Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
J. Experimental Child Psychology 87 (2004) 1–32
Journal of Experimental Child Psychology
www.elsevier.com/locate/jecp

Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth

Candice M. Mills* and Frank C. Keil

Yale University, P.O. Box 208205, New Haven, CT 06520-8205, USA

Received 9 June 2003; revised 26 September 2003

Abstract

Adults overestimate the detail and depth of their explanatory knowledge, but through providing explanations they recognize their initial illusion of understanding. By contrast, they are much more accurate in making self-assessments for other kinds of knowledge, such as for procedures, narratives, and facts. Two studies examined this *illusion of explanatory depth* with 48 children each in grades K, 2, and 4, and also explored adults' ratings of the children's explanations. Children judged their understanding of mechanical devices (Study 1) and procedures (Study 2). Second and fourth graders showed a clear illusion of explanatory depth for devices, recognizing the inaccuracy of their initial impressions after providing explanations. The illusion did not occur for knowledge of procedures.

Keywords: Metacognition; Explanation; Understanding; Knowledge; Cognitive development

Introduction

As adults, we have all encountered situations where we thought we understood how something worked or why a phenomenon occurred only to be caught up short by a question that illustrated a huge gap in understanding. Despite our initial intuitions, it becomes apparent in these situations that our assessments of our own explanatory

* Corresponding author. Fax: 1-203-436-1546.
E-mail address: candice.mills@yale.edu (C.M. Mills).

0022-0965/$ - see front matter
doi:10.1016/j.jecp.2003.09.003

knowledge are not always well founded. This report examines how self-assessments of our naíve explanatory understandings develop in the elementary school years and how they contrast with self-assessments of other kinds of knowledge. We argue that insight into one's own explanatory prowess is governed by distinctive factors that lead to a different pattern of self-assessments than other kinds of knowledge, and that even young children are influenced by these features when judging their knowledge.

Hundreds of studies have explored children's knowledge about their own memory processes. In general, research has shown that children are poor at predicting their own performance in cognitive tasks, most often overestimating the level of their comprehension or abilities. In pioneering work on metacognitive development, for instance, large mismatches occurred between children's estimates of their ability to recall a sequence of serially presented pictures and their actual ability to recall. In some cases, preschoolers and kindergartners are confident about being able to recall well more than a dozen items while actually recalling only two or three (Flavell, Freidrichs, & Hoyt, 1970). The younger children are, the more they tend to overestimate their own memory for lists of items (Schneider & Pressley, 1997); second and fourth graders tend to be much more accurate than younger children, although their performance estimations are still not perfect (Flavell et al., 1970).

With more complex assessments involving comprehension of information, similar effects emerge. For example, children under 12 years old are often inaccurate when monitoring their own understanding of text passages; they think they have grasped a great deal more from a passage than they really have (Markman, 1977, 1979). Similarly, in the *judgment of learning* paradigm, children overpredict their own performance after studying materials. Second graders and younger children are more prone to being less accurate in both pretest and posttest estimations of test performance than older children; they tend to overestimate the level of their performance (Pressley, Levin, Ghatala, & Ahmad, 1987).

The judgment of learning work, however, is limited by its focus on predictions for recently learned information, such as for pairs of words and pictures or lists of items. In many other contexts, children must also reflect on their own understanding of more complex, long-standing knowledge that they already have. A child might be asked to estimate her understanding of how the sun disappears at night or how to make cookies—knowledge that may have been gradually built up over several years. How accurate is she in her initial estimates of how much she knows, and to what extent can she become aware of the gaps in her knowledge? To what extent do children have an appropriate sense of the depth and detail of their own intuitive theories of the world around them?

Intuitive or folk theories may underlie our understandings of natural phenomena (such as how the sun disappears at night), biological phenomena (such as how growth occurs), or mechanical processes (such as a how a crossbow works). Folk theories have been argued to be present throughout much of development; from intuitive theories in infancy (Spelke, Breinlinger, Macomber, & Jacobson, 1992) and childhood (Gelman & Koenig, 2003), to our own adult theories in specific domains such as biology or psychology, these theories are seen as helping us interpret causal relation and engage in conceptual change.

In reality, though, folk theories are often skeletal and incomplete (Wilson & Keil, 1998), and yet people usually remain unaware of the inconsistencies and holes in their own theories (Ahn & Kalish, 2000; Dunbar, 1995). Recent studies suggest that although adults may think they know how things work far better than they actually do (i.e., mechanical devices or biological phenomena), attempting to provide explanations gives them insight into their knowledge, and they become aware of the gaps and lack of detail in their theories (Rozenblit & Keil, 2002).

These studies used a successive rating technique to explore participants' perceptions of their own understanding of various phenomena: adults were trained on a 7-point scale to indicate whether they had a deep, partial, or shallow understanding of a particular phenomenon (such as how a crossbow works or how earthquakes occur), and they used this rating scale through a variety of tasks designed to require subsequently closer examination of their own knowledge. Participants were asked to use this scale to rate their level of understanding for a variety of phenomena within a particular domain. Then, they were asked to explain a subset of the phenomena, rerating their initial level of understanding after providing the explanations. Further reratings of their knowledge occurred after other manipulations that varied across tasks.

For example, in one study, adults were asked to use the scale to rate their understanding of each of 48 items. The participants were then asked to write a step-by-step explanation for each of four test phenomena, such as how a cylinder lock works. After providing each explanation, participants rerated how well they understood that phenomenon. Next, they answered a diagnostic question about each of the four phenomena. Diagnostic questions were designed to require knowledge about the mechanism of each phenomenon to provide a response. In the cylinder lock example, participants were asked to explain how to pick a cylinder lock. After answering the diagnostic questions, participants once again rerated how well they understood the phenomena. Finally, participants read an "expert" explanation of the four test phenomena, and rerated their prior level of understanding compared with this expert explanation.

Each step of these studies probed adults' perceptions of their own long-term knowledge in more detail, which led them to reconsider and adjust their self-ratings downward. After providing an explanation for a phenomenon, participants tended to drop their ratings of their initial knowledge for that item. They recognized that gaps existed in their understanding, often very large ones, and that they had initially overestimated the detail and depth of their own knowledge. Through providing explanations, they realized that their understandings of the phenomena were both superficial and missing important information (superficial understanding is almost always correlated with large gaps in understanding). For example, a participant might know that entering the right key with a certain shape into a cylindrical lock turns a latch, but when explaining how this might work, the participant recognizes a lack of awareness of the mechanism. This effect is known as the *illusion of explanatory depth*: the impression that one's own explanatory knowledge is deeper than it actually is (Rozenblit & Keil, 2002).

If the illusion were just an issue of overconfidence, then one would expect that adults would exhibit this illusion regardless of the type of knowledge. However, adults do not miscalibrate with other types of knowledge or do so to a much lesser extent. The illusion of explanatory depth occurs in knowledge areas that involve complex causal patterns (such as mechanical devices and biological processes), knowledge that can be described as explanatory or theory-like. For other kinds of knowledge (such as procedural or narrative), adults are more accurate in their original assessments. For example, when asked to evaluate their knowledge about how to make chocolate chip cookies from scratch, or about the plot of the movie *Forrest Gump*, participants showed much smaller or nonexistent drops. The contrast between explanatory and other kinds of knowledge can even be seen for the same item; for instance, the explanation of the mechanism behind how a toaster works is a fundamentally different question from the procedure of how to use the toaster. This effect was not driven by familiarity or embarrassment over ignorance for the information (see Rozenblit & Keil, 2002, for more information); rather, the effect seemed to be most influenced by the properties of the type of knowledge. Additionally, contrary to classic ability confidence estimates in which people are underconfident for easy tasks and overconfident for difficult tasks (Lichtenstein & Fischhoff, 1977), across a wide range of difficulties the illusion of explanatory depth goes in the same direction, namely, a tendency to think one understands things, whether relatively complex or relatively simple, in more detail than one really does (Rozenblit & Keil, 2002).

Several factors may converge to create an especially strong illusion for explanations compared with other kinds of knowledge. First, explanatory knowledge is more causally complex and less clearly defined than most other types of knowledge. Explanatory knowledge is heavily layered: there are deeper and deeper levels at which something can be understood, and therefore people may initially confuse their insights at a broad high level of understanding with a more detailed mechanistic understanding. When asked to explain a cylindrical lock, for example, one may confuse an understanding that key rotation opens a latch with a deeper understanding of how pins are raised by the key into an aligned position. It is also difficult to know in advance what a successful explanation will look like and thereby know if one has achieved success.

These factors are less important for other kinds of knowledge, such as for knowledge of procedures. Knowledge of procedures is less layered, and the end states are relatively clear. When explaining how to make a cheese pizza, for example, one knows that the end state is the cooked cheese pizza, and the steps to arriving at the end state are apparent. Knowing beyond those main steps (how to make cheese or how the oven works) is unnecessary for knowledge of procedures; in fact, it is more efficient simply to focus on the steps at that main level.

Second, in evaluating our knowledge about mechanical devices, for instance, we often underestimate how much we rely on environmental support. If adults or children are explaining how a device works while looking at it, they can infer causal relations from how the features of the device connect together. However, neither adults nor children may realize how much more difficult it may be to

provide an explanation without such environmental support. This effect may be analogous to one noted in the change blindness literature, in which people forget how much information they normally recover by looking at a scene again and thus overestimate how much they will remember, a phenomenon known as "change blindness blindness" (Levin, Momen, Drivdahl, & Simons, 2000). For knowledge of procedures, in contrast, environmental support is not as important. When making a cheese pizza, seeing the ingredients sitting on the counter does not provide as much information on the relation between the ingredients and the final product.

