had been made by a separate group of scientists.

For each animal behavior, children heard four corresponding observations: same animal/relevant, same animal/irrelevant, different animal/relevant, and different animal/irrelevant. Same animal/relevant observations involved information that directly related to the causally anticipated outcome in the proposed explanation (e.g., scientists have observed that predators who see a snapping turtle standing up on its legs get scared and run away). Same animal/irrelevant observations involved behaviors that had no connection to the explanation (e.g., scientists have observed that a snapping turtle prefers to live in lakes, rivers, or streams that have a muddy bottom). The different animal observations described the same member of a different species and either involved a behavior relevant to the causally anticipated outcome of the proposed explanation (e.g., scientists have observed that predators who see a blowfish who has expanded to twice its usual size will swim away) or a causally irrelevant behavior (e.g., scientists have observed that a blowfish will swim closer to the surface of the ocean when it is looking for a mate). After the experimenter stated each observation, children viewed the rating scale on the screen and were prompted to indicate how helpful that observation was for figuring out if the character's explanation of the target behavior was correct.

Children completed six sets of four trials, where each set involved a different animal and four subsequent observations. The target animal behaviors were presented in one of two random orders, and the order in which the observations were presented was also randomized and counter-balanced between participants.

Results and Discussion

Children's responses on the evidence evaluation task were converted to scores of 0 to 4, where higher scores correspond to ratings of greater helpfulness. Preliminary analyses suggested no effects of participant gender or trial order so these factors were excluded from further analyses.

A 4 (Observation type) × 2 (Age group: 7-8-year-olds, 9-10-year-olds) mixed-measures ANOVA was conducted on the average ratings. Using the Greenhouse-Geisser correction for inhomogeneity of variance, the ANOVA revealed a main effect of Observation type, *F*(1.92, 88.31) = 37.06, *p* < .001, partial η^2= .458. Children's mean helpfulness ratings were highest for same animal/relevant observations (*M* = 2.94, SD = .47) and lowest for the different animal/irrelevant observations (*M* = 1.30, SD = 1.28; see Figure 2). Post-hoc pairwise comparisons revealed significant differences between children's ratings for all four types of observations, *p* ≤ .004, except for between the different animal/relevant observation ratings compared with the same animal/irrelevant observation ratings, *p* = .154.

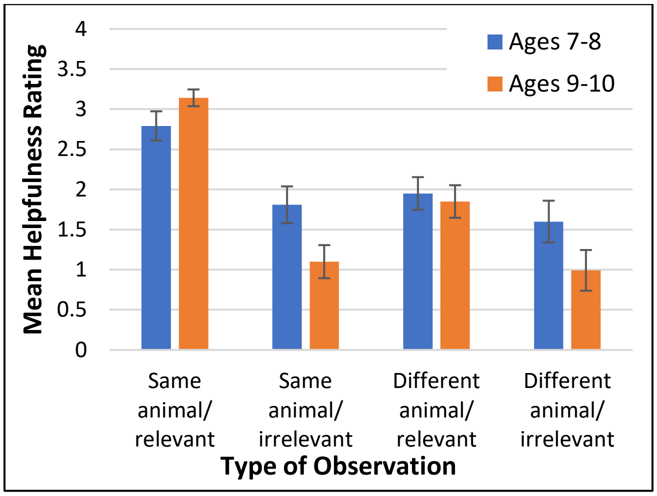

Figure 2: Mean helpfulness ratings for each type of observation by younger and older children. (Error bars indicate standard error.)

There was no significant main effect of Age group, *F*(1, 46) = 1.60, *p* = .24, partial η^2= .034, but there was a significant interaction between Observation type and Age group, *F*(1.92, 88.31) = 4.11, *p* = .021, partial η^2= .082. Post-hoc t-tests showed that older children assigned lower ratings than younger children to same animal/irrelevant observations, *t*(46) = 2.30, *p* = .026, but otherwise ratings for each type of observation did not significantly differ between age groups, *ts* < 1.67, *ps* > .100. Further supporting this finding, older children assigned ratings to the same animal/irrelevant observations that were significantly lower than the midpoint of the scale (rating of 2 or "sort of helpful"), *t*(23) = 4.27, *p* < .001, but younger children's ratings did not differ from the midpoint, *t*(23) = .83, *p* = .413.

