Cognitive Development, 11, 523–550 (1996)

The Development of Evidence Evaluation Skills

Eric Amsel
Vassar College

Susan Brock
University of Saskatchewan

To assess developmental differences in evidence evaluation skills, 77 second- and third-grade students, 85 sixth- and seventh-grade students, 36 non-college-educated adults, and 40 college students were presented with four data sets depicting plants grown by each of four people. The data sets presented a perfect positive or zero correlation between plant health and the presence or absence of one variable, believed by participants to have a causal influence on growing healthy plants, or another, believed to have no causal influence. In each of three missing data conditions, the data sets depicted instances in which the status of the variable, outcome, or both were unknown in addition to the contingency data. After each data set was presented, participants judged (and justified) the causal status of the variable. Although demonstrating a basic competence, the two groups of children were more strongly influenced by prior beliefs and missing data than were the two adult groups. There were also age or educational differences in participants' tendency to justify judgments on the basis of the contingency data. The implications of the results for conceptualizing the continuity or discontinuity of children's, adults', and scientists' evidence evaluation skills are discussed.

Like other metaphors in science, likening the child to a scientist is a powerful image that informs researchers, theorists, and practitioners alike. The child-as-scientist metaphor is typically taken to mean that children and scientists are similar in the manner by which they acquire knowledge about the world (Gruber, 1973; Ross, 1981; Rosser, 1994). The image associated with the metaphor is of children, like scientists, forming, testing, and revising causal theories about the world. Through such processes, it is supposed that

The research was funded by a Social Sciences and Humanities Research Council of Canada (SSHRCC) postdoctoral fellowship and a Natural Science and Engineering Research Council of Canada (NSERC) operating grant awarded to Eric Amsel and an NSERC summer grant awarded to Susan Brock. Portions of the results were presented at SRCD, Seattle 1991. We thank Judi Amsel, Merry Bullock, Joyce Gastel, Richard Kitchener, Deanna Kuhn, Brenda McLaren-Cooley, Robert Sternberg, and two anonymous reviewers for their help and comments on various phases of the project.

Correspondence and requests for reprints should be sent to Eric Amsel, Department of Psychology, Weber State University, 1202 University Circle, Ogden, UT 84408-1202. E-mail: <eramsel@vassar.edu>.

children and scientists acquire an increasingly adequate understanding of the world.

The child-as-scientist metaphor has motivated a wide range of research studies that assess whether or not there is a continuity between how children and scientists acquire knowledge. Research has generally supported the view of a continuity between children and scientists in that both form and revise theories of the world (Brewer & Samarapungavan, 1991; Wellman & Gelman, 1992). Children form and revise intuitive theories, defined as integrated networks of explanatory concepts and causal beliefs regarding phenomena in specific domains such as psychology (Gopnik & Astington, 1988; Gopnik & Wellman, 1992), biology (Carey, 1985b; Gelman & Markman, 1986), and physics (Amsel, Goodman, Savoie, & Clark, 1996; Smith, Carey, & Wiser, 1985). These intuitive theories allegedly have the same explanatory and predictive functions for children that formal theories have for scientists, thereby allowing children and scientists alike to go beyond mere observations and descriptions of phenomena, to explanations and predictions of them as well (Carey 1985a; Wellman & Gelman, 1991).

Although this research converges on the view of children as theoretical scientists who explain and predict phenomena on the basis of intuitive theories, research addressing their status as experimental scientists is less conclusive. Children's status as experimental scientists has been tested, among other ways, by their ability to evaluate evidence from the world and its bearing on hypotheses about the world. *Evidence* is defined as information that serves to confirm or disconfirm hypotheses (Hemple, 1961). In many of these studies, children evaluate evidence bearing on causal hypotheses (e.g., whether a variable is causally or noncausally related to an outcome), and their status as experimental scientists is assessed by whether the children judge causal hypotheses on the basis of evidentially relevant information alone, without any influence of evidentially irrelevant information. For example, Kuhn, Amsel, and O'Loughlin (1988, Studies 1a and 1b) assessed whether children, adolescents, and adults evaluate multiple causal hypotheses (e.g., four variables presented as possible causes of an outcome) on the basis of sequentially presented instances of contingency data (e.g., the association between each variable and the outcome) alone, ignoring their prior beliefs about the causal status of the variables. Kuhn et al. demonstrated that children, unlike adults, often justified their judgments of the causal status of variables on the basis of their prior beliefs of the causal or noncausal status of the variables (belief-based judgments) and not the contingency data (evidence-based judgments).

Kuhn et al. (1988) also found that children continued to be inappropriately influenced by their prior beliefs despite being specifically requested to evaluate the causal hypotheses on the basis of the contingency data. Although more evidence-based judgments were made in such circumstances, children, adolescents, and even some adults interpreted the contingency data in a biased manner. For example, they would interpret contingency data showing that a variable and outcome are independent of each other as confirming their prior belief that the variable was causal by pointing to only those instances in which the variable and outcome are both present.

Kuhn et al. (1988) also noted that children confronted by data that were inconsistent with their prior beliefs would sometimes articulate a new, often incredible, belief to explain the contingency data. For example, one ninth-grade boy who initially believed that getting colds or no colds was unrelated to the type of relish (mustard or catsup) eaten was shown that children who ate mustard got few colds, whereas children who ate catsup got lots of colds. The ninth grader judged that the type of relish was causal but justified the judgment by claiming that mustard keeps people healthier than catsup because it has more ingredients (a belief that was refined in a variety of ways over the course of the interview). Although the child's judgment of the causal status of the variable was consistent with the data, his justification of the judgment was based on a (perhaps newly formed) belief about how the variable produces the outcome. Kuhn et al. considered such cases as beliefs mediating the interpretation of the contingency data because the beliefs provided participants with an explanation of what would have been otherwise inconsistent data.

Thus, Kuhn et al. (1988) suggested that children fail to evaluate evidence independently of their causal and noncausal beliefs because their beliefs influenced and/or mediated if, when, and how the evidence was evaluated. That is, instead of evaluating causal hypotheses solely on the basis of contingency data alone (i.e., what was described as reasoning about a theory and coordinating it with evidence), children evaluated the evidence in light of available (prior or newly formed) causal and noncausal beliefs (i.e., what was described as reasoning with a theory and merging it with information from the world). More recent research has suggested that with sufficient practice, children are capable of evaluating evidence independently of beliefs. Microgenetic studies have revealed that, over a few short weeks, children become increasingly likely to correctly interpret evidence and its bearing on causal hypotheses (Kuhn, 1995; Kuhn, Schauble, & Mila-Garcia, 1992; Schauble, 1990). However, children have more difficulty evaluating evidence than adults, whose evidence evaluation skills showed more rapid improvement with practice than children's (Kuhn, 1995; Schauble & Glaser, 1990).

The conclusion that children are poor experimental scientists because their evaluation of causal hypotheses tend to be influenced and/or mediated by available beliefs has been challenged on methodological and normative grounds. Methodologically, young children have been shown to interpret correctly contingency evidence and its bearing on causal hypotheses independently of their beliefs if they are given a less cognitively demanding evidence evaluation task than the one used by Kuhn et al. (1988, Bullock, 1991; Bullock, Ziegler, & Martin, 1992; Ruffman, Perner, Olson, & Doherty, 1993; Sodian, Zaitchik, & Carey, 1991). On these tasks, children interpreted contingency evidence that was presented all together (rather than Kuhn et al.'s procedure to present the data sequentially) and then judged the causal status of typically one or two variables (in contrast to judging the causal status of four or more in Kuhn et al.). But perhaps the most important methodological innovation in these studies has been the assessment of children's judgments of the causal status of variables separately from their justifications of those judgments. Kuhn et al. assessed children's judgments of the variables together with their corresponding justification and coded each judgment–justification pair as being either belief-based or evidence-based. However, Bullock (1991; Bullock et al., 1992) found that on a simplified evidence evaluation task, a majority of second and third graders ignored their prior beliefs and correctly judged the causal status of a variable on the basis of contingency data. That is, children's judgments of the variable were consistent with the contingency (covariation and noncovariation) data and inconsistent with their prior belief regarding the variable. However, only a minority of the same children justified their judgments on the basis of the data. Bullock et al. concluded that it was children's tendency to justify judgments on the basis of data rather than their fundamental reasoning processes that change with age and education on evidence evaluation tasks.