Finally, another difference between explanatory knowledge and other types of knowledge is that people might have less experience in self-testing their explanatory knowledge. To be able to perform a procedure, one must be able to walk through the steps, and even children undoubtedly have some experience with doing this. It is also possible to examine one's memory as to whether one has performed the procedure in the past. With explanatory knowledge, people rarely have to give full explanations, and as a result they may remain unaware of their own shallow understandings and have little in the way of past performance to compare against.

These factors converge to create an initial illusion of the depth of one's explanatory knowledge; to the extent that other kinds of knowledge share these features, an illusion of depth may result in other areas as well. For example, historical events may be considered causally complex, but some people may have more experience explaining them; art criticism, in contrast, may not be as causally complex, but people may not often reflect on their knowledge of art. There may well be individual differences in people's experiences with different domains that could influence their initial assessments of their knowledge, but the key is that when certain factors converge, people are likely to suffer from an illusion of depth. From research so far, explanatory knowledge seems to invoke the strongest illusion. Indeed, using a wide range of examples for several other categories of knowledge, knowledge of procedures, knowledge of facts, and knowledge of narratives, the illusion is either not present or much smaller in magnitude (Rozenblit & Keil, 2002).

This illusion of depth is an effect above and beyond general overconfidence; people are not *just* overestimating their knowledge in general. Instead, they are sensitive to properties of the type of knowledge when making judgments about their own understanding. Explanatory knowledge generally leads people to initially overestimate the depth of their knowledge, and the act of providing an explanation leads them to recognize both the gaps in their knowledge and their superficial level of understanding. Supporting this idea, when independent judges rate the quality of the explanations offered by other participants in these studies, the judges' ratings are closer to matching participants' postexplanation ratings compared with the initial ratings. Thus, participants' ratings were indeed more accurate after providing an explanation. By contrast, judges' ratings of the quality of descriptions of procedures accord with people's initial ratings of their own knowledge of procedures (Rozenblit & Keil, 2002), showing that in this case, participants are right on target when initially estimating their knowledge.

These studies with adults raised two questions for developmental research. First, when do children become aware of the inadequacies of their own knowledge? Given the findings from previous research on metacognitive development, we proposed that children would likely overestimate the depth of their own knowledge, with younger children most strongly overestimating their own knowledge. The factors that influence adults' judgments of their own knowledge (such as causal complexity, amount of environmental support, and experience providing explanations or descriptions) are ones that should influence children as well, persuading them to overestimate their knowledge. However, we thought that in providing explanations, at least the older children would be likely to recognize their illusion; other research has found that older children are capable of realizing when their performance on a certain task was not as accurate as they had initially predicted (Pressley et al., 1987).

Second, how might the illusion of knowing vary across kinds of knowledge? With adults, we know that people are selectively miscalibrated for explanatory understanding. We predicted that once children are able to recognize the inadequacy of their own knowledge as a consequence of having to provide that knowledge in detail, they tend to do so selectively for explanatory knowledge, just as adults do. Therefore, we also predicted that children would not overestimate their knowledge for procedures, and so they would not drop their ratings after providing descriptions of the procedures. The convergence of properties that create an illusion of deep knowledge for explanations for adults is shown to be at work for children as well. This is different from the more domain general overconfidence in one's mental abilities, which is predicted to be strongest for the younger children. Thus, two different factors are influencing judgments of knowledge that should be distinguishable in patterns of development.

In short, recent studies of adult assessments of the depth of their own longstanding knowledge reveal an illusion of knowing that is specific to explanatory understanding. Given the extensive literature on the emergence of metacognitive abilities in the school years and given consistent claims that younger children have inflated views of their capacities, it is important to ask whether the particular structural properties of explanatory knowledge that cause a specific effect in adults also cause a comparable effect in school-aged children, or whether other developmental changes that occur during that period make the effect disappear.

Study 1 examined children's assessment of their own understanding of how mechanical devices work. Study 2 used the same methods as Study 1 but instead involved procedural descriptions. (We chose to use knowledge of procedures as opposed to other types of knowledge, such as of facts and narratives, because it offered one of the clearest contrasts with explanatory knowledge in work with adults.) To determine the calibration of the children's ratings and to provide an alternative means of examining the illusion of explanatory depth, an independent group of adult participants rated all the explanations and descriptions of the children for each study. These ratings were compared with the children's ratings to determine how well calibrated the children were at different points in the study.

Study 1: Devices

Method

Participants

Twenty-four kindergartners, 24 second graders, and 24 fourth graders participated in the study. (An additional 15 kindergartners, 8 second graders, and 4 fourth graders participated but were excluded from analysis because they did not pass a critical comprehension screening test that is described subsequently.) Participants were primarily from middle-class backgrounds. The mean age of the kindergartners was 6 years, 3 months (*range* = 5 years, 9 months to 7 years, 6 months; 10 males, 14 females); the mean age of the second graders was 8 years, 3 months (*range* = 7 years, 9 months to 9 years, 2 months; 14 males, 10 females); and the mean age of the fourth graders was 10 years, 2 months (*range* = 9 years, 6 months to 10 years, 7 months; 8 males, 16 females).

Children were recruited either from a local elementary school or from the greater New Haven area by calling families based on birth records. Children were tested in a quiet room either in our laboratory or at the elementary school. The parents of all children who participated provided written consent and the children themselves agreed to participate. The children received stickers and a certificate for their participation. All sessions were recorded on audiotape to be transcribed for further analysis; children granted consent for this as well.

Seventeen undergraduate students also participated in this study, compensated with experimental credit for their introduction to psychology course. They were recruited through sign-up sheets posted at the university, and they were tested in a quiet room in the laboratory.

Materials

The study was designed to be as similar as possible to the original illusion of explanatory depth studies with adults to make the comparison between the performance of adults and children more valid. In preparing the materials for this study, children's books and web sites discussing how things work were examined to create a list of approximately 20 devices with which most children would be familiar. We then piloted the study with children in kindergarten, second grade, and fourth grade to pick the final set of items to use (choosing the ones children were most familiar with) as well as to make sure the children understood the task before conducting the study. Fourteen items were selected for the initial items, and from among those items, four (of six possible items) constituted the actual measures for each participant.

Procedure

The experiment was divided into three parts: training, self-rating, and posttest. It was conducted as an oral interview of one 20- to 30-min session. A fourth part of the experiment was conducted in a separate session with adults rating the children's explanations.

Training. Children were trained on a 5-star scale to rate their knowledge about a given topic. A 5-star rating meant that the person knew all of the parts of a device and how they worked together. Three stars meant that they knew only some of the parts, and one star meant that they knew what the device did but not how it worked. The experimenter emphasized that the more someone knows, the more stars they get. Children could choose any number between 1 and 5 for their ratings, and they were reminded of this throughout the training.

The goal of the training was to make sure that the children understood that there were different levels to which a person could know something, and also how to use the scale to assess their own knowledge. Three examples were given: a can opener, a zipper, and an elevator.

The can opener was used first solely as an instructive example. Five pencil drawings of a can opener were placed on the table in front of the child, with each picture showing more detail. The experimenter gestured to the most detailed picture, and then told children that someone who knows all about how a can opener works might say the following: "A can opener has a sharp-edge wheel that slices into the lid. First you squeeze the handles together to attach the can opener to the can. The sharp wheel is on top and a toothed wheel is under the edge of the lid and behind the sharp edge. When you turn the knob, that turns the wheels to move the can around until the lid is cut off." After hearing the explanation, children were informed that because it tells about all the parts of a can opener and how they work together, that should get 5 stars. The experimenter placed a small strip of laminated paper showing 5 stars underneath the corresponding picture.

Children were then told that they were going to hear a few things other kids said who did not understand as well how a can opener works. The explanation meant to serve as a 1 star was read to the children next: "A can opener cuts the lid off a can. After you turn the knob around and around for a while, that cuts off the lid of the can." Children were then told, "That answer says what a can opener does, but it doesn't talk about the parts involved and how they work together. So that should get 1 star!" This time, the experimenter gestured to the least detailed picture while reading the explanation, and then placed a strip of paper showing 1 star underneath that picture.

Finally, the experimenter read the 3-star explanation. "When you squeeze the handles together, you attach the can to the can opener. Then you turn the knob around, which makes a sharp wheel cut off the lid of the can." Children were then told, "The person who says that understands more than just what a can opener does, but not really how all of the parts work together. They don't understand how exactly turning the knob makes the lid get cut off. They know more than the 1 star and less than the 5 stars. So that should get 3 stars!" The experimenter gestured to the picture in the middle, and then placed 3 stars underneath it. Children were told that they could also give 2 or 4 stars, depending on how much someone knew. Two stars and four stars were placed underneath the appropriate pictures. The experimenter reemphasized that the more someone knows, the more stars they get, and 5 stars meant that someone understood all of the parts and now they work together.

