Developmental Psychology
1997, Vol. 33, No. 1, 70–78

0012-1649/97/$3.00

Children's Differential Performance on Deductive and Inductive Syllogisms

Kathleen M. Galotti, Lloyd K. Komatsu, and Sara Voelz
Carleton College

At what age and in what ways do children distinguish deductive and inductive problems? In Experiment 1, students from kindergarten and from Grades 2, 4, and 6 were presented with deductive or inductive inference problems and asked to draw an inference and rate their confidence. By 4th grade, confidence ratings for deductive problems were higher than those for inductive problems, and responses were faster for deductive than for inductive items. In Experiment 2, students from Grades 2, 3, 4, and 5 responded to the same problems used in Experiment 1 but were asked to provide explanations for their responses. Again, confidence was higher with deductive than with inductive problems, and latency to respond was faster for deductive than for inductive items. Further, explanations differed as a function of the type of problem. These findings help fill in gaps in the emerging picture of the development of children's reasoning skills.

Recent research on children's reasoning abilities and competence has presented a fascinating but complex picture. On the one hand, accumulating evidence seems to suggest that, at least in some circumstances, even preschoolers are capable of drawing deductively valid conclusions (Dias & Harris, 1988; Hawkins, Pea, Glick, & Scribner, 1984). On the other hand, some work suggests that sophisticated mastery of deductive inference, as well as full, explicit understanding of the idea of logical necessity, does not emerge before early adolescence (Galotti & Komatsu, 1989; Kuhn, 1989; Markovits, 1993; Moshman, 1990; Moshman & Franks, 1986; Moshman & Timmons, 1982; Overton, Ward, Black, Noveck, & O'Brien, 1987; Ward & Overton, 1990; but see Girotto, Blaye, & Farioli, 1989, and Light, Blaye, Gilly, & Girotto, 1989, for opposing views).

To determine whether a child of a given age can or cannot reason (logically, deductively, inductively, etc.), one must first provide some definition of what it means to reason. In our view, reasoning of any sort involves, at a minimum, the ability to draw an inference in accordance with a system of rules, if any, that govern that type of inference (Overton, 1990). Mature reasoning goes beyond this ability and includes the ability to reflect on and to explain the basis for the inference drawn (Moshman, 1990). This type of reasoning includes the recognition of different types of inferences and the construction and adoption of different criteria for different inferences.

In this research, we were specifically interested in the intermediate steps between being able to draw a deductive or inductive inference and some sort of recognition of the difference between them, a more reflective metareasoning ability. It is likely that this distinction begins implicitly in performance, before full, explicit understanding and articulation of the nature of the different kinds of inference.

A brief review of relevant terms may be in order. A conclusion is said to be *deductively valid* if it is true whenever the premises are true (Rips, 1990). Thus, a deductive inference is guaranteed to yield true conclusions whenever the original premises are true. In most everyday kinds of reasoning, of course, such inferences are rarely encountered (Galotti, 1989). More frequent are inductive inferences. A conclusion is said to have *inductive strength* if it is likely to be true when the premises are true. In other words, inductive inferences produce conclusions that vary in the likelihood of their being true. The greater the inductive strength, the more likely it is that the conclusion is true, given the truth of the premises (Rips, 1990).

Inductive inferences typically demand that the reasoner consult her or his practical world knowledge to assess the likelihood of a conclusion being true. Deductive inferences, in contrast, often force the reasoner to ignore practical world knowledge and focus analytically upon the premises supplied. Presumably, this analytical mode of reasoning, calling upon an ability to abstract from pragmatic experience, is more difficult for younger children (Bucci, 1978). Indeed, within a Piagetian framework, the ability to draw deductive inferences is seen as requiring (at a minimum) concrete operations (Gellatly, 1987). Thus, a study by Hawkins et al. (1984) produced some very surprising results.