Ruffman et al. (1993, Study 1) also used a simplified task to assess Kuhn et al.'s (1988) claim that children fail to evaluate evidence independently of prior beliefs. They found that 5-year-olds correctly judged that two protagonists would arrive at different conclusions about the causal status of the same variable if each protagonist witnessed a different association between the variable and outcome. Moreover, because only the first protagonist was presented as witnessing the "true" data, children were able to correctly interpret the "false" data for the second protagonist despite the data conflicting with the hypothesis that the children believed to be true. Finally, Ruffman et al. (Study 3) found that although a large majority of young children correctly judged the "true" and "false" data, only a small minority of them justified their judgments about the causal status of variables on the basis of the data. They concluded that although children do not justify their judgments about a causal hypothesis on the basis of evidence, they are nonetheless able to correctly evaluate evidence and its bearing on hypotheses independently of their beliefs.

Bullock et al. (1992) and Ruffman et al. (1993) suggested that Kuhn et al.'s (1988) assessment of children's justifications is not only unnecessary but also potentially misleading because children do judge the causal status of variables consistently with contingency data even though they may not refer to the data when justifying their judgments. However, children's belief-based justifications of otherwise consistent interpretations of the data were the basis for Kuhn et al. to claim that the children's judgments were mediated by beliefs that serve to explain the data. From Kuhn et al.'s perspective, Ruffman et al. and Bullock et al. may have merely demonstrated that children explain novel data by forming new beliefs, rather than demonstrating that they evaluate evidence independently of their beliefs.

Other researchers have challenged Kuhn et al.'s (1988) assertion that causal hypotheses ought to be evaluated on the basis of contingency data alone, independently of beliefs. For example, Koslowski, Okagaki, Lorenz, and Umbach (1989) assessed the kinds of information that influence sixth-grade, ninth-grade, and college students' evaluation of causal hypotheses. Koslowski et al. varied whether or not students were told of a plausible mechanism explaining how a variable produces an outcome (e.g., how impurities in a gas additive can lower a car's gas mileage) in addition to the presence or absence of covariation data about the variable and outcome (e.g., how all cars given the gas additive had lower gas mileage than cars not given the additive) among other types of information. They found that all students were more certain of the causal status of a variable when it was described as covarying than not covarying with the outcome and that their causal certainty was greater when a plausible causal mechanism was given than when it was not given. This latter effect was absent for sixth graders when judging the causal status of variables covarying with the outcome. These children were equally (and strongly) certain of the causal status of a variable whether or not they were given an explanation of how it influences the outcome. The authors suggested that sixth graders may have assumed that a plausible causal mechanism exists that explains the connection between the variable and outcome, even though they were not directly told about such a mechanism.

Koslowski et al. (1989) concluded that children, like adults, treat both contingency data and mechanism information as evidence when evaluating causal hypotheses. Moreover, they cited philosophers of science who made the same claim about how practicing scientists actually evaluate causal hypotheses. As such, they argued that Kuhn et al. (1988) were wrong in asserting that children are poor experimental scientists because their evaluation of evidence was influenced or mediated by their beliefs. However, Kuhn et al. (p. 4) never supposed that they were studying how practicing scientists actually evaluate causal hypotheses. Rather, they claimed to be studying whether children would respond like scientists to a challenge regarding why one causal hypothesis was accepted over others. Kuhn et al. supposed that scientists would respond by reflecting on and carefully distinguishing between representations of theories (i.e., their causal beliefs and explanatory concepts) on the one hand and the evidence (i.e., information confirming or disconfirming hypotheses) on the other. It was argued that a failure to reflect on and carefully distinguish between representations of theory and evidence would lead to a melding of the two into a single account of "the way things are" (Kuhn et al., p. 221). Koslowski's sixth-grade students appear to have experienced such a melding of theory and evidence. When presented with data of a covariation between a variable and outcome, the sixth-grade students simply assumed the existence of a causal mechanism connecting the variable to the outcome. That the covariation between a variable and outcome gave only children license to infer a causal mechanism between them goes to the heart of Kuhn et al.'s claim that children, unlike college students and scientists, merge representations of theory and evidence rather than seeking to distinguish between them.

These critiques of Bullock et al. (1992), Koslowski et al. (1989), and Ruffman et al. (1993) may be taken to suggest that even on simplified evidence evaluation tasks, there may be developmental changes in how children interpret contingency data and its bearing on causal hypotheses. Although adults may interpret covariation data and its bearing on causal hypotheses independently of their prior beliefs, children may not and instead interpret such data in light of available (prior or newly formed) beliefs about the causal connection between variable and outcome. In the study presented here, children were assessed for their ability to evaluate contingency data bearing on a causal hypothesis independently of (i.e., not influenced or mediated by) causal or noncausal beliefs regarding variables.

The task employed in this study incorporates three methodological innovations on Kuhn et al.'s (1988) procedure to better assess the development of children's evidence evaluation skills. First and foremost, like Bullock et al. (1992) and Ruffman et al. (1993), an evidence evaluation task was used that was less cognitively demanding than the one used by Kuhn et al. In the task presented here, participants were assessed for their judgments and justifications regarding the causal status of each of two variables that appeared in each of two data sets. Each data set presented perfect covariation or noncovariation evidence for one variable believed to causally related to an outcome and one variable believed to be noncausally related. Thus, children were presented with four data sets but assessed the influence of only one variable in each data set. Moreover, the instances in each data set were presented all at once rather than sequentially. We presumed that children's ability to correctly interpret the evidence and its bearing on causal hypotheses would be enhanced by reducing the number of variables reasoned about and presenting the instances of contingency data all at once rather than sequentially.

Second, like Bullock et al. (1992), children's causal judgments were assessed independently of their justifications. This is in contrast with Kuhn et al. (1988), for whom judgments and justifications were coded together. As a result of coding the data in such a manner, children in the study presented here may be shown to make judgments of the causal status of variables that are consistent with the contingency data but fail to justify such judgments by reference to the data. Third, like Koslowski et al. (1989) participants' judgments were assessed on a scale reflecting their certainty that the variable was causal, not causal, or neither causal nor noncausal (see the Methods section). Such a coding technique was thought to be a more realistic and sensitive measure of participants' judgments of the causal status of a variable than would be having children make forced-choice causal or noncausal judgments (Acredolo & O'Conner, 1991).

The study presented here employs three ways of measuring whether or not participants' evaluation of evidence was influenced or mediated by their available beliefs. First, they were assessed for whether their judgments of the causal status of variables reflected the influence of prior causal or noncausal beliefs. Although Bullock et al. (1992) found no such influence in 8 and 9-year-olds, the study presented here corrects what are taken to be a number of features of Bullock et al.'s study that may have led to an overestimation of children's evidence evaluation skills. Bullock et al. presented the covariation evidence to participants grouped according to outcome (e.g., all the positive outcome instances were grouped together on one side of a page and the negative outcome instances on the other). Such an organization may have unduly highlighted the covariation or noncovariation structure of the data for children. In addition, children's prior beliefs regarding the causal or noncausal status of variables were not assessed. On the face of it, children's prior beliefs were probably held without the conviction arising from knowledge of the task domain. It is likely that children had only a limited understanding of two task domains used by Bullock et al.: the aerodynamics of flying a kite and the design of lanterns. Thus, Bullock at al.'s conclusion that children correctly evaluate contingency evidence independently of prior beliefs may not generalize to cases where there is no highlighting of the contingency structure of a data set and children's causal and noncausal beliefs are held with a strong conviction.

In the study presented here, children and adults were selected for having strongly held prior causal and noncausal beliefs regarding the influence of variables on an outcome. Pilot research indicated that children and adults alike strongly believe that growing healthy plants requires giving it sun and can cite a causal mechanism (albeit not always photosynthesis) connecting the cause to the effect. Children and adults also strongly believe that a "charm," represented by a four-leaf clover, makes no difference in growing healthy plants because there is no causal mechanism connecting the variable to the outcome. To ensure that the same prior beliefs regarding the variables were held by all participants, each was initially interviewed about their beliefs. On a subsequent interview, each subject was presented with four data sets, representing plants grown by four different people. Each data set contained at least four instances of a plant (which was either healthy or sick) with each instance associated with either the presence or absence of sunshine or a clover. The instances were presented all together and there was no additional organization or grouping of the contingency data. On the presentation of each data set, participants were asked to make judgments of the causal status of the variable.