Next, the examples of the zipper and the elevator were given to provide the children a chance to choose the level of stars; the children were read an explanation and told that it was a 5-star explanation, and then they were asked how many stars several other explanations should get. Each successive example became more interactive. For the zipper example, five pictures of a zipper drawn in more detail were placed in front of the child to provide additional cues. The elevator example did not include any pictures. If the children were incorrect about any of the ratings (rating a 3-star explanation as worth 1 star, for example), the experimenter reread the explanation, briefly described the rating scale once again, and prompted for a rating. To give a rating, either children could say the number out loud, or they could gesture to the appropriate number of stars. The majority of children chose to do the former.

For the training portion of the experiment, all the explanations were roughly comparable in length as well as the complexity of language used. The 5-star explanations tended to be a bit longer than the other two because to perfectly equate the length of the three levels of explanations was not feasible, given that the 1-star explanations by definition provided less information, and thus adding too many extra words as filler made those explanations sound bizarre.

Self-rating. In this portion of the study, children gave a series of four ratings of their own level of knowledge about different devices. Immediately following training, children were told, "Now we're going to do something more fun. Now I'm going to ask you to tell me how many stars you know about how something works."

The initial rating was taken as children were asked to rate their own knowledge given a list of 14 different devices (e.g., piano keys, a tricycle). The experimenter told the child, "Think about how much *you* know about how a ____ works. How many stars would you give for what *you'd* say about how a ____ works?" The experimenter filled in the blank with the name of each device and paused for the child to give a self-rating. If the child did not respond, the experimenter prompted with the question, "How many stars do you know about how a ____ works?" If the child did not know what the item was, the item was skipped. This was rare, but did happen more frequently with the kindergartners (approximately one item was unknown for every 28 items rated).

The second rating was taken after the children gave explanations. Children were asked to give a detailed explanation of how four of the following devices worked (all of which were on the original list of 14): a toaster, a gumball machine, a water faucet handle, a stapler, a toilet, or a music box. Of these six items, four were chosen randomly, making sure that the children were familiar with those items. For each item, the child was told, "Think about all you know about how a ____ works. Now tell me *all* that you know about how the parts of a ____ work together." After providing each explanation, children were asked to rerate their level of knowledge. The experimenter said the following, inserting the number of stars given on the initial rating: "At the beginning, you said you knew X stars about how a ____ works. Now that you've told me how it works, do you think X stars was right, or do you think you should've said a different number for what you knew back then?" The experimenter

randomly reversed the order within the prompt, sometimes stating, "Do you think you should've said a different number for what you knew back then, or do you think *X* stars was right?"

The third rating followed a diagnostic question. For each item that was administered, a question was asked that was designed to make the children apply their knowledge about the mechanical workings of the item. After answering the question, children were once again prompted to rerate their level of knowledge.

Finally, children were provided with a child's expert explanation of the device. After hearing the expert explanation, children were told, "Now you've heard the expert who knows 5 stars about how a ____ works." They were then prompted to rerate their level of knowledge. (See Appendix A for stimuli used in this study.)

Posttest. All too often, younger children perform less accurately than older children because they do not understand the task. At the conclusion of the interview, therefore, a posttest was administered to ensure the children had clearly understood the scale and how to use it. They were told the following: "Remember the toaster? Well, I'm going to read you several different things people said for how a toaster works, and I want you to tell me which should get 5 stars, which should get 3, and which should get 1. You can change your mind after I read all of them if you want." Each explanation was designed to represent a different star rating level: 5, 3, and 1. The explanations were read in random order, and after each one, the child was prompted for a rating. If the child was able to differentiate between the different qualities of explanations using the rating scale, it was assumed that he or she understood the scale. In this study, children were told only once that the explanations were designed to represent different star rating levels, and thus they were not prompted to give different ratings if they assigned the same star rating to two explanations. The children who did not clearly understand the scale (differentiating the quality of the three explanations) were eliminated from the analysis. Such a strict screening procedure might result in underestimates of developmental change by causing a larger percentage of children of younger ages to be excluded, but it was important to make sure that the children who were included in the study completely understood how to apply the rating scale.

During the entire study, the experimenter was careful to keep her body language and comments as neutral as possible. While the child was talking, the experimenter kept her eyes down on her paper, and after the child was finished, the experimenter said, "Okay." The purpose of this was to attempt to keep children from using social cues to make decisions about how well they had performed.

Adult ratings. The children's explanations from all sessions in Study 1 were transcribed, with "ums" and other miscellaneous comments removed. The explanations were randomly assorted into six packets, with approximately 48 items in each packet (both grade and item were mixed within each packet).

Participants were instructed that they would be reading explanations given by children and rating them on a 5-star scale. They then received the same can opener training example as the children, except the instructions were written instead of oral. After reading the training example, each participant was randomly given four packets of explanations with the instructions to rate the sets of explanations given by

elementary school children on the 5-star scale. Each participant rated approximately 200 explanations during the 30-min session.

Results

Child data: ratings

For each participant, averages were calculated for each of the four rating tasks for the four items that the participant rated during the study: the initial rating (initial), after the explanations (postexplanation), after the answer to the diagnostic question (postquestion), and after the expert explanation (postexpert). These average ratings are coded as the variable *rating task*. Participants' responses over successive rating tasks are shown in Fig. 1.

A repeated-measures ANOVA with rating task (initial–postexpert) as a within-subject factor and grade as a between-subject factor showed a significant difference in rating task, $F(3, 207) = 6.870$, $p < .001$, $\omega^2 = .091$, with no rating task × grade interaction. There was, however, a trend for the ratings to decrease as grade increased, $F(2, 69) = 2.784$, $p = .069$, $\omega^2 = .075$. Planned LSD paired comparisons showed that the only significant difference between the grades was that the average ratings for kindergarten were significantly greater than those for fourth grade, $p = .022$.

Overall paired t tests examining differences between the rating tasks showed a significant drop between the initial rating and the postexplanation rating, $t(71) = 4.313$, $p < .001$. There was also a significant drop between the initial rating and the postquestion rating, $t(71) = 3.979$, $p < .001$.

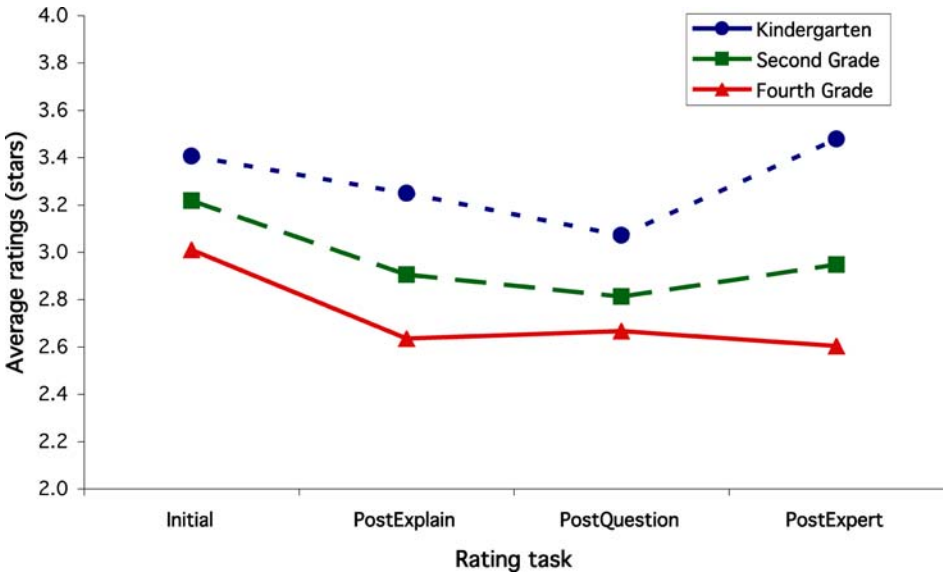

Fig. 1. Study 1: devices. Mean ratings over rating tasks by grade.

Taking the grades separately, we found some different trends. Second and fourth graders both had significant differences in rating tasks as measured by a repeated-measures ANOVA, $F(3, 69) = 4.521$, $p = .006$, $\omega^2 = .164$, and $F(3, 69) = 8.197$, $p < .001$, $\omega^2 = .263$, respectively. For both second and fourth grade, there was a significant drop between the initial rating and the postexplanation rating, $t(23) = 4.505$, $p < .001$, and $t(23) = 5.196$, $p < .001$, respectively. Finally, fourth graders showed a significant drop between the initial rating and the postexpert rating, $t(23) = 3.462$, $p = .002$.

For kindergarten, there was no significant effect of rating task as measured by a repeated-measures ANOVA, $F(3, 69) = 1.593$, $p = .199$. Planned within-subject contrasts showed that the only significant difference was between the postquestion rating and the postexpert rating, $F(1, 23) = 4.343$, $p = 0.048$, $\omega^2 = 0.159$.

To determine the percentage of the children in each grade who dropped their ratings after providing explanations, the average drop across all items was calculated for each child. Thirty-three percent of the kindergartners dropped their ratings after providing explanations (8 of 24), whereas more than 70% of the second and fourth graders dropped their ratings (17 of 24 and 19 of 24, respectively).

Child data: explanations

The explanations were on average 34 words long. The explanations by kindergartners were slightly shorter than the explanations by second and fourth graders (average explanation length by kindergartners = 26 words; second graders = 35 words; fourth graders = 41 words). Kindergartners also gave extremely short answers or said, "I don't know," more often than the older children (approximately 8% of explanations were shorter than 7 words for kindergartners; approximately 3% of explanations were shorter than 7 words for second and fourth graders). Sample explanations are included in Table 1.