Within each age group, planned t-tests showed that 7- and 8-year-olds gave significantly higher ratings to same animal/relevant observations than to the other three types of observations, *ts* ≥ 2.97, *ps* ≤ .007, but otherwise their ratings for the other types of observation did not significantly differ from each other, *ts* ≤ 1.76, *ps* ≥ .092. Nine- and 10-year-olds ratings significantly differed between all observation types, *ts* ≥ 4.47, *ps* ≤ .001, except for between the same animal/irrelevant and different animal/irrelevant observations, *t*(23) = .891, *p* = .382.

Overall, children in both age groups recognized that observations of the same animal that support the causal outcome in the proposed explanation (e.g., observing predators leaving) would be very helpful for evaluating the explanation's accuracy, and that observations of causally irrelevant behaviors in a different animal would be less helpful. However, when evaluating observations about the same animal that were not causally relevant to the explanation (i.e., because they involved a different biological process), older children assigned lower helpfulness ratings than younger children. Older children might have been better able to look past the presence of the same animal topic word (e.g., "snapping turtle") in the same animal/irrelevant observation and realize that the observation involved a different biological process or causal mechanism. In addition, older children also seemed more sensitive than younger children to the fact that the different animal/relevant explanation was more helpful than either type of causally irrelevant explanation.

The differences between older children's and younger children's responses on the same animal/irrelevant items might also reflect a difference in calibration when using the helpfulness rating scale. In making their ratings, younger children may have believed that all observations that were about the same animal and that involved the causal outcome proposed in the explanation would be very helpful, and, thus, any observations that did not meet *both* of these criteria would only be somewhat helpful. Study 2 was designed to explore this possibility.

Study 2

In order to ensure that children's ratings of the observations were not influenced by the presence of much stronger evidence (e.g., the same animal/relevant observations) or much weaker evidence (e.g., the different animal/irrelevant observations), these trials were omitted from Study 2. Additionally, four trials were added after the evaluation task in order to check whether children were using the helpfulness rating scale appropriately.

Participants

Participants were 24 children ages 7 and 8 (M_{age} = 8.05; 12 males and 12 females) and 26 children ages 9 and 10 (M_{age} = 9.94; 13 males and 13 females) who had not participated in Study 1. Parents identified 84 percent of the children in the sample as White, 4 percent as Black/African American, 2 percent as Asian American, 2 percent as Native Hawaiian and 4 percent as more than one race (parents of 4 percent of children chose not to indicate their child's race). Eighty-eight percent of the sample were identified as non-Hispanic, 8 percent were Hispanic, and the remaining 4 percent chose not to identify their ethnicity. Children were tested individually in a university laboratory in the Louisville, Kentucky area.

Procedure

Scale introduction and evidence evaluation task Children were introduced to the rating scale and the evidence evaluation task in the exact same manner as Study 1. The evidence evaluation task used the same animal stimuli presented in the same two randomized orders as Study 1; however, the same animal/relevant and different/animal irrelevant observations were omitted, so there were only two trials per animal.

Check questions Following the evidence evaluation task, children heard about two additional familiar animal behaviors (a squirrel burying nuts in the ground, a monkey carrying bananas into a tree). After each behavior was introduced with an accompanying photo, a character offered an explanation for the behavior (e.g., the squirrel does this to save food for the winter). Children then rated two observations in terms of helpfulness for evaluating the explanation: one that was about the same animal and causally relevant to the explanation (e.g., scientists have observed that squirrels dig up nuts from the ground to eat them during the winter) and one that was about a different animal and was not causally relevant to the explanation (e.g., scientists have observed that beavers build dams across rivers using sticks and trees).

Results and Discussion

As in Study 1, children's responses on the evidence evaluation task were converted to scores of 0 to 4. Preliminary analyses suggested no effects of participant gender or trial order. Thus, these factors were excluded from further analyses.

A 2 (Observation type: same animal/irrelevant, different animal/relevant) × 2 (Age group: 7-8-year-olds, 9-10-year-olds) mixed-measures ANOVA was conducted on the mean helpfulness ratings. The ANOVA revealed a main effect of Observation type, $F(1, 48) = 30.27$, $p < .001$, partial $\eta^2 = .387$. Children rated the different animal/relevant observations more highly (M = 1.95, SD = .99) than the same animal/irrelevant observations (M = 1.22, SD = 1.13; see Figure 3). There was also a significant main effect of Age group, $F(1, 48) = 4.93$, $p = .031$, partial $\eta^2 = .093$, where younger children (M = 1.88, SD = .19) gave higher ratings than older children (M = 1.31, SD = .18). There was no significant interaction of Observation type X Age group, $F(1, 48) = .39$, $p = .538$, partial $\eta^2 = .008$.