The second measure of the influence or mediation of belief on children's evaluation of evidence was their justifications. As previously noted, justification data were used by Kuhn et al. (1988) to identify children who failed to evaluate evidence independently of beliefs. However, Ruffman et al. (1993) and Bullock et al. (1992) argued that justification data were misleading about children's true evidence evaluation skills, a concern echoed by others (Sodian et al., 1991). As a result, a third measure was developed to augment the other two as a means to assess whether or not children interpret evidence independently of prior beliefs. If children's evaluation of evidentially relevant contingency data is influenced or mediated by beliefs, then those beliefs may influence or mediate their evaluation of other, evidentially irrelevant data. For example, children's available beliefs may influence or mediate their evaluation of instances of missing data, the presence of which neither confirms nor disconfirms a causal hypothesis.

In the study presented here, participants in each age group were assigned to either the control condition and presented with contingency data in each of the four data sets or to one of three missing data conditions and given instances of missing data in addition to the contingency data in each of the four data sets. Participants assigned to a missing data condition were told that the person who grew the plants forgot the status of the variable (e.g., whether the sun or charm was present or absent), and/or the outcome (e.g., whether the plant was healthy or sick) on the "missing data" instances, but correctly remembered the status of the variable and outcomes on the other instances. There were a total of three missing data conditions that varied whether the status of the variable, outcome, or both variable and outcome were unknown.

If children evaluate evidence independently of prior beliefs, then their judgments of the causal status of variables in control and missing data conditions should be no different from each other. Participants in the missing data conditions would judge the causal status of variables on the basis of the evidentially relevant contingency data alone, ignoring the evidentially irrelevant missing data. However, children may be influenced by prior beliefs such that they use their beliefs to "read into" the instances of missing data and explain them. For example, children who believe that sun makes plants healthy may use such a belief to explain cases where there is a healthy plant but the status of sun is unknown. The children may reason that the sun had to have been there for the plant to be healthy. If participants treat missing data as additional instances that can be explained by their prior beliefs, those in the missing data conditions would be more likely to judge variables consistently with their prior beliefs (e.g., that sun is causal and charm is noncausal) than children in the control condition who have no missing data instances to reason about.

In contrast, instead of being influenced by prior beliefs, children's evaluation of evidence may be mediated by (perhaps newly formed) beliefs that are used to explain the contingency data. If such beliefs are used to explain the contingency data, they may also be used to "read into" and explain the instances of missing data. For example, children observing charm covarying with healthy plants may form the belief that the charm really does do something to make plants healthy. Such a belief can be used not only to account for the contingency data but also to reason that a healthy plant with an unknown status of charm must have had charm to make it turn out the way it did. If participants explain missing data in light of mediated beliefs, then participants in the missing data conditions should judge variables more consistently with the contingency data (e.g., causal when they covary with plant health and noncausal when they do not) than those in the control condition, who have no instances of missing data to reason about.

In summary, this research addresses children's ability to evaluate evidence independently of beliefs. It was stressed that beliefs may not only influence but also mediate participants' evaluation of evidence. In this study, the influence, mediation, or both of belief on the evaluation of evidence was assessed by soliciting children's, non-college-educated adults', and college students' judgments and justifications of the causal status of variables—one believed causally and one noncausally related to an outcome—that were covarying or not covarying with an outcome. In each age or education group, participants were assigned to one of three conditions in which they were given instances of missing data in addition to instances of contingency data or a control condition in which they received no instances of missing data. The task was designed to minimize cognitive demands on participants by presenting the contingency data all together and in a univariable context. However, participants were selected for holding strong prior beliefs regarding the variables, and the data were presented without any additional organization of the instances.

METHOD

Participants

Two hundred thirty-six people took part in the study, 161 children and 75 adults. Two age groups of children were recruited from elementary schools in a western Canadian province. There were a total of 76 second- and third-grade students (41 boys and 35 girls, *M* = 8;6 years; range = 7;4–10;3 years). There were a total of 85 sixth- and seventh-grade students (35 boys and 50 girls, *M* = 12;5 years; range = 10;8–14;0 years). There were two groups of adult participants: 40 college students, 22 males and 18 females (*M* = 24 years; range = 18–45 years); and 35 non-college-educated adults, 17 males and 18 females (*M* = 29 years; range = 18–55 years). The non-college-educated adults were solicited from members of an electrical union and the maintenance staff of the University of Saskatchewan. The mean ages of participants in the non-college- and college-educated groups were not statistically different.

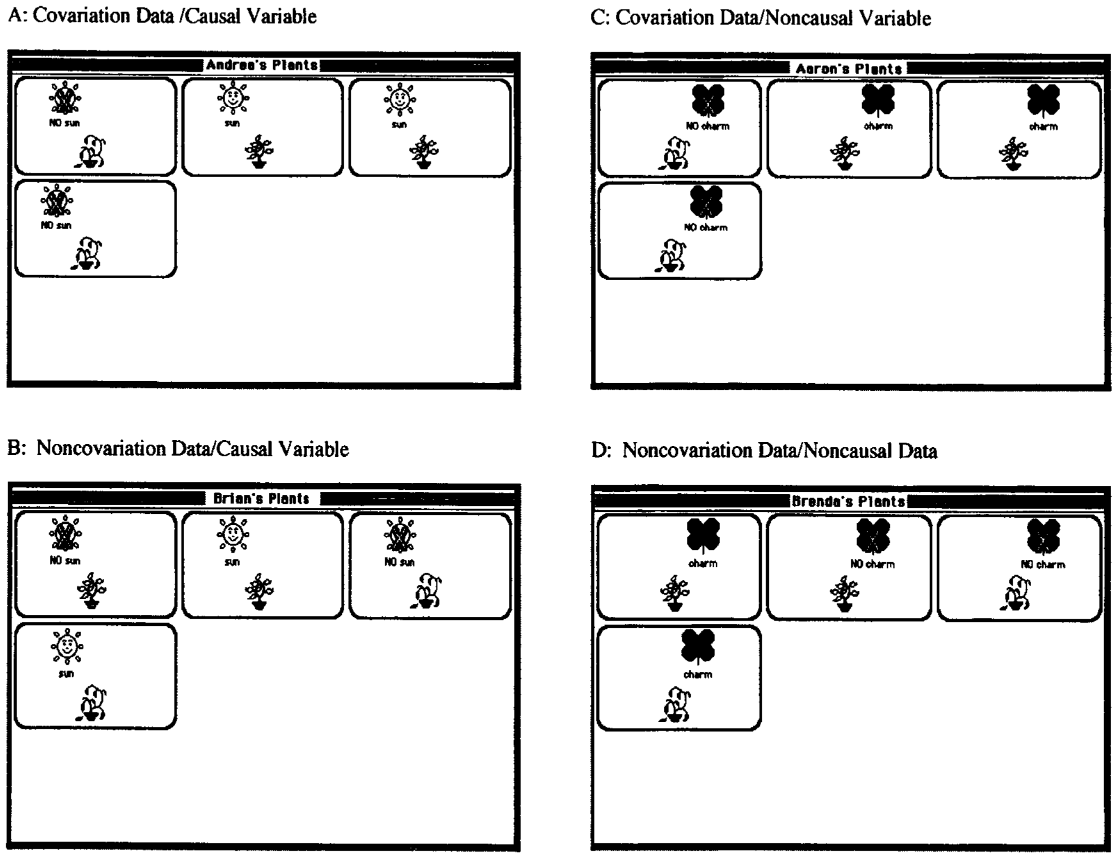
Task

Participants were initially interviewed regarding their prior beliefs that the presence of the sun is causally related to the growth of healthy plants and that the presence of a charm is not. At a second interview, they were asked to make judgments regarding the causal status of the charm or sun variables solely on the basis of information regarding the status of plants (healthy or sick) grown with or without the target variable. They made judgments for each of four data sets, with each data set composed of line drawings depicting instances of healthy and sick plants associated with the presence or absence of a variable.