Qualitatively, the explanations by children in kindergarten, second grade, and fourth grade varied significantly. Whereas kindergartners focused more on listing things involved with each item (e.g., "you do this and then it does this"), older children used more causal relations in explaining the items (e.g., "pushing this makes the heat come on which toasts the bread"). Each explanation was coded for the causal language used (involving action words demonstrating the relation between two or more steps, such as "if/then", "do *X* so *Y* can happen", and "*X* causes *Y*"). Approximately 10% of the explanations given by kindergartners used this sort of language, compared with 25% of the explanations by second graders and 40% of the explanations by fourth graders.

Adult data

Each explanation was rated between 11 and 13 times. The averages were calculated for each child's explanations in each grade, and gamma correlations were calculated between the adult rating of the explanation and the children's ratings for each rating task. Gamma correlations are generally used for calibration research; they are more conservative than Pearson correlations, taking into account ties in calculating the coefficient (see Nelson, 1984, for more information). The mean ratings are shown in Table 2; correlations between adults' and children's ratings appear in Table 3.

Table 1
Samples of children's explanations from Study 1

Grade	Item	Explanation
Kindergarten	*Toaster*	"You put something in it and then you press a button and then—you press the button. Push it down and leave it there. And then it heats and then it comes up too."
	Stapler	"You push it down and then what comes out staples stuff."
	Gumball machine	"It goes like... When you press a button like something bumps it and then it goes around and around to the bottom then you pick it up and put it in your mouth."
Second grade	*Toaster*	"Well, you put the bread in and you push this little lever down so then there you go. It'll heat rays inside and it'll make the bread really really hard and stuff and it'll just pop out."
	Stapler	"There's a bottom part and a top part and you press the top part down and it staples. The staples are in the top part."
	Gumball machine	"You put a quarter in it and then you spin the knob, and a gumball rolls down and it comes out the bottom."
Fourth grade	*Toaster*	"Ok. A toaster is made by electricity. You plug it in. There's a cord it comes electricity and then you put bread in. And then you press a button down. When you hit that button all the way down red lights which is heat comes out which is from the electricity and it heats the bread and when it comes out it's toast."
	Stapler	"You take the stapler, and there is this little place where the staples are and when you push down the staples are held up here and they're turned downwards when you push it down one staple goes in the edge and it pushes down any thing you want and there's a thing on the bottom that flattens the point of the stapler so it becomes flat and connect it .... and I think that's it."
	Gumball machine	"You have ... sometimes you put a penny or a quarter and you put it in a little slot and then you turn the knob and the penny goes down and when the knob is almost done, the gumball—there's something on the knob that pushes the gumball up and it comes out."

Table 2
Study 1: means and standard deviations for children's and adults' ratings of explanations for mechanical devices

Grade	Adult rating	Child ratings			
		Initial	PostExplain	PostQuestion	Final
Kindergarten	1.43 (0.45)	3.41 (0.90)	3.25 (0.88)	3.07 (1.16)	3.48 (1.09)
Second	1.79 (0.74)	3.22 (0.89)	2.91 (0.91)	2.82 (0.90)	2.95 (0.78)
Fourth	2.26 (0.84)	3.01 (0.86)	2.64 (1.00)	2.67 (1.05)	2.60 (0.91)

Note. Mean ratings are based on a 5-point scale; standard deviations are in parentheses.

Table 3
Study 1: correlations between adults' ratings of knowledge and children's ratings

Grade	Rating			
	Initial	PostExplain	PostQuestion	Final
Kindergarten	.057	.234*	.154	.133
Second	.196*	.300**	.253**	.248**
Fourth	.340**	.426**	.443**	.384**

Note. Values are gamma correlations.
* $p < .05$
** $p < .01$

Using the adults' average ratings as an approximation of the true quality of the children's explanations, older children were better calibrated at all time points (i.e., the more highly correlated the children's ratings were to the adults' independent ratings). The fourth graders were best calibrated, $p < .001$, for all correlations (initial rating–postexpert rating). The second graders were not calibrated as well, although the correlations were still all significant. The kindergartners were significantly correlated only at postexplanation.

Additionally, all children were better calibrated after providing an explanation than they were initially. The correlations for fourth and second graders were higher postexplanation than at the initial rating. Even kindergartners, who did not show a significant drop between their initial ratings and their postexplanation ratings, were better calibrated after providing an explanation than before.

Discussion

We found a trend for an overall developmental effect in ratings, in that the younger the children, the higher their ratings. However, beyond that effect, there is a clear awareness of an illusion of explanatory depth for children as young as second grade, as shown by the drop between the initial rating and the postexplanation rating. At least by second grade, they are able to become aware of their initial overestimates by providing explanations. Supporting this, in debriefing, some of the children expressed surprise at how little they knew about the items in the study, remarking that they had thought they knew more.

The explanations themselves varied in quality, but above and beyond these differences in detail there was a general effect across all items of an illusion of explanatory depth at both second and fourth grade. Thus the degree of illusion may have varied across items, as it does for adults, but it was present in the same direction to varying degrees for all items and therefore was not carried by one or two items.

Kindergartners did not show a significant drop in their rating between those two rating tasks, although the graph suggests this trend. Looking at patterns of performance for individual children across the study, 8 of the 24 kindergartners showed a drop in ratings (on average) after providing explanations, suggesting that providing an explanation does not help them recognize the gaps in their own understanding.

An additional finding emerges from these data: unlike the older children, the ratings from kindergartners *increased* between the postquestion rating and the postexpert rating. In fact, after hearing the expert explanation, some of the children commented that they knew all of that already (whereas their own mediocre explanations suggested otherwise). For example, after giving all 1's for a particular item, hearing the expert explanation caused one kindergartner to say, "Oh! I think I should've said 4 or 5." These results suggest that kindergartners may be having a difficult time listening to the expert explanation and reflecting back to their own original knowledge and explanations. This raises some questions about children's ability to monitor the sources of information that are addressed in the General Discussion.

The ratings from the adults provide additional support to the idea that the process of providing explanations helps children to become aware of their initial illusion. If the children were initially overestimating their knowledge, independent ratings of the explanations should be more highly correlated with self-ratings given after providing explanations than with the initial ratings. This is exactly what we found: children in all grades were better calibrated after providing explanations than initially.

Together, the results from adults and children also support the previous developmental research finding that the younger the children, the more they overestimate their knowledge. This is demonstrated both by the trend that the younger the children, the higher their average ratings, and by the finding that the younger the children, the more less accurately calibrated they were at all time points.

Study 2: Procedures

Study 1 demonstrated that children can become aware of their illusion of explanatory depth by second grade (as suggested by the significant drop between the initial rating and the postexplanation rating) if not earlier (as suggested by the better calibration for *all* children postexplanation as compared with initially, seen with the adult ratings). In adults, an awareness of an illusion of depth was found to be selectively strong for explanatory knowledge (knowledge about mechanical devices or biological phenomena, for example). Adults were much more accurate at judging their knowledge in different areas, such as for facts, narratives, and procedures; they generally did not initially overestimate their knowledge, and therefore they did not drop their ratings after answering questions or giving descriptions.

As mentioned in the Introduction, there are several differences between explanatory knowledge and other types of knowledge that seem to contribute to the illusion of explanatory depth in adults. This study explored the specificity of the illusion of depth in children by asking them to assess their own knowledge of procedures. Our aim in this study was to find out if children have an illusion of depth for their knowledge about procedures, and if they do, when they are able to recognize it.

Methods

Participants

Twenty-four kindergartners, 24 second graders, and 24 fourth graders participated in the study. (Nine kindergartners were excluded from analysis due to a failure to meet the posttest criteria. See below for more information.) Participants were predominantly from middle-class backgrounds. The mean age of the kindergartners was 5 years, 9 months (*range* = 4 years, 11 months to 6 years, 6 months; 9 males, 15 females), the mean age of the second graders was 7 years, 7 months (*range* = 7 years, 0 months to 8 years, 3 months; 8 males, 16 females), and the mean age of the fourth graders was 9 years, 9 months (*range* = 9 years, 0 months to 10 years, 7 months; 15 males, 9 females).

Children were recruited from a local elementary school or from the greater New Haven area by calling families based on birth records. Children were tested either in a quiet room in our laboratory or in a quiet room at the elementary school. The parents of all children who participated provided written consent and the children themselves agreed to participate. The children received stickers and a certificate for their participation. All sessions were recorded on audiotape to be transcribed for further analysis; children granted consent for this as well.

Twenty undergraduate students participated in this study, compensated with experimental credit for their introduction to psychology course. They were recruited through sign-up sheets posted at the university, and they were tested in a quiet room in the laboratory.

Materials

This study was designed to be as similar as possible to the original illusion of explanatory depth studies with adults. Approximately 20 real-world procedures were chosen with which most children would be familiar and yet which had multiple steps: we wanted the expert descriptions for this study to be approximately the same length as the expert explanations in Study 1 to make the two studies as equivalent as possible. We then piloted the study with children in kindergarten, second grade, and fourth grade to pick the final set of items to use (choosing the ones children were most familiar with) as well as to make sure children understood the task before conducting the study. Fourteen items were selected, and from among those items, four (of six possible items) constituted the actual measures for each participant.

Procedure

The experiment was almost identical to Study 1, except that two experimenters were involved in testing. Each session was divided into three parts: training, self-rating, and posttest.