Planned t-tests showed that both age groups gave higher ratings to the different animal/relevant observations than the same animal/irrelevant observations, $ts \geq 2.99$, $ps \leq .007$. Compared with the results of Study 1, these findings suggest that younger children were better able to recognize the relative helpfulness of different animal/relevant observations over same animal/irrelevant observations when they were faced with a more limited set of observations that did not include observations that were stronger on both dimensions (i.e., same animal/relevant) or weaker on both dimensions (i.e., different animal/irrelevant).

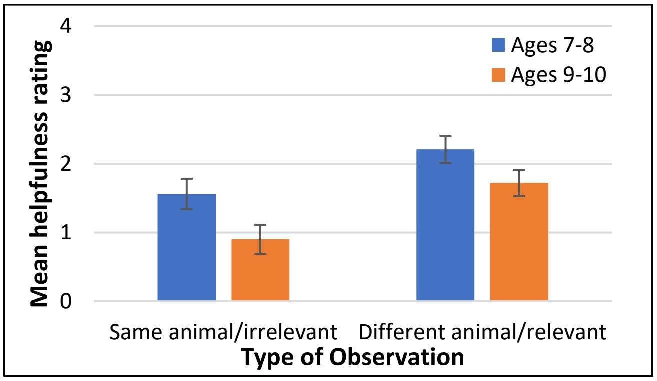

Figure 3: Mean helpfulness ratings for each type of observation by younger and older children. (Error bars indicate standard error.)

Check items Overall, children gave high helpfulness ratings to the same animal/relevant check items (M = 3.34, SD = .89) and low helpfulness ratings to the different animal/irrelevant check items (M = .82, SD = 1.31), suggesting that they were using the rating scale appropriately. These ratings were much higher and much lower, respectively, than the ratings for the same animal/irrelevant items and different animal/relevant items in the preceding evidence evaluation task.

General Discussion

Two studies examined children's judgments of the helpfulness of different types of evidence for determining whether a proposed causal explanation was accurate. When the evidence involved the same animal along with the same causal mechanism as the proposed explanation, children ages 7-10 consistently deemed it very helpful. Children also rated observations involving the same causal mechanism as the target explanation as relatively more helpful than observations involving an unrelated process. However, when the evidence involved the same animal and a causal mechanism that was unrelated to the proposed explanation, 7- and 8-year-olds gave it higher helpfulness ratings than 9- and 10-year-olds did.

As suggested by Study 2, the disparity between younger and older children's helpfulness ratings for the same animal/irrelevant items did not seem to be related to differences in understanding or using the rating scale. Rather, younger children may have found it challenging to distinguish between evidence that did not involve the same animal and the same causal mechanisms when their ratings were anchored by clearly superior or inferior observations. These findings may seem surprising given that 7- and 8-year-olds were capable of distinguishing between causally relevant and irrelevant evidence about the same topic in Johnston et al.'s (2019) studies. In Johnston et al.'s studies, children evaluated how helpful pieces of evidence would be for generating an explanation of a mechanical process, but they did not have to actually develop or evaluate an explanation, which may have allowed them to be fairly selective about the evidence they chose. In contrast, in the current studies, children evaluated how helpful each piece of evidence was for determining if an explanation *that already existed* was accurate. Seven- and 8-year-olds might have found the latter task more challenging as it required recognizing the underlying causal process involved in the proposed explanation and then identifying the similarities between it and the causal processes involved in each observation. Thus, although 7- and 8-year-olds are capable of evaluating evidence to some extent (i.e., they rated the same animal/irrelevant observations lower than the same animal/relevant ones in Study 1, and they rated the same animal/irrelevant observations lower than different animal/relevant ones in Study 2), they may still be less adept than their older counterparts at evaluating evidence when faced with multiple possible mechanisms and multiple types of observations.

We also found that younger children gave higher ratings to the three weaker types of evidence than older children, and that younger children's ratings hovered around the midpoint of the scale (corresponding to "sort of helpful") while older children's ratings were often below the midpoint (closer to the "not helpful at all" end of the scale). This pattern of results might reflect a positivity bias among younger children that influenced them to label all types of observations as somewhat helpful (see Boseovski, 2010). Younger children may also have been influenced by the fact that each observation was attributed to a group of scientists, and they may have been reluctant to reject such "scientific evidence" as unhelpful. Future research could explore this possibility by varying the source of the evidence and the number of people described as having made the observation. It would also be informative to see if younger children assign lower ratings to evidence involving entirely different types of information (e.g., physical or chemical reactions), rather than information involving a different, yet still somewhat taxonomically-related entity (e.g., a different animal).