Each data set was described as the drawings of plants grown by a friend of the experimenter. In any given data set, there were drawings of at least two healthy and two sick plants. The healthy plants were drawn with many leaves and branches whereas the sick plants were drawn with no leaves and drooping branches. Associated with each plant in a data set was a drawing of a sun or a charm (a four-leaf clover). Those who participated were told that each of the plants was treated similarly by the person except for whether or not the target variable was present or absent. A plant drawn with a variable (charm or sun) meant that the variable was present when that plant was grown, whereas a plant drawn with an "X" over the variable meant that the variable was absent when the plant was grown (see Figure 1).

Each of the four data sets presented one of two patterns of contingency data. The patterns presented either a perfect positive correlation (covariation data) or a zero correlation (noncovariation data) between the presence or absence of a variable and the health or sickness of the plants. For the covariation data, the healthy plants in a data set were associated with the presence of the variable, and the sick plants were associated with the absence of the variable. For the noncovariation data, the healthy and sick plants in a set were each associated equally often with the presence and absence of the variable.

The four data sets that each participant evaluated represent a complete factorial of two levels of prior belief (causal belief in sun and noncausal



A: Covariation Data /Causal Variable

C: Covariation Data/Noncausal Variable

B: Noncovariation Data/Causal Variable

D: Noncovariation Data/Noncausal Data

Figure 1. Items used in the control condition.

belief in charm) and two levels of contingency data (covariation and noncovariation). As seen in Figure 1, Andrea's plants present covariation data regarding sun (the variable believed to be causal), Brian's plants present noncovariation data for sun, Aaron's plants present covariation data for charm (the variable believed to be noncausal), and Brenda's plants present noncovariation data for charm.

These four sets of plants presented in Figure 1 were given to participants in the control group. There were three other conditions in which subjects received data sets each of which contained instances of missing data in addition to the contingency data. Participants in the variable unknown condition received the four data sets, each of which included the contingency data and two additional instances of a healthy and a sick plant. However, whether the plants did or did not get the target variable (sun or charm) on these additional instances was unknown, as represented by a "?" over the drawing of the variable. Participants in the outcome unknown condition received two extra instances in each data set, one associated with the presence of a target variable and the other with its absence. However, the health or sickness of the plant in each of these additional instances was unknown, represented by a "?" replacing the drawing of the plant. Participants in the both unknown condition also received two additional instances in each data set, but the health or sickness of each plant was unknown as was the presence or absence of the target variable.[1]

Procedure

The task involved two separate interviews (Prior Belief and Causal Judgment Interviews) approximately 3 days apart. All interviews were carried out individually and were tape recorded for later coding. All children were tested in school during the school day. College students were interviewed on campus and adults were interviewed at their home or place of work.

Prior Belief Interview. The first interview consisted of questions regarding participants' beliefs about what does and does not make a difference in whether a plant grows to be healthy or sick and the consistency and strength of those beliefs. To elicit their beliefs, participants were presented

[1] To equate the features of the data sets presented to participants in each condition, plant instances of the same size were drawn within a standard-sized (7 in. × 4.25 in.) rectangular border such that the drawing of either four (control condition) or six (missing data conditions) instances of plants within a border did not completely fill up the space. That is, each data set drawn for each condition was such that there was remaining space for additional plant instances within the border. Participants could have inferred that because there was additional space within the border of a given data set, other plants were grown and their results were unreported. There was no indication from participants' justifications that they made any such inference. However, even if such inferences were made, they should have been made for each data set in each condition, therefore not biasing the results.

with two pairs of line drawings. One pair of drawings was of a healthy and a sick plant and the second pair of drawings was of two healthy plants. Participants were first presented with drawings of the one healthy and one sick plant, and told the following:

> These are pictures of two of my plants. These are the same kind of plants and both started out healthy. But now, this one is healthy, and this one is sick. I must have treated these plants differently for them to grow so differently. For example, maybe I watered this healthy plant but did not water this sick plant and that is why they grew so differently. What are some other ways I may have treated this healthy plant differently than I treated this sick plant?

If a participant did not initially mention the role of sun in promoting the growth of healthy plants, he or she was prompted with the following question, "Is there any other way I may have treated these plants differently that would have made them grow so differently?" Typically, participants arrived at the possible causal influence of the sun with only this minimal amount of prompting. However, if the prompts failed to induce them to spontaneously discuss the role of sun in the growth of healthy plants, they were asked, "What about sun? Does giving one plant sun and not giving the other plant sun make a difference in whether the plants are healthy or sick? Yes, No, or Maybe?"

The interview continued only if a participant believed that sun made a difference in the growth of plants. Then participants' consistency was assessed by asking, "So, does giving a plant sun make a difference in how the plant grows? Yes or no?" After assessing participants' consistency, their certainty was assessed by asking, "How sure are you that sun makes a difference whether plants are healthy or sick: a little sure, pretty sure, or very sure?"

A similar sequence of questions was then asked regarding participants' belief in the noncausal status of charm. They were shown the second pair of pictures and were told the following:

> Here I have pictures of two more of my plants. These are the same kind of plants, and they started out healthy. I treated these plants differently, but it didn't make a difference in how they turned out—see, they are both healthy. For example, maybe I grew this plant next to a picture of my brother, and there was no picture next to this other plant. But they both grew to be healthy. What are some other ways I may have treated these plants differently that didn't make a difference in how they turned out?

If a participant did not spontaneously discuss the noncausal status of charm (which was often the case because there were an infinite number of variables not influencing the growth of healthy plants), he or she was asked, "What about a charm? Does putting a charm next to one plant and not putting a charm next to the other plant make a difference in whether the plants are healthy or sick? Yes, No, or Maybe?" If a participant believed that charm was noncausal, she or he was asked the consistency question, "So, putting a charm next to a plant makes no difference in how the plant grows? Yes or no?" Then those who were consistent in believing that charm is a noncausal variable in the growth of healthy plants were asked the certainty question, "How sure are you that putting charms next to plants makes no difference in how plants grow: a little sure, pretty sure, or very sure?"

Participants were excluded from the second interview based on the criteria of not consistently judging that sun makes a difference and that charm makes no difference in the growth of healthy plants and not being at least "pretty sure" of their judgment. A total of eight second- and third-grade students, two sixth- and seventh-grade students, and one non-college-educated adult were dropped from the study because of their responses during the prior belief interview. They were not included in the description of the sample.

Causal Judgment Interview. Following the first interview, participants in each age group were randomized into one of four conditions (a control group or one of the three experimental groups) and interviewed a second time. Participants in each condition were told that they were going to see drawings of the plants of the interviewer's friends and that these people know nothing about growing plants except what happened to their own plants. The interview began with a practice trial, which featured drawings of five plants that "Terry" grew and the status of the variable "water" associated with each plant: Healthy plant with water present, sick plant with water absent, healthy plant with water status unknown, unknown plant outcome with water present, unknown plant outcome with water status unknown. The experimenter presented the five drawings of plant and variable instances on a single sheet of paper and then described each instance. Each participant was then asked to pick out each instance (i.e., can you show me the plant that Terry watered but forgot whether it was sick or healthy?). Finally, in the presence of the data, those who participated were asked a practice causal judgment question. They did not receive any feedback regarding their judgment on the practice trial.

Following the practice trial, participants were presented with the plants of Andrea, Brian, Aaron, and Brenda. The order of presentation of each set of plants was randomized. On presentation of a set of plants, the experimenter described each instance, and participants were reminded to answer the questions only on the basis of the information in front of them and not on the basis of what they know about growing plants. In the presence of the first data set, participants were first asked the judgment question: "Does giving plants (sun, a charm) or no (sun, charm) make a difference in the plants being healthy or sick: Yes, No, or Maybe?" Following the judgment question, participants were asked the justification question, "Why do you say that (sun, charm) (makes, does not make) a difference?" If no justification was elicited by the first way of posing the question, a second way was used: "How do you know that (sun, charm) (makes, does not make) a difference?" Then participants were asked the certainty question: "How sure are you that (sun, charm) (makes does not make) a difference: A little sure, pretty sure, or very sure?" If a participant was initially uncertain about the causal status of the variable (i.e., responded "maybe"), she or he was asked the justification question but not the certainty question. If the participant was unable to explain his or her uncertainty when asked the justification question, he or she was asked, "Would someone who knows a lot about growing plants be able to tell whether giving plants (sun, a charm) or no (sun, charm) makes a difference in these plants being healthy or sick?" This question was used as a basis for distinguishing between participants who were confused about how to interpret the data and those who truly believed that there is no determinate answer to the question. We supposed that participants who experienced true indeterminacy would hold that even an expert would also be indeterminate as well. This question was asked infrequently (4% of all data sets). After completing the questioning for one data set, it was removed and a new data set was randomly selected and presented to participants and the sequence of questions were asked again. This procedure continued until all four data sets were presented.