Training. Participants were trained on a 5-star scale to rate their knowledge about a given procedure. A 5-star rating meant that the person knew all of the steps of how to do a certain procedure. Three stars meant that they knew only some of the steps, and one star meant that they knew what the procedure was but not the steps of how

to do it. The experimenter emphasized that the more someone knows, the more stars they get. Participants could choose any number between 1 and 5 for their ratings.

From piloting the study, we found that two examples were sufficient for children to understand the scale. First, an example of folding a flag "soldier-style" was explained as an instructive example, using pictures to help make the points. Five pictures depicting the steps of folding a flag soldier-style were placed on the table in front of the child, with each picture showing more detail and more steps. Children first were told that someone who does not know how to fold a flag soldier-style might say something like the following: "You have a flag and you fold it in a special way and then it is folded soldier-style." They were instructed that because that person knows that you have to fold the flag, but does not really know anything about the steps of how to do it, the explanation should get 1 star. The experimenter pointed to the picture with the least amount of detail, and then placed a strip of paper with 1 star underneath the picture.

Next, children were told that someone who knows *some* of the steps to folding a flag soldier-style might say something like this. "You need two people to work together to fold the flag. The flag is first folded in half, and then you do some more folds. At the end, the flag ends up folded into a small triangle with stars showing." They were told that because that person knows some of the steps, but not all of them, the explanation should get 3 stars. The experimenter pointed to corresponding picture and placed 3 stars underneath it.

Finally, children were told that a person who knew all of the steps of how to fold a flag soldier-style might say the following. "You need two people to work together to fold the flag. Have two people stand on either side of the flag, holding a corner in each hand. Fold the lower half of the stripe part of the flag over the star part, and then fold the flag again with the star part on the outside. Starting at the striped end, fold one corner into the opposite side of the flag, forming a triangle. Keep doing this triangular folding until only a small strip of the star part shows. Then tuck the last strip into the triangle. At the end, the flag ends up folded into a small triangle with stars showing." Children were informed that because that person knows all the steps to folding a flag soldier-style, the explanation should get 5 stars. Once again, the experimenter pointed to the corresponding picture and placed 5 stars underneath it. The experimenter then reminded children that they could also give 2 or 4 stars, depending on how much someone knew. Two stars and four stars were placed underneath the appropriate pictures.

Next, children were read a 5-star description of how to use a washing machine and were asked how many stars it should get. Several other descriptions were then read to the children, and they were asked how many stars each should get. If the children were incorrect about any of the ratings, the experimenter reread the description, briefly described the rating scale once again, and prompted for a rating. For this portion of this experiment, we once again tried to balance the descriptions as much as possible for length as well as the complexity of language used.

Self-rating. In this portion of the study, children gave a series of three ratings of their level of knowledge about different devices. We eliminated the diagnostic question and rerating step to be comparable to the original illusion of explanatory depth studies with adults, which did not include diagnostic questions for procedures.

Immediately following training, children were told, "Now we're going to do something more fun. Now I'm going to ask you to tell me how many stars *you* know about the steps to do certain things."

The initial rating was taken as children were asked to rate their own knowledge given a list of 14 different procedures (e.g., how to plant a flower seed, how to make a cake from a mix). The experimenter told the child, "Think about all you know about the steps of how to ___. How many stars would you give for what *you'd* say for the steps of how to ____?" The experimenter filled in the blank with the name of each device and paused for the child to give a self-rating. If the child did not respond, the experimenter prompted with the question, "How many stars do you know about how to ____?" If the child did not know what the item was, the item was skipped.

The second rating was taken after the children gave descriptions. Children were asked to give detailed descriptions for the steps of the following procedures: how to make cheese pizza, how to get money out of an ATM machine, how to change a flat tire, and how to catch a fish with a fishing rod. If the children were unfamiliar with one of those items, the following items were substituted: making up the bed from scratch and carving a pumpkin. For each item, the child was told, "Think about all you know about the steps of how to ___. Now tell me *all* that you know about the steps of how to _____." After providing each description, children were asked to rerate their level of knowledge in the same manner as in Study 1.

Finally, children were provided with a child's expert description of how to do each procedure. After hearing the expert description, they were then prompted to rerate their level of knowledge. (See Appendix B for stimuli used in this study.)

Posttest. At the conclusion of the interview, a posttest was administered to ensure the children understood the scale and how to use it. They were told the following: "So, now we're going to try one more. I'm going to read you several different things people said for how to make a cheese pizza, and I want you to tell me which should get 5 stars, which should get 3, and which should get 1. You can change your mind after I read all of them if you want." They were then read three descriptions of how to make a cheese pizza. Each description was design to represent a different star level: 5, 3, and 1. If the child was able to differentiate between the different qualities of descriptions using the rating scale, it was assumed that he or she understood the scale.

One problem with the posttest from the last study is that we did not reemphasize that each explanation was supposed to represent a different star level, excluding several children from analysis who gave two explanations the same ratings who might have been able to differentiate the explanations. Therefore, for this study, we emphasized that each description was meant to represent a different star level. The children who did not pass the posttest were eliminated from the analysis.

Adult ratings. The children's descriptions from all sessions in Study 2 were transcribed, with "ums" and other miscellaneous comments removed. The descriptions from 3 fourth grade participants were unable to be transcribed due to tape failure, so only the descriptions from 21 fourth graders were rated in this study, along with the descriptions from 24 kindergartners and 24 second graders. The descriptions were randomly assorted into six packets, with approximately 48 items in each packet (both grade and item were mixed within each packet).

Results

Child data: ratings

For each participant, averages were calculated for each of the three rating tasks for the four items the participant rated in this study: the initial rating (initial), after the descriptions of the procedures (postdescription), and after the expert description (postexpert). These average ratings are coded as the variable *rating task*. Participants' responses over successive rating tasks are shown in Fig. 2.

A repeated-measures ANOVA with rating task (initial–postexpert) as a within-subject factor and grade as well as experimenter as between-subject factors showed no significant effect of experimenter. Therefore, we dropped that factor from further analyses.

A repeated-measures ANOVA with rating task (initial–postexpert) as a within-subject factor and grade as a between-subject factor showed a significant difference in rating task, $F(3, 138) = 4.452$, $p < .013$, $\omega^2 = .061$. There was no rating task × grade interaction. However, there was a significant effect of grade, in that the younger the children, the higher their ratings, $F(2, 69) = 11.280$, $p < .001$, $\omega^2 = .246$. Post hoc LSD paired comparisons showed that the mean ratings for kindergarten were significantly higher than those for second grade ($p < .003$) as well as fourth grade ($p < .001$). Overall paired t tests showed a significant *rise* between the initial rating and the postdescription rating, $t(71) = 3.592$, $p < .001$.

Taking the grades separately, we found no significant effect of rating task in a repeated-measures ANOVA. Planned linear contrasts for second grade showed a significant increase between the initial rating and the postdescription rating,

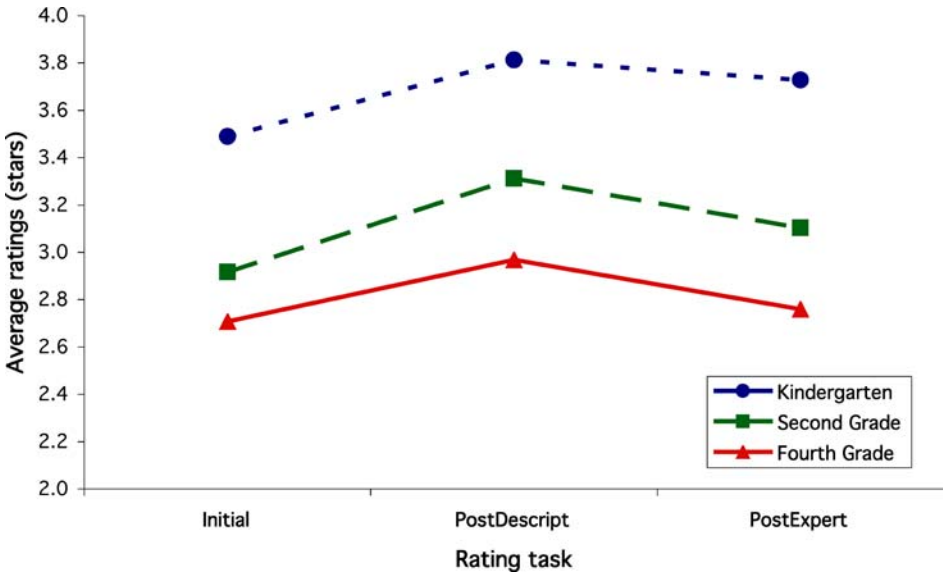

Fig. 2. Study 2: procedures. Mean ratings over rating tasks by grade.

$F(1, 23) = 3.760$, $p = .039$, and no other effects. Fourth graders showed a trend for the same comparison, $p = .052$, and a significant drop between the postdescription rating and the postexpert rating, $F(1, 23) = 4.600$, $p = .043$. There were no significant differences shown in linear contrasts for kindergartners, although perhaps a trend for an increase between the initial rating and postexplanation rating, $p = .062$.

To determine the percentage of the children in each grade who dropped their ratings after providing descriptions, the average drop across all items was calculated for each child. Twenty-five percent of the kindergartners dropped their ratings after providing procedural descriptions (6 of 24), while 33% (8 of 24) of second graders and 29% (7 of 24) of fourth graders dropped their ratings.