Taken together, the current findings suggest that elementary school age children recognize when a piece of evidence is clearly relevant to a proposed causal explanation, but their ability to recognize that evidence involving the same entity as the explanation can sometimes be causally irrelevant may still be somewhat fragile or inconsistent. Children in this age range are also capable of recognizing when information about a different entity is causally relevant to an explanation, although younger children may find this more challenging than older children. These data suggest that the underpinnings of the ability to critically evaluate evidence in relation to a proposed hypothesis or explanation are present by age 7 or 8, yet elementary school age children would still benefit from instruction in how to evaluate disparate types of evidence. For example, providing a chart that children can use to visually represent the relevance of different pieces of evidence to a target explanation (as in Rinehart et al., 2016) might enable children to more effectively compare observations that vary on multiple dimensions. Ultimately, science educators should keep in mind that although understanding how complex and indirect pieces of evidence inform explanations is essential for successful scientific reasoning, children may require substantial time and experience to develop this skill.

Acknowledgments

The authors thank members of the University of Louisville KID Lab for their assistance with data collection and entry, as well as the staff and students at St. Gabriel the Archangel School for their support.

This work was supported by National Science Foundation Grant DRL-1551862, awarded to Judith Danovitch, and DRL-1551795, awarded to Candice Mills.

References

Danovitch, J. H., & Keil, F. (2004). Should you ask a fisherman or a biologist?: Developmental shifts in ways of clustering knowledge. *Child Development*, *75*, 918-931.

Duncan, R. G., Chinn, C. A., Barzilai, S. (2018). Grasp of evidence: Problematizing and expanding the Next Generation Science Standards' conceptualization of evidence. *Journal of Research in Science Education*, *55*, 907-937.

Eskritt, M., Whalen, J., & Lee, K. (2008). Preschoolers can recognize violations of the Gricean maxims. *British Journal of Developmental Psychology*, *26*, 435-443.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts Vol. 3*. New York, NY: Academic Press.

Haack, S. (2007). *Defending science: Between scientism and cynicism*. Amherst, NY: Prometheus Books.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Cambridge, MA: Harvard University Press.

Johnston, A. M., Sheskin, M., & Keil, F. C. (2019). Learning the relevance of relevance and the trouble with truth: Evaluating explanatory relevance across childhood. *Journal of Cognition and Development*, *20*, 555-572.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227-254.

Kelemen, D. (2019). The magic of mechanism: Explanation-based instruction on counterintuitive concepts in early childhood. *Perspectives on Psychological Science*, *14*, 510-522.

Lane, J. D. (2018). Children's belief in counterintuitive and counterperceptual messages. *Child Development Perspectives*, *12*, 247-252.

Legare, C. H. (2014). The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives, 8*, 101-106.

LoBue, V., Bloom Pickard, M., Sherman, K., Axford, C., & DeLoache, J. S. (2013). Young children's interest in live animals. *British Journal of Developmental Psychology*, *31*, 57-69.

McNeill, K. L., & Berland, L. (2017). What is (or should be) scientific evidence use in K-12 classrooms? *Journal of Research in Science Teaching*, *54*, 672-689.

Mills, C. M., Danovitch, J. H., Rowles, S. P., & Campbell, I. L. (2017). Children’s success at detecting circular explanations and their interest in future learning. *Psychonomic Bulletin and Review*, *24*, 1465-1477.

Rinehart, R. W., Duncan, R. G., & Chinn, C. A. (2014). A scaffolding suite to support evidence-based modeling and argumentation. *Science Scope*, *38*, 70-77.

Rinehart, R. W., Duncan, R. G., & Chinn, C. A., Atkins, T., & DiBenedetti, J. (2016). Critical design decisions for successful model-based inquiry in science classrooms. International *Journal of Designs for Learning*, *7*, 17–40.

Ruffman, T. (1999). Children's understanding of logical inconsistency. *Child Development*, *70*, 872-886.

Sandoval, W. A., Sodian, B., Koerber, S., & Wong, J. (2014). Developing children's early competencies to engage with science. *Educational Psychologist*, *49*, 139-152.

Shtulman, A., & Carey, S. (2007). Improbable or impossible? How children reason about the possibility of extraordinary events. *Child Development*, *78*, 1015-1032

Sperber, D., & Wilson, D. (1996). *Relevance: Communication and Cognition.* Cambridge, MA: Harvard University Press.