Coding

Responses to the judgment and certainty questions for each data set were combined into a single 7-point scale of causal certainty. Each point on the scale reflects a particular answer to the judgment and certainty questions: 1 (*very sure that the variable is not causal*), 2 (*pretty sure the variable is not causal*), 3 (*a little sure the variable is not causal*), 4 (*the variable is neither causal nor noncausal*), 5 (*a little sure the variable is causal*), 6 (*pretty sure the variable is causal*), and 7 (*very sure that the variable is causal*).

Following Kuhn et al. (1988), responses to each justification question were coded as either evidence-based or belief-based, independently of their responses to the judgment and certainty questions. A justification was coded as evidence-based if it involved a direct or indirect reference to the contingency data. An example of a direct evidence-based justification is, "The sun makes a difference because each time a plant in the picture had sun it was healthy," whereas an example of an indirect evidence-based justification is, "The charm is causal because healthy plants have charm." The latter does not make direct reference to the contingency data in the data set, although there is no reason to believe the participant was referring to any plants other than those in the data set. Included as evidence-based justifications were indeterminate judgments that were justified by direct or indirect reference to the data. A example of an indeterminate judgment justified by a direct reference to the data is, "Can't tell (whether or not sun makes a difference) . . . because sun made the plant healthy here (pointing to an instance) but made it sick here (pointing to another instance)." An indeterminate judgment justified by an indirect reference to the data was one in which participants claimed that an expert could not tell whether or not a variable has a causal influence on the plants. Such a justification was credited as evidence-based because it reflected a judgment about the indeterminacy of the data that no one including an expert could resolve.

All other justifications were coded as belief-based, including ones involving a reference to the participant's or others' knowledge (or lack thereof) or beliefs about growing plants. A failure to respond or an irrelevant response to the justification question (i.e., one having nothing to do with the variable or with growing plants, e.g., "because the pot was big") were also coded as belief-based. Coding all these responses as belief-based is in keeping with Kuhn et al. (1988), as each reflects a failure to consider or make reference to the contingency data alone in making judgments of the causal status of variables. Interrater reliability for judging 40 randomly selected participants (160 justifications) was 92%. The two scorers resolved their disagreements through discussion.

RESULTS

The data regarding age or educational differences in participants' responses to the judgment and certainty questions on each of the four data sets (coded together on a 7-point scale of causal certainty) were analyzed first, followed by an analysis of their responses to the justification questions.

Causal Certainty Judgments

Causal certainty scores for each of the four data sets were subjected to a 4 (Group: Second and Third Grade, Sixth and Seventh Grade, Adults, and College Students) × 2 (Prior Belief: Causal vs. Noncausal) × 2 (Contingency: Covariation vs. Noncovariation Data) × 4 (Condition: Control, Variable Unknown, Outcome Unknown, Both Unknown) mixed-model analysis of variance (ANOVA). Prior belief and contingency data were within-subject variables, and group and condition were between-subject variables. This analysis, like all other ones, employed a correction for unequal cell sizes (Levine, 1991). The results of the ANOVA are reported in terms of age or educational differences in the effect of (a) contingency data, (b) prior belief, and (c) condition on participants' causal certainty scores.

Contingency Data. With respect to the influence of contingency data on causal certainty scores, the ANOVA revealed a main contingency data effect, $F(1, 220) = 318.76, p < .001$, and a Group × Contingency Data interaction, $F(3, 220) = 34.30, p < .001$. Table 1 presents each group's mean causal certainty score for the sun and charm variables when they covaried with plant health (covariation data) and when they did not (noncovariation data). When the variables covaried with plant health, the mean causal certainty score for each group was above 4, reflecting a causal judgment on the scale, and when the variables did not covary, each group's mean score was below 4, reflecting noncausal judgments. Simple effects analysis revealed that each group had a higher mean causal certainty score for the variables when they covaried with plant health than when they did not covary, $F(1, 220)$ ranging from 40.77 to 260.60, all *ps* $< .001$. However, a Newman–Kuels post hoc test ($p < .05$) showed that the difference between mean scores for covariation and noncovariation data was significantly larger for college students than for participants in the other groups, who did not differ from each other (see Table 1). Thus those in each group judged the causal status of variables differently when they covaried than when they did not covary with plant health, although, compared to all other participants, the college students were most different in their judgment of the variables' causal status. This result supports Ruffman et al.'s (1993) finding that children can arrive at different conclusions about the causal status of a variable if they observe different data regarding the variable's contingency with an outcome.

Prior Belief. The ANOVA also revealed a main effect of prior belief, $F(1, 220) = 284.62, p < .001$, and a Group × Prior Belief interaction, $F(3, 220) = 6.16, p < .001$. Table 1 also presents each group's causal certainty score for judgments regarding the sun and charm variables, averaged over judgments made for each variable covarying and not covarying with plant health. Each group's mean causal certainty score for the sun variable was above 4, reflecting a causal judgment on the scale, and the score for the

Table 1. Mean Causal Certainty Score by Group for Contingency Data and Prior Belief

		Contingency Data			Prior Belief		
Group	*n*	Cov.	Noncov.	Difference	Causal	Noncausal	Difference
Second and third	76	4.93	3.93	1.00	5.63	3.24	2.39
Sixth and seventh	85	4.93	3.76	1.17	5.72	2.97	2.75
Adults	35	5.29	3.61	1.68	5.67	3.23	2.44
College	40	6.51	2.99	3.52	5.28	4.22	1.06

Note: The causal certainty score ranges from 1 *very sure the variable is noncausal,* to 4 *the variable is neither causal nor noncausal,* to 7 *very sure the variable is causal.* A score below 4 reflects a noncausal judgment, and a score above 4 reflects a causal one. Cov. = covariation; Noncov. = noncovariation.

charm variable was below 4, reflecting a noncausal judgment. Simple effects showed that participants' mean causal certainty score was higher for the sun than the charm variable, $F(1, 220)$ ranging from 10.03 to 146.48, all $ps < .01$, although a Newman–Kuels post hoc test ($p < .05$) revealed that the college students had a significantly smaller difference between mean scores for the sun and charm variables than the other groups, who did not differ from each other (see Table 1). Thus, participants in each group tended to have an overall bias to judge sun as causal and charm as noncausal, despite each variable having been presented with the same evidence. However, the judgments of college students were less biased than all the other participants.

A three-way interaction between group, prior belief, and contingency data, $F(3, 220) = 3.35, p = .02$, also proved to be significant in the ANOVA. The mean causal certainty scores of each age group for each data set is presented in Figure 2, along with the hypothetical causal certainty scores of an ideal reasoner for each data set. The scores of the ideal reasoner reflect judgments of the causal status of a variable for each data set made on the basis of the evidence alone, with no influence or mediation of beliefs. An ideal reasoner would be very certain of the causal status of the sun and charm variables (score of 7) when each covaried with the health of plants and equally certain of the noncausal status of each (score of 1) when they did not covary with plant health.

Figure 2 shows that the causal certainty scores of each group are generally similar to the ideal reasoner for the data sets where the contingency data confirmed participants' prior beliefs. Prior beliefs were confirmed in two data sets: when sun covaried with the plant health (covariation/sun: Andrea's plants) and when charm did not (noncovariation/charm: Brenda's plants). Each group was on average "pretty sure" that sun was causal when it covaried with plant health (an ideal reasoner would be "very sure") and was between "pretty sure" and "a little sure" of the noncausal status of charm when it did not noncovary with plant health (an ideal reasoner would be "very sure"). A series of t tests confirmed that each age group had significantly higher mean causal certainty scores for sun when it covaried with plant health than they did for charm when it did not covary with plant health, second- and third-grade: $t(75) = 15.5, p < .001$; sixth- and seventh-grade: $t(84) = 20.36, p < .001$; adults: $t(34) = 12.87, p < .001$; college students: $t(39) = 17.74, p < .001$.