Child data: descriptions

The descriptions were on average 39 words long. The older the children, the longer their descriptions (average description length for kindergartners = 24 words, second graders = 38 words, fourth graders = 54 words). Kindergartners were also more likely to give extremely short answers or say "I don't know" than the older children in this study (approximately 1 of 10 explanations were shorter than 7 words for kindergartners; approximately 1 of 35 explanations were shorter than 7 words for second and fourth graders). Sample descriptions are shown in Table 4.

Most of the descriptions involved lists of steps, and rarely referred to causal relations between the steps. Children in each grade used approximately the same amount of causal language, with all children referring to causes and the relation between two or more items very rarely (2% of the descriptions included causal language, saying things like "the jack makes the car go up when you pump it"). Thus, qualitatively, the structure of the descriptions and their patterns of developmental change were quite different from those found for explanations in Study 1.

Adult data

Each description was rated between 9 and 11 times. The averages were calculated for each child in each grade, and gamma correlations were calculated between the adult rating of the description and the children's ratings at each time point. (Mean ratings are shown in Table 5; correlations between children's and adults' ratings appear in Table 6.)

Using the adults' average ratings as an approximation of the true measure of the quality of the children's descriptions, the second graders were better calibrated than the fourth graders for the initial rating and the postdescription rating (for second graders, $p < .01$ for all correlations; fourth graders were insignificantly correlated initially, but $p < .01$ for postdescription and postexpert ratings). However, there were fewer descriptions by fourth graders rated in this study, so the correlations might have been larger had we been able to include descriptions from all 24 of the children.

Children were better calibrated after providing descriptions as compared with their initial impressions. For the second graders, the correlations were higher postdescription than initially. For fourth graders, we see the same effect: correlations were higher postdescription compared with initially. The only time kindergartners

Table 4
Samples of children's descriptions from Study 2

Grade	Item	Description
Kindergarten	*Making pizza*	"You take dough and you make a big round circle. And then you put cheese on it. And then you throw it up in the air and put sauce on it. And then you put it in the oven."
	Changing flat tire	"Put some air in it, and then you pump it up until it becomes really big. And then it's good for you to drive again."
	Fishing	"You first put a worm on the fishing rod. Then you wait a while then you've caught a fish and you bring the hook back in."
Second grade	*Making pizza*	"You get a tray, and put the dough on top of the tray. You put some sauce on, and then you put cheese on. And then if you want other stuff on it you can put stuff on top. Then you put it in the oven and cook it."
	Changing flat tire	"You get another tire. You take the tire that is flat off. And then you get a jack and you put it up on the jack. And then you put the other tire on."
	Fishing	"First you get a fishing rod, and you put a worm or whatever you want to put on it. And then you put it in the water and you wait a little while until the fish comes. And when the fish comes you have to roll it for it to come up and see what you got."
Fourth grade	*Making pizza*	"You buy the dough and you add the sauce and whatever toppings you want. You let the dough rise for a while. Then you put the sauce on it and you roll it out. Then you put toppings on. And you set the oven to the temperature you want and cook it for probably a half an hour. Then take it out."
	Changing flat tire	"First, when you know that you have a flat because you hear the sound, you jack up the flat tire. Then take out the spare and take off the flat tire so the little peg that holds the tire on is sorta leaning on the jack. Then you kinda lift up that part of the car. You put the tire on and screw the bolts in and that's it."
	Fishing	"First you have to bait it up with a worm. Then you have to cast. Then you have to wait until you see the bobber go under. Then you have to reel—then you have to tug it a little bit to get it the fish's mouth to hook on. Then you have to reel it in medium fast, or medium slow sort of. Then, you have a fish."

were significantly correlated with the adults' average ratings is after providing the description.

Discussion

In rating their knowledge of procedures, children actually slightly *increased* their assessments of how much they know between giving an initial rating and after

Table 5
Study 2: means and standard deviations for self and adult ratings of descriptions for procedures

Grade	Adults' rating	Children's ratings		
		Initial	PostDescript	Final
Kindergarten	1.65 (0.74)	3.49 (0.87)	3.81 (0.77)	3.73 (1.08)
Second	1.93 (0.67)	2.92 (0.86)	3.31 (0.81)	3.10 (0.74)
Fourth	2.83 (0.91)	2.71 (0.80)	2.97 (0.78)	2.76 (0.77)

Note. Mean ratings are based on a 5-point scale; standard deviations are in parentheses.

Table 6
Study 2: correlations between adults' ratings of knowledge and children's ratings

Grade	Rating		
	Initial	PostDescript	Final
Kindergarten	.172	.376**	.102
Second	.356**	.484**	.284**
Fourth	.170	.393**	.479**

Note. Values are gamma correlations.
*$p < .05$; **$p < .01$.

providing a description. This finding is not driven by one or two items: the pattern of ratings was similar across all items, with no outliers showing the pattern seen in Study 1. This suggests that children do not suffer from an illusion of depth for knowledge of procedures, and providing descriptions of procedures does not lead children to think that their initial ratings of their knowledge were overestimates. Children were also more accurate in their ratings after providing descriptions in this study, demonstrated by the higher correlation with the adult ratings after providing descriptions than initially.

In the original studies on knowledge of procedures with adults, there was also a slight but nonsignificant increase in the rating tasks between the initial rating and the postdescription rating. A possible reason for the increase in ratings seen in the adult studies as well as here is that when describing the procedures, the information participants started with may have cued other knowledge, helping them realize that they actually know more than they originally had thought. This does not seem to occur with explanatory knowledge, such as examined in Study 1, in part because of the lack of a clear end state for explanations as well as the multiple layers of causal complexity.

Finally, unlike when reflecting on explanatory knowledge about devices, kindergartners do not show an increase in ratings after hearing the expert description. A possible explanation for this finding is that the ability to monitor the source of information differs depending on the kind of knowledge. We discuss this option in the General Discussion.

In sum, these results suggest that children are better calibrated about their knowledge after providing an explanation or a description. Both Study 1 with devices and

Study 2 with procedures suggest that the younger the children, the higher the ratings in general, but the critical pattern of a drop of ratings over tasks does not occur for procedures as it does for explanations, suggesting a different set of influences on self-ratings than with the case of explanations. In fact, unlike when providing explanations, children were likely to increase their self-ratings after providing the description (as opposed to decreasing them), and this new higher rating was closer than the initial rating to the adults' rating of the description. The initial ratings in Study 1 and Study 2, as shown in Fig. 3, are not significantly different, discounting the possibility of floor or ceiling effects for the current study; the initial ratings for individual grades are also not significantly different. Therefore, these results support the hypothesis that children have knowledge-specific illusions of their understanding similar to those of adults: the structural properties of explanatory knowledge that differ from knowledge of procedures seem to influence adults and children in equivalent ways. In some ways, it seems like children initially suffered from an illusion of ignorance for their knowledge of procedures, underestimating the depth of their knowledge, a marked contrast to the case of explanatory knowledge.

General discussion

These studies explored several questions relating to how children assess their own understanding for different kinds of knowledge. First, we examined children's assessments of their own knowledge for devices, hypothesizing that children would show an illusion of depth for their knowledge, and at least the older children would

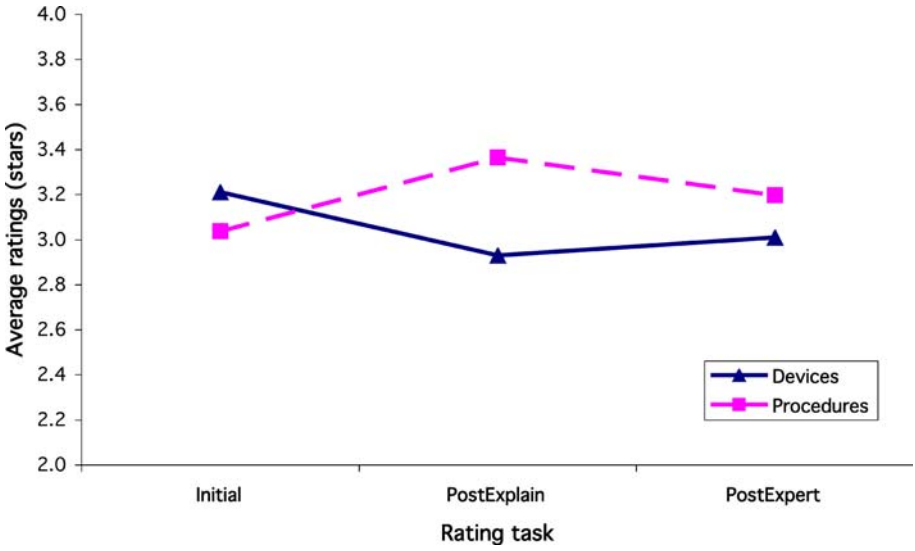

Fig. 3. Mean ratings over time collapsed across grade for Study 1 and Study 2.

recognize it after providing explanations. The results from Study 1 showed a trend that the younger the children, the higher they rate their knowledge. However, beyond the initial level of knowledge rating, all children show a similar pattern of responses in that they drop their self-ratings after providing explanations. Children as young as second grade clearly exhibit an awareness of their overestimates for their explanations of mechanical devices, initially suffering from what has been described in adults as the illusion of explanatory depth.