Figure 2 also shows that when the contingency data disconfirmed participants' prior beliefs, there were differences in the pattern of judgments for the children, adults, and college students. The college students were the group most similar to an ideal reasoner because, like the ideal reasoner, the college students tended to ignore their prior beliefs and make causal certainty judgments for variables in a manner consistent with the contingency data. On average, the college students judged that charm was causal when

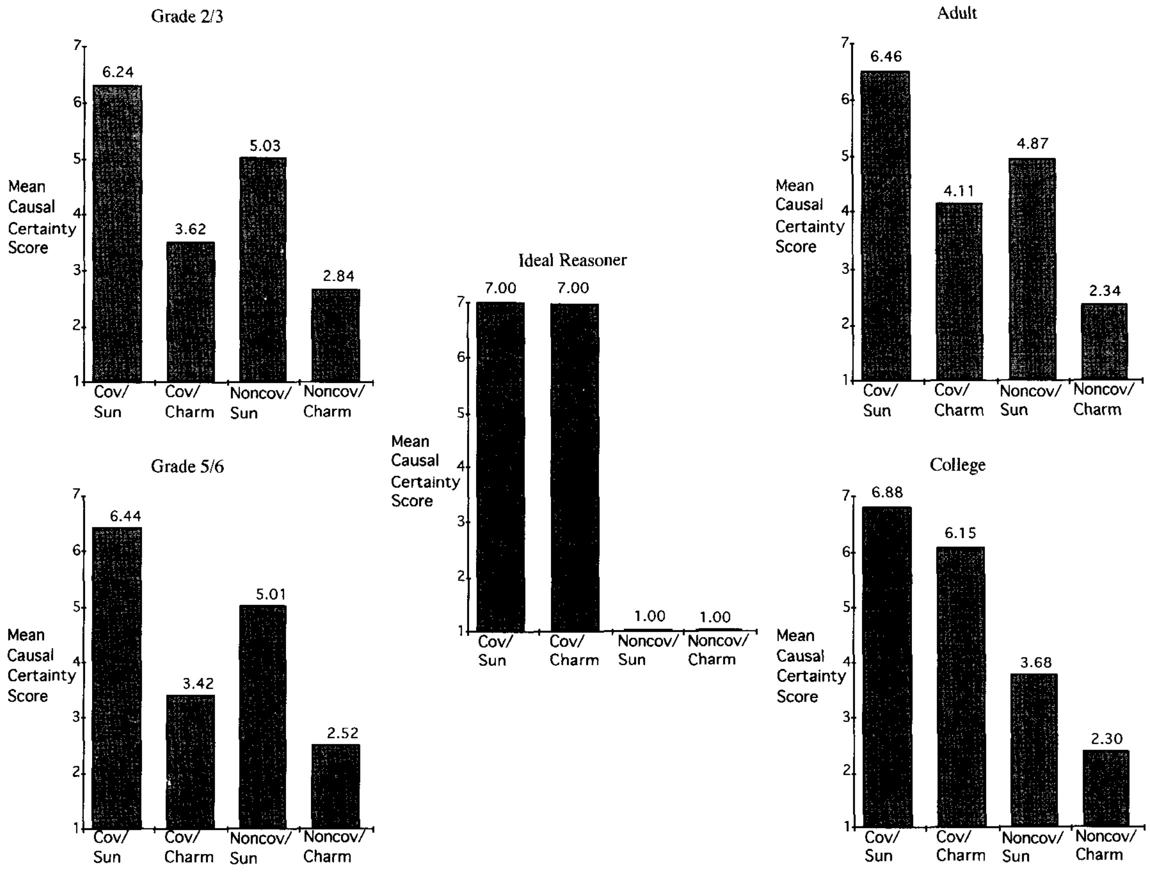


Figure 2. Mean causal certainty scores for each data set by group.

it covaried with plant health and that sun was noncausal when it did not covary. The difference in causal certainty judgments for charm given covariation evidence and sun given noncovariation evidence was significant, $t(39) = 5.52$, $p < .001$. In contrast, the second and third and sixth and seventh graders tended to judge the causal status of variables disconfirmed by the data in a manner consistent with their prior beliefs. On average, the children were "a little sure" that *sun* was causal even though it did not covary with plant health and were between "a little sure" and "pretty sure" that charm was noncausal even though it did covary with plant health. Separate *t* tests showed that each group of children had significantly higher causal certainty scores for sun given noncovariation data than for charm given covariation data, which is the opposite pattern of scores on the same data sets from college students, second- and third-grade: $t(75) = 3.90, p < .001$; sixth- and seventh-grade: $t(84) = 5.24, p < .001$. Finally, the judgments of the adults were unlike the other groups in that their causal certainty scores for sun not covarying with the health of plants was no different than their scores for charm covaring with plant health, $t(34) = 1.23$, n.s.

Thus, the Age × Prior Belief × Contingency interaction effect is explained by children and adults making judgments about the causal status of variables that were similar to each other (and to an ideal reasoner) when participants' prior beliefs about the variables were confirmed by the contingency data. However, when the contingency data disconfirmed participants' prior beliefs about the variables, children tended to judge the variables consistently with their beliefs (unlike the ideal reasoner), the college students tended to judge the variables consistently with the evidence (like the ideal reasoner), and the judgments of the adults tended to be in between the children and the college students. The pattern of judgments suggests that, on this task at least, there are differences associated with age and education in judging the causal status of variables independently of the influence of prior beliefs.

Condition. With regard to the effect of missing data on participants' causal judgments, the prediction was that children may not simply ignore instances of missing data (as an ideal reasoner would) but "read into" them additional instances explained by either their prior or newly formed beliefs about the causal status of the variables. As a result, children in the missing data conditions (variable unknown, outcome unknown, and both unknown) were expected to judge the variables to be more consistent with either their prior beliefs or the contingency data than children in the control condition, who receive no missing data instances to reason about.

The results of the ANOVA revealed a significant Condition × Contingency interaction effect, $F(3, 220) = 6.07, p = .001$. Table 2 presents participants' mean causal certainty scores for variables covarying (covariation data) and not covarying (noncovariation data) with plant health and the difference between those means. Inspection of Table 2 shows that the difference between participants' mean causal certainty scores for variables covarying and not covarying with the outcome is larger in the control condition (in which the data sets contained no instances of missing data) than in each of the missing data conditions. This observation was largely confirmed by a series of *t* tests, control vs. variable unknown: $t(119) = 3.21$, $p < .01$; control vs. outcome unknown: $t(122) = 1.92$, $p = .057$; both unknown: $t(117) = 3.29, p < .001$. The results suggests that, contrary to initial predictions, participants were not explaining the missing data. Instead, the presence of instances of missing data in the data sets made them less certain that variables covarying with the outcome are causal and those not covarying are noncausal. As a result, there is a smaller difference between participants' mean scores of variables covarying and not covarying with the outcome in each missing data condition than in the control condition.

Notably, the presence of instances of missing data did not additionally confuse participants about the meaning or significance of the contingency data presented in the data sets. The mean number of "can't tell" responses given by control participants (who had no instance of missing data in their data sets) over the four data sets ($M = .86$) was no different than the number given by participants in each other condition (variable missing: $M = 1.00$; outcome missing: $M = .88$; both missing: $M = .98$), $F(3, 234) = .34$, n.s. Moreover, as will be discussed later, participants in the control condition were no more likely to give evidence-based responses to the justification question than were children in the other conditions. Thus, participants in the missing data conditions were not more confused by the data although they were less certain of their causal and noncausal judgments than those in the control condition.

In addition to the Condition × Contingency interaction effect, there was also a three-way Group × Condition × Contingency interaction effect that approached significance, $F(9, 220) = 1.79, p = .07$. Follow-up *t* tests computed within each age group compared the difference in mean scores for variables when they covaried and did not covary with plant health for

Table 2. Mean Causal Cetainty Score by Contingency Data and Condition

		Contingency Data		
Condition	***n***	**Covariation**	**Noncovariation**	**Difference**
Control	64	5.59	3.39	2.20
Variable unknown	57	5.08	3.83	1.25
Outcome unknown	60	5.28	3.69	1.59
Both unknown	55	5.02	3.78	1.24

Note. The causal certainty score ranges from 1 *very sure the variable is noncausal,* to 4 *the variable is neither causal nor noncausal,* to 7 *very sure the variable is causal.* A score below 4 reflects a noncausal judgment, and a score above 4 reflects a causal one.

participants in the control and experimental conditions. To specifically test for an effect of missing data, the difference scores of participants in the control condition were compared to the difference scores of those in all three experimental conditions combined (See Table 3). The results demonstrated that second and third graders had a higher difference score in the control condition ($M = 1.74$) than in the combined missing data conditions ($M = .72$), $t(74) = 3.30, p < .001$. Similarly, the sixth and seventh graders had a higher difference score in the control ($M = 1.91$) than the missing data conditions ($M = .91$), $t(83) = 3.65, p < .001$. The adults' difference scores in the control ($M = 1.77$) and the missing data ($M = 1.64$) conditions were not significantly different, $t(33) = 0.19$, n.s. Similarly, the college students' mean evidence scores in the control ($M = 3.83$) and the experimental ($M = 3.39$) conditions were not significantly different, $t(38) = 0.78$, n.s. The results suggest that the presence of instances of missing data influenced children's, but not adults' or college students', certainty of the causal or noncausal status of variables when they covaried and did not covary with plant health. The effect was to make the children but not the adults or college students less certain of their causal or noncausal judgments.

Justifications

To assess whether or not participants referred to the data when they justified their judgments of the causal status of variables, their responses to each justification question were coded as belief-based (scored as 0) or evidence-based (scored as 1). (Six participants, two sixth and seventh graders, three adults, and one college student, were dropped from this analysis because their justifications were not recorded.) The dichotomous data were subjected to a 4 (Group) × 2 (Prior Belief) × 2 (Contingency Data) × 4 (Condition) mixed-model ANOVA, with prior belief and contingency data as

Table 3. Mean Causal Certainty Score by Group and Contingency Data for Participants in the Control and Combined Missing Data Conditions

	Control				Missing Data			
Group	*n*	Cov.	Noncov.	Difference	n	Cov.	Noncov.	Difference
Second and third	21	5.43	3.96	1.47	55	4.75	4.03	0.72
Sixth and seventh	22	5.23	3.32	1.91	63	4.83	3.92	0.91
Adults	9	5.33	3.56	1.77	26	5.27	3.63	1.64
College	12	6.71	2.88	3.83	28	6.43	3.04	3.39

Note. The causal certainty score ranges from 1 *very sure the variable is noncausal,* to 4 *the variable is neither causal nor noncausal,* to 7 *very sure the variable is causal.* A score below 4 reflects a noncausal judgment, and a score above 4 reflects a causal one. Cov. = covariation; Noncov. = noncovariation.

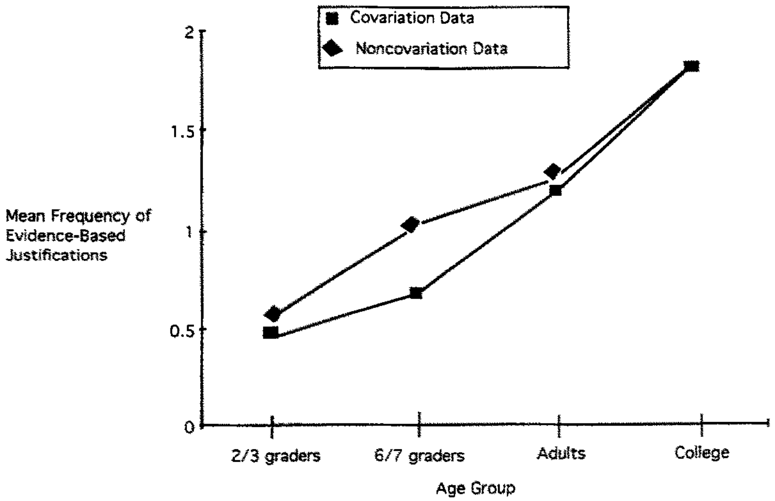


Figure 3. Frequency of evidence-based judgments by contingency data for each group.

within-subject variables and group and condition as between-subject variables. There was a main group effect, $F(3, 214) = 34.35, p < .001$. A Newman–Kuels post hoc test ($p < .05$) revealed that there was a significant difference in the mean frequency (maximum = 4) of evidence-based justifications among adjacent age or education groups (second and third graders: $M = 1.01$; sixth and seventh graders: $M = 1.71$; adults: $M = 2.44$; college students: $M = 3.69$).

There was also a contingency data main effect, $F(1,214) = 18.40, p < .001$, due to participants giving more evidence-based justifications for the noncovariation ($M = 1.04$ out of two noncovariation data sets) than the covariation ($M = 0.84$) data. However, the main effect was modulated by a significant Group × Contingency Data interaction effect, $F(3, 214) = 3.96, p = .009$. An analysis of simple effects revealed that only the sixth and seventh graders had a significantly higher mean frequency of evidence-based justifications when the variables did not covary than when they did covary with plant health, $F(1, 214) = 27.29, p < .001$ (see Figure 3). The results suggest that there are differences associated with age and education in participants' tendency to justify causal judgments on the basis of contingency data, even though such judgments may have been made on the basis of the data.

DISCUSSION

The results of this research point to sources of both continuity and discontinuity between children and adults in their evidence evaluation skills. On the one hand, the mean causal certainty score for each age or education group of participants was significantly different for variables when they covaried than when they did not covary with plant health. Overall, members of each age or educational group judged that variables were causal when they covaried with plant health and that the same variables were not causal when they did not covary with plant health. Such a continuity between the performance of children and adults confirms the results of other studies (Bullock, 1991; Bullock et al., 1992; Koslowski et al., 1989; Ruffman et al., 1993) showing that children evaluate the same variables differently depending on their contingent relation with an outcome.

On the other hand, there were three clear examples of discontinuity in the evidence evaluation skills of the two groups of children compared to the two adult groups. First, the two groups of children made evidence-based justifications for only a minority of their judgments, whereas the adults and college students gave evidence-based justifications for a majority of their judgments. This finding also confirms the results of other researchers (Bullock, 1991; Bullock et al., 1992; Kuhn et al., 1988; Ruffman et al., 1993), although Bullock et al. and Ruffman et al. denied Kuhn et al.'s claim that the difference in children's and adults' tendency to give evidence-based justifications reflects a fundamental change in their process for evaluating evidence.

Second, when the contingency between a variable and plant health disconfirmed participants' prior belief regarding the variable, the two groups of children, unlike the adults and college students, tended to make judgments regarding the variable that were consistent with their prior beliefs. However, when the contingency data confirmed participants' prior beliefs, the children and adults made similar judgments. This result confirms the results of Klahr, Fay, and Dunbar (1993), who found that third and sixth graders were less likely than community college and college students to successfully test hypotheses about the correct function of a repeat key of a programmable robot when the function was believed to be implausible (13% vs. 56%). However, when the function was believed to be plausible, the success rate of children (75%) and adults (88%) were similar.

The influence of prior beliefs on children's evaluation of evidence may have been detected in the research presented here but not in other studies in part because only in the research presented here did participants have rich knowledge of the task domain and hold strong beliefs about the variables they were evaluating. One methodological conclusion to draw in comparing the results of this study with the others is that the nature of the task can highlight either children's evidence evaluation competencies or their confusions and biases.

Third, only the judgments of the two groups of children, unlike the adults and college students, were influenced by the presence of evidentially irrelevant instances of missing data. The effect of missing data was contrary to the prediction that children would "read into" the missing data further instances that could be explained by their (prior or newly formed) beliefs about the data. Rather, the influence of missing data on children was to make them less certain of the difference in the causal status of variables covarying and not covarying with the outcome. It was not simply that children became confused by the presence of instances of missing data in the data set. Rather, the presence of such instances made children in the missing data condition modulate their certainty that variables covarying or not covarying with plant health were (respectively) causal or noncausal. The finding that evidentially irrelevant data made children more uncertain is contrary to other findings that children rely on irrelevant task features in order to resolve uncertainty in a task situation (Scholnick & Wing, 1988; Somerville, Hadkinson, & Greenberg, 1979). It seems that, unlike adults, children do not just ignore or dismiss irrelevancies (Amsel et al., 1996), although the precise effect of such data on children's reasoning depends on the task.