Second, we compared children's assessments about their own knowledge of devices to their knowledge of procedures to examine how the illusion varies with different types of knowledge. Our hypothesis was that after children are able to recognize the inadequacy of their own knowledge, they will have similar specificity in their illusions of their understanding as adults. Our results support this hypothesis: children have an illusion of depth for explanatory knowledge of mechanical devices (Study 1), but not for procedures (Study 2). Even kindergartners make different kinds of assessments for knowledge that is explanatory than for knowledge that is procedural: the difference in kindergartners' performance between the two tasks is significantly different, $t(46) = 2.037$, $p < .05$.

Although there are within-subject differences in how much one child knows about a particular item, as there were in the adult studies on this topic, the same general pattern is seen across all items. The differences therefore do not seem to be driven by individual items, nor can they be explained by floor effects in the second study. The differences should also not be due to demand characteristics, as one would expect that repeated questioning would affect judgments about both devices and procedures equally. Instead, there may be something fundamentally different about assessing one's own knowledge for devices as compared with procedures. Table 7 shows one way of visualizing the difference between ratings of procedures and explanations by showing the percentages of children in both studies who showed drops in ratings of their knowledge after providing an explanation or a description as well as those who showed increases in ratings.

As mentioned in the Introduction, explanatory knowledge differs from most other types of knowledge (such as knowledge of procedures) in at least three ways: how the knowledge is organized, the way information is used in expressing that knowledge, and the experience we have with self-testing the knowledge. Regardless of which specific factors influence children's assessments the most, these differences in the

Table 7
Percentage of participants decreasing or increasing this rating between the initial rating and after providing an explanation or description

Grade	Study 1: devices		Study 2: procedures	
	Decrease	Increase	Decrease	Increase
Kindergarten	33	25	25	50
Second	71	4	33	50
Fourth	79	13	29	63

Note. Percentages do not add up to 100, as some participants did not change between the two ratings.

characteristics of explanatory knowledge as compared with knowledge of procedures (as well as other kinds of knowledge) are so strong that children as young as second grade are clearly influenced by them and perform similarly to adults. To the extent that other kinds of knowledge share these factors, both adults and children should experience an illusion of depth.

The language that the children use in the different studies demonstrates that they think about explanatory knowledge very differently from knowledge of procedures; second and fourth graders used many phrases referring to causal relations in explaining devices, while they hardly ever referred to causal relations when talking about procedures. This adds additional support to the idea that these are very different kinds of knowledge, invoking different kinds of explanations and descriptions.

The current research focused on children from middle-class backgrounds; however, if the illusion of explanatory depth is widespread, one would expect to see it in different populations. Preliminary results in our laboratory suggest that these results generalize to children in less affluent groups and are not influenced by claimed differences in self-esteem of those groups (Skinner, Mills, & Keil, unpublished honors thesis).

The conclusions for kindergartners are less clear, suggesting a need for further investigation with young children. To make sure that the children completely understood the rating scale, the data from quite a few of the kindergartners had to be discarded during screening procedures. This suggests that the tasks in this study were quite difficult for kindergartners in general, and that the developmental changes that are found here may actually be conservative. Thus, a more extensive initial training session with kindergartners might include a larger percentage in the sample and would likely reveal a more dramatic developmental change.

Younger children may not be as sensitive to certain features of explanatory knowledge, and this difference would have affected both the performance on the posttest screening procedure and their evaluations of their own knowledge. Kindergartners may also not reliably monitor the information provided in their own explanations, and thus, they have a more difficult time recognizing gaps in their understanding. Across all our studies (as well as in previous metacognitive research), kindergartners are relatively inaccurate at assessing their own knowledge, and so it is difficult to determine how much of their self-ratings is driven by purely overestimating their knowledge or by being influenced by certain aspects of explanatory knowledge. The kindergartners do not seem to be as affected by the quality of their answers from one step of the study to the next: in Study 1, they did not significantly drop their self-ratings after providing explanations, as the older children did. Even so, as shown by adult ratings of the children's responses, the kindergartners in our study were better calibrated after providing explanations than initially, and the difference in the self-ratings for kindergartners between the initial rating and after providing the explanation resembles the drop shown by the older children. Also, as mentioned previously, the performance of the kindergartners in the two studies differs significantly, showing that these young children can make different sorts of assessments in these studies, even if it is difficult for them to do so in general. With a less verbally laden task, kindergartners might also be shown to more consistently be able to

recognize the inadequacies of their explanatory knowledge as well as evaluate the explanations and descriptions of others.

The only time kindergartners significantly changed their rating from one step to another was between the postquestion rating and the postexpert rating, and this occurred only when explaining devices and not procedures. These younger children may have been listening to the expert explanation and having difficulty differentiating it from their own. Still, because this confusion did not seem to occur for knowledge of procedures, this pattern begs the question of whether the ability to monitor the source of information varies across types of knowledge. Research with 4- and 5-year-olds has found that children often think they have always known something that they have just learned. However, they are more aware of learning new information when the information is behavioral as opposed to when it is factual (Esbensen, Taylor, & Stoess, 1997; Taylor, Esbensen, & Bennett, 1994). Then there is a drastic improvement between the ages of 4 and 6 in children's ability to monitor the source of factual information (Drummey & Newcombe, 2002). Perhaps preschoolers first have a behavioral understanding of knowledge acquisition (Perner, 1991), and children are later able to monitor their learning about other material, such as facts. After that point, causal complexity (which relates to both how the material is organized and the way information is used in expressing that knowledge) may play a major role in children's source monitoring ability: the more causally complex the information, the more difficult it is to monitor the source of the information. Performance in this study supports this idea in general: kindergartners were more likely to increase their ratings after hearing an expert explanation for devices than expert descriptions for procedures (which are less causally complex). Second graders also show this pattern to a lesser degree.

Another possible explanation is that young children have different ideas about what makes a good explanation; they may not understand the difference between a deep explanation and a shallow one for some domains. Perhaps younger children are most influenced by key terms used in explanations, and their sensitivity to causal complexity comes in later. In our study, then, young children might conclude that they knew just as much as the expert after hearing the expert explanation containing some of the same key words that the children themselves used. This seems improbable given that children had to successfully differentiate between three levels of explanations to be included in the study; therefore, they demonstrated some sensitivity to causal complexity. However, previous research suggests that young children are more accurate in assessing the performance of others than of themselves (Stipek, 1984), so they might use different criteria for evaluating their own explanations as compared with evaluating a specific other (which is the scenario posed during the training as well as the posttest). This may also explain why kindergartners do not significantly drop their ratings between the initial rating and the postexplanation rating: they are being asked to evaluate their own explanations, which might be difficult for them to do. Future research should explore these issues with kindergartners and younger children to determine both what they understand about causal complexity in explanations and how they evaluate their own knowledge.

Errors made by children and adults in their initial judgments of the extent of their knowledge have important implications for education. Learners may make many mistakes in determining when they have fully understood a concept, which self-testing may reveal. They nod their heads in agreement when the teacher asks if they understood a topic, not always just to get the teacher to be quiet, but because they often truly think they understand.

What then motivates children to ask questions, if they are often concluding that they deeply understand something even when they do not? One factor may involve the presence of interested adults. Thus, young children engage in more thorough exploration of scientific museum exhibits when their parents are present and shaping the way their children think through comparisons and usage of causal terms (Crowley et al., 2001). Explicit training and modeling both within and outside of the classroom can guide children's future question-asking and self-regulatory skills as well (Henderson & Garcia, 1973; Schunk & Zimmerman, 1997). These sorts of activities may push children to reflect further on their own knowledge, which allows them to experience the cognitive disequilibrium that leads to question asking of their own accord (Graesser & Olde, 2003). Young children may occasionally recognize gaps in their understandings that drive them to ask questions, but they may still be overestimating the depth of their knowledge. In fact, an illusion of explanatory understanding may actually provide them with some degree of confidence that they know enough to ask reasonable questions.

One implication of these findings for education is that children by at least the second grade can be made aware of their illusions of explanatory depth, given the appropriate circumstances. Our research shows that one way to make children reflect on the gaps of their knowledge is by asking them to explain themselves, but other tools may allow this sort of reflection, such as showing gaps and missing details in explanations given by peers, focusing on contradictory details, and illustrating techniques of local self-testing of knowledge. Teachers might want to use these sorts of tools to enable children to see the incompleteness in not just their own understandings but also in the minds of other children and adults. As a full insight into the imperfections and incompleteness of science continues to develop during late childhood and adolescence (Kuhn, 1996), efforts to teach younger children of such limitations may help accelerate that awareness.

However, other factors may also help maintain part of the illusion even when careful efforts are made to dispel it. In general, young children may often be optimistic about their performance (Stipek, 1984) and their futures (Lockhart, Chang, & Story, 2002) compared with older children and adults, and indeed such optimism may be adaptive. Persisting despite failure or overestimating how well one will perform a certain task allows young children to move beyond the typical difficulties they face, influencing how much they learn (Bjorklund & Bering, 2002; Bjorklund & Green, 1992). The illusion of explanatory depth may also be seen as adaptive: it is difficult if not impossible to have complete explanations and detailed intuitive theories for all aspects of knowledge, and thus the illusion of explanatory depth keeps both adults and children generally satisfied with skeletal theories about the world. At the same time, however, it is important to be able to know the limits of one's

understanding so as to seek out more information and better explanations from others. This line of research suggests that on reflection, both children and adults are able to realize the limits of their own understanding. Future research should explore how much reflection is enough for people to recognize the gaps and inconsistencies in their own folk theories.