The findings of the study presented here support both continuity and discontinuity in the development of evidence evaluation skills. The two groups of children demonstrated both inferential competence (like the adults) and biases and confusions (unlike the adults) in how they used contingency data to evaluate causal hypotheses. The finding that children have the competence to evaluate hypotheses on the basis of contingency data despite demonstrating inferential biases and confusions is not uncommon in the literature (Klahr et al., 1993). Schauble (1990) found that fifth and sixth graders correctly revised many of their causal beliefs, despite the fact that 67% of their inferences regarding the evidence were invalid. Similarly, Karmiloff-Smith and Inhelder (1974) found not only inferential biases and unsystematic responses by children who failed to balance blocks at their geometric center (a weight was hidden in the block) but also an ability to use such repeated failures as evidence disconfirming the generality of their geometric-center theory of block balancing.

The results of the research presented here have identified ways in which children's evidence evaluation skills are both continuous and discontinuous with those of adults and scientists, a conclusion also reached by Klahr et al. (1993). Such a conclusion calls into question whether there can be an unambiguous answer to the question about children's status as experimental scientists. As it stands now, the debate about whether or not children are like experimental scientists depends on the kind of evidence evaluation task used to assess children and the assumptions made about the appropriate model of scientists' evidence evaluation. Perhaps a better, more productive, way of thinking about the status of children as experimental scientists is not to assume that evidence evaluation skills are historically fixed in scientists.

We propose that just as there are parallels between ontogenesis and the history of science in the content of theories that children form and revise (cf. McCloskey, 1983; Piaget & Garcia, 1989), there may also be parallels at the level of the process of scientific experimentation in general and evidence evaluation in particular. Hypothetico-deduction, or the 'method of hypothesis,' was not always accepted as appropriate scientific practice in the history of science (Laudan, 1981). In such a method, hypotheses about the world are first proposed and data that would confirm or disconfirm them are deduced from the hypotheses. Only then are observations made of data from the world (Kyburg, 1970). The preferred methods of scientific experimentation in the 17th and 18th centuries were inductive or abductive ones (Laudan, 1981). In these methods, laws or explanations are inferred only after observations of the world are made, rather than being proposed hypothetically prior to making any observation (Kyburg, 1970). The inferred law or explanation is justified on the basis of being the best account of all the observations made (i.e., an inference to the best explanation).

The factors associated with the children's but not adults' evidence evaluation performance—belief-based justifications, prior beliefs, and missing data—are consistent with the notion that they are poor hypothetico-deductivists, but reasonably competent inductivists or abductivists. Children employing an inductive or abductive method may in fact be validly justifying their judgments when they refer to their beliefs—a justification based on children's explanations being their best account of the available information. Similarly, children employing an inductive or abductive method may be influenced by prior beliefs when evaluating information because such beliefs direct which laws or explanations are proposed to account for the information. Finally, the influence of missing data on children's evidence evaluation performance may also reflect their employing an inductive or abductive method. In such a method, laws or explanations are proposed to account for all the data. However, children's explanations could only be applied to a subset of the data in the data sets presented in the missing data conditions (i.e., four of the six instances) as opposed to all the instances (i.e., all four instances) in the data sets presented in control condition. Thus, children in the missing data conditions may have been less certain of the causal status of variables covarying and not covarying with plant health than children in the control condition because of the proportion of instances in the data sets that children's explanations could account for in the conditions.

It remains to be seen whether or not children revise the methodological principles and procedures that guide their everyday inquiries, and, if so, whether children and scientists reject old and accept new methods for the same reason. Answers to such questions will require more careful analyses of children's use of such methods of scientific experimentation as induction, abduction, falsification, and hypothetico-deduction (and all their variants). Identifying such developmental differences in children's evidence evaluation processes may prove to be more productive than separate examinations of children's ability to reason according to particular models of scientific reasoning.

REFERENCES

Acredolo, C., & O'Conner, J. (1991). On the difficulty of detecting cognitive uncertainty. *Human Development. 34,* 204–223.

Amsel, E., Goodman, G., & Savoie, D., & Clark, M. (1996). The development of reasoning about causal and noncausal influences on levers. *Child Development, 67,* 1624–1646.

Brewer, W., & Samarapungavan, A. (1991). Children's theories vs. scientific theories: Differences in reasoning or differences in knowledge. In R. Hoffman & D. Palermo (Eds.), *Cognition and the symbolic processes: Applied and ecological perspectives* (pp. 209–232). Hillsdale, NJ: Erlbaum.

Bullock, M. (1991, April). *Scientific reasoning in elementary school: Developmental and individual differences.* Paper presented at SRCD, Seattle, WA.

Bullock, M., Ziegler, A., & Martin, S. (1992). Scientific thinking. In F.E.W. Weinert & W. Schneider (Eds.), *LOGIC Report 9: Assessment procedures and results of wave 6.* Munich: Max Planck Institute for Psychological Research.

Carey, S. (1985a). Are children fundamentally different kinds of thinkers and learners than adults? In S.F. Chipman, J.W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Research and open questions* (Vol. 2, pp. 485–517). Hillsdale, NJ: Erlbaum.

Carey, S. (1985b). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Gelman, S., & Markman, E. (1986). Categories and induction in young children. *Cognition, 23,* 183–209.

Gopnik, A., & Astington, J. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance/reality distinction. *Child Development, 59,* 26–37.

Gopnik, A., & Wellman, H. (1992). Why the child's theory of mind really *is* a thory. *Mind and Language, 7,* 145–171.

Gruber, H. (1973). Courage and cognitive growth in children and scientists. In M. Schwebel & J. Ralph (Eds.), *Piaget in the classroom* (pp. 73–105). New York: Basic Books.

Hemple, C. (1961). *Philosophy of natural sciences.* Englewood Cliffs, NJ: Prentice-Hall.

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition, 3,* 195–212.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25,* 111–146.

Koslowki, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough. The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development, 60,* 1316–1327.

Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science, 6,* 133–139.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills.* San Diego: Academic.

Kuhn, D., Schauble, L., & Mila-Garcia, M. (1992). Cross domain development of scientific reasoning. *Cognition and Instruction, 4,* 285–327.

Kyburg, H. (1970). *Probability and inductive logic.* London: Collier-Macmillan.

Laudan, L. (1981). *Science and hypothesis: Historical essays on scientific methodology.* Dordrecht, Holland: Reidel.

Levine, G. (1991). *A guide to SPSS for analysis of variance.* Hillsdale, NJ: Erlbaum.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A.L. Stevens (Eds.), *Mental models* (pp. 299–322). Hillsdale, NJ: Erlbaum.

Piaget, J., & Garcia, R. (1989). *Psychogenesis and the history of science.* New York: Columbia University Press.

Ross, L. (1981). The "intuitive scientist" formulation and its developmental implications. In J. Flavell & L. Ross (Eds.), *Social cognitive development: Frontiers and possible futures* (pp. 1–42). New York, NY: Cambridge University Press.

Rosser, R. (1994). *Cognitive development: Psychological and biological perspectives.* Boston: Allyn & Bacon.

Ruffman, T., Perner, J., Olson, D., & Doherty, M. (1993). Reflecting on scientific thinking: Children's understanding of the hypothesis-evidence relation. *Child Development, 64,* 1617–1636.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49,* 31–57.

Schauble, L., & Glaser, R. (1990). Scientific thinking in children and adults. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills: Contributions to human development* (Vol. 21, pp. 9–27). Basel: Karger.

Scholnick, E.K., & Wing, C.S. (1988). Knowing when you don't know: Developmental and situational considerations. *Developmental Psychology, 24,* 190–196.

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight, and density. *Cognition, 21,* 177–237.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62,* 753–766.

Somerville, S., Hadkinson, B.A., & Greenberg, C. (1979). Two levels of inferential behavior in young children. *Child Development, 50,* 119–131.

Wellman, H., & Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43,* 337–375.