In sum, by second grade, if not earlier, children are ensnared by the structural properties of explanatory knowledge that influence adults: they have an illusion of depth for explanations and can become aware of it, but they do not have an illusion of depth for procedures. Future research should move beyond the approach of examining overconfidence of "general knowledge" to more clearly understand how the properties of different types of knowledge impact our assessment of what we know. Identifying these features and their roles in structuring our knowledge would in turn provide insight into the development of our intuitive theories about the world.

Acknowledgments

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship to the first author, as well as under National Institutes of Health Grants R01-HD23922 and R-37-HD023922 to Frank Keil.

The authors are grateful to the staff, parents, and students of St. Mary's School and St. James' Elementary School as well as to the families bringing in children to our laboratory. The authors thank Paul Bloom, Karen Wynn, and the members of the Yale Cognition and Development Laboratory for their insightful comments. Finally, special thanks to Judith Danovitch, James Detzel, Megan Huth, and Harlyn Skinner for their assistance with this research.

Appendix A. Stimuli used in ratings for Study 1

Item	Diagnostic question	Expert explanation
Toaster	What makes the toast pop up after it's done cooking?	A toaster makes bread brown and crispy. When you press the button that sticks out on the side of the toaster, the slices of bread go down. When you do that, things inside the toaster get red hot and the heat cooks all sides of the bread. When it gets hot enough inside the toaster (depending on how dark you want the toast to get), the heat makes a switch push a hook that pops the toast back up.

Appendix A. (*continued*)

Item	Diagnostic question	Expert explanation
Water faucet handle	Why does a water faucet sometimes leak?	The water faucet handle is attached to a screw, which has a plug at the other end. This plug covers up a hole, and the water is right underneath that hole. So when you turn the handle, the screw turns and pulls the plug up from the hole, which gives the water a place to come out.
Stapler	Why does a stapler sometimes not work if you don't press down hard enough?	When you push down on the stapler, the top part of the stapler pushes the first staple in line through the papers. The dents on the bottom of the stapler force the end of the staple to clip together. Then when you stop pressing the stapler, a spring makes the top part bounce back, and another spring pushes the next staple into place.
Gumball machine	Why don't all of the gumballs fall out when you turn the handle?	When you put the right coin in the slot, you can turn the handle. There's a little wheel inside right next to the handle that can move around, and it has little slots to hold a gumball. So when you turn the handle, that little wheel turns around and moves a gumball right over a hole at the bottom of the gumball machine. Then a gumball can come out.
Flushing the toilet	How does pulling on the handle make the water go down?	When you pull the handle on the side of the tank, it pulls a chain, which is connected to a rubber plug, which covers a hole inside the tank. The plug moves out of the way and then water rushes from the tank into the toilet bowl. This makes what was in the bowl before go down a hole at the bottom of the bowl. Then, when the tank is empty the plug falls back down and covers the hole, and the tank is refilled with water for the next flush.

Appendix A. (*continued*)

Item	Diagnostic question	Expert explanation
Music box	How is music made in a music box?	A music box plays music. When you turn the handle, it makes a little piece shaped like a can move around in the box. That piece has little bumps on it of different lengths, and when the piece moves around, the bumps hit the teeth of a metal comb. Then the teeth shake to make different sounds, which makes the music.

Appendix B. Stimuli and expert explanations for Study 2

Item	Expert explanation
Making a cheese pizza	Place a rack in the center of the oven and preheat the oven to 350 degrees. Next, put the pizza crust on a pan and spread the tomato sauce on it. Then you sprinkle cheese on top of the sauce. After the oven is heated, put the pizza in the oven. Bake the pizza until the cheese is bubbling and the crust is a crisp brown, about 10 minutes. Remove the pan from the oven and place it on a wire rack to cool for 10 minutes. Slice with a special roller cutter that makes neat slices in the pizza.
How to get money out of an ATM (cash) machine	To get money out of an ATM cash machine, you have to insert your special card into the machine. After you do that, you type in your secret code number and press enter. Then you choose the ''withdraw'' button, and decide which account to take the money from. You enter in the amount of cash that you want and then you take the money out of the slot. Finally, make sure to get your receipt and your card back.
How to change a flat tire	Get the spare tire out of the trunk of the car. Use a special wrench to loosen each of the nuts that hold the flat tire on JUST a bit. Then, use a special lifting tool called a jack to raise up the car so it's up off the ground. Then you take the nuts all the way off. Now you can grab the flat tire and pull it straight out and off. Then, you put the spare tire on instead. Put the nuts back on, and tighten them all the way. Then, use the jack to lower the car to the ground. Put the old flat tire in the trunk so that you can throw it out, and that's it!

Appendix B. (*continued*)

Item	Expert explanation
How to catch a fish with a fishing rod	You need the following things: a fishing rod, some really thin rope that is called "fishing line," a hook, and some food that the fish likes, which is called "bait." First, make sure that the fishing line is attached to the fishing rod. Put a hook at the end of the line that is big enough to catch the kind of fish you want, and put some bait on the hook. Then, put the hook and the fishing line in the water. Move the rod around a little bit so it seems like the food you put on the hook is alive. When you feel a tug on the fishing line, pull the line back up so you can catch the fish on the hook. Then take the hook out of the fish and you can either keep the fish or throw it back.
Make up the bed from scratch	First, put on the fitted sheet, which has stretchy corners. Make sure that this bottom sheet is on really smoothly. Next put the flat sheet on top. Tuck in the part where your feet would go under the bottom of the mattress, and fold down the part of the sheet near your head so that you will be able to get under the sheet later. After that, make sure to put your pillows in the pillowcases. Then put on the blanket, put the pillows on, and you're done.
Carving a pumpkin	You get a pumpkin that is ripe, has no bruises or cuts. You have to make sure to place your pumpkin on several layers of newspaper before carving. You cut a circle around the stem of the pumpkin using a special "pumpkin carving" knife. Make sure the hole is large enough to reach in. Next you pull out the seeds and other stuff inside the pumpkin, and throw that stuff away. If you want, you can draw an outline of a face on the pumpkin with a marker where you want the face to be. Then you carefully cut out the design with your special knife. When you are finished cutting, simply push out the pieces to see your design.

References

Ahn, W., & Kalish, C. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Cognition and explanation* (pp. 199–225). Boston: MIT Press.

Bjorklund, D. F., & Bering, J. M. (2002). The evolved child: Applying evolutionary developmental psychology to modern schooling. *Learning and Individual Differences, 12*, 347–373.

Bjorklund, D. F., & Green, B. L. (1992). The adaptive nature of cognitive immaturity. *American Psychologist, 47*, 46–54.

Crowley, K., Callanan, M. A., Jipson, J. L., Galco, J., Topping, K., & Shrager, J. (2001). Shared scientific thinking in everyday parent–child activity. *Science Education, 85*, 712–732.

Drummey, A. B., & Newcombe, N. S. (2002). Developmental changes in source memory. *Developmental Science, 5*, 502–513.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *The nature of insight* (pp. 365–395). Cambridge: MIT Press.

Esbensen, B. M., Taylor, M., & Stoess, C. (1997). Children's behavioral understanding of knowledge acquisition. *Cognitive Development, 12*, 53–84.

Flavell, J. H., Freidrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology, 1*, 324–340.

Gelman, S. A., & Koenig, M. A. (2003). Theory based categorization in early childhood. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming buzzing confusion* (pp. 330–339). New York: Oxford University Press.

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology, 95*, 524–536.

Henderson, R. W., & Garcia, A. B. (1973). The effects of a parent-training program on question asking behavior of Mexican-American children. *American Educational Research Journal, 10*, 193–201.

Kuhn, D. (1996). Is good thinking scientific thinking?. In D. R. Olson & N. Torrance (Eds.), *Modes of thought: Explorations in culture and cognition* (pp. 261–281). New York: Cambridge Univ. Press.

Levin, D. T., Momen, N., Drivdahl, S. B., & Simons, D. J. (2000). Change blindness blindness: The metacognitive error of overestimating change-detection ability. *Visual Cognition, 7*, 397–412.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Decision Processes, 20*, 159–183.

Lockhart, K. L., Chang, B., & Story, T. (2002). Young children's beliefs about the stability of traits: Protective optimism? *Child Development, 73*, 1408–1430.

Markman, E. M. (1977). Realizing that you don't understand: A preliminary investigation. *Child Development, 48*, 986–992.

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development, 50*, 643–655.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43*, 96–111.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*, 521–562.

Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty* (2nd ed). Mahwah, NJ: Erlbaum.

Schunk, D. H., & Zimmerman, B. J. (1997). Social origins of self-regulatory competence. *Educational Psychologist, 32*, 195–208.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*, 605.

Stipek, D. (1984). Young children's performance expectations: Logical analysis or wishful thinking?. In J. G. Nicholls (Ed.), *Advances in motivation and achievement, Vol. 3: The development of achievement motivation* (pp. 35–56). Greenwich, CT: JAI Press.

Taylor, M., Esbensen, B. M., & Bennett, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child Development, 65*, 1581–1604.

Wilson, R. A., & Keil, F. C. (1998). The shadows and shallows of explanation. *Minds & Machines, 8*, 137–159.