Hawkins et al. (1984) gave 4- and 5-year-old children a number of verbal syllogisms (e.g., "Pogs wear blue boots. Tom is a pog. Does Tom wear blue boots?") that required deductive inferences. Contrary to Piagetian predictions, Hawkins et al.

Kathleen M. Galotti, Lloyd K. Komatsu, and Sara Voelz, Department of Psychology, Carleton College.

Some of this research was presented at the 1995 Annual Meeting of the Psychonomic Society, Los Angeles, California.

We thank the principal, teachers, students, and parents at the Longfellow School, Greenvale School, Sibley School, St. Dominic's School, and Northfield Middle School, all in Northfield, Minnesota, and the Nerstrand School in Nerstrand, Minnesota. We thank Beth Lavin, Kate Ainsworth, and Karen Jacobs for assistance in data analysis. We thank David Moshman and Robert Siegler for comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Kathleen M. Galotti or Lloyd K. Komatsu, Department of Psychology, Carleton College, Northfield, Minnesota 55057.

found that their preschool children could answer many of the syllogisms correctly, and could provide appropriate justifications for their answers. The syllogisms used varied in content: Some contained premises that were expected to be congruent with the children's practical knowledge (e.g., "Bears have big teeth. Animals with big teeth can't read books. Can bears read books?"); other premises were expected to be incongruent with practical knowledge (e.g., "Glasses bounce when they fall. Everything that bounces is made of rubber. Are glasses made of rubber?"); and others, called *fantasy problems*, contained premises describing mythical creatures foreign to practical knowledge, as in the pogs example above.

Hawkins et al. (1984) found that children performed best on congruent problems (averaging about 94% correct responses), and worst on incongruent problems (averaging about 13% correct responses), with intermediate performance on fantasy problems (averaging 73% correct responses). Performance on the fantasy problems was taken as most indicative of reasoning ability because these problems necessarily prevented children from answering on the basis of preexisting empirical knowledge. It is interesting that the order in which children solved the syllogisms had large effects on performance: Those children who received the fantasy problems first showed higher overall levels of performance (and especially on the fantasy problems relative to problems with congruent or incongruent content) than all other groups of children. Hawkins et al. speculate that this order of presentation communicated to the children that empirical knowledge should not be used in this task. Consistent with this conclusion, Dias and Harris (1988, 1990) found that 4- and 6-year-olds can even make deductive inferences with incongruent syllogisms if they are clearly encouraged to answer in the spirit of play, but not necessarily if they are just verbally encouraged to pretend that the incongruent premises are true.

On the other hand, Markovits, Schleifer, and Fortier (1989) demonstrated that children ages 6 and 8 years have trouble distinguishing between syllogisms with related premises (e.g., "Every Ploque is tall. No tall thing is green. Are Ploques green?") and syllogisms with unrelated premises (e.g., "Risomes cannot sing. Zapp is a Touki. Can Zapp sing?"). Markovits et al. argue that this result implies that even if younger children can sometimes draw a deductive conclusion, they do not show a full understanding of the relationship between the premises and the conclusion before age 11 years or so (see related arguments by Freeman & Sepahzad, 1987).

Moshman and Franks (1986) presented additional evidence that even when children can draw deductively valid conclusions, they lack a full appreciation and understanding of the idea of logical necessity and validity. In three studies, students in fourth grade, seventh grade, and college were given a number of arguments that varied in the truth of the premises, the truth of the conclusions, and the validity of the arguments. To understand the concept of validity, Moshman and Franks assert, is to ". . . [differentiate] it from empirical truth and [generalize] across some range of content areas and argument types" (1986, p. 163). Children were asked to sort the arguments in as many ways as they could. Results showed that, even when specifically prompted to do so, fourth graders had great difficulty sorting on the basis of validity. Moshman and Franks argue that, before age 12 years or so, children have little explicit awareness of validity, especially under less than ideal conditions.

The procedures used by Hawkins et al. (1984) and Dias and Harris (1988, 1990) can be described as ones that facilitate maximally children's ability to draw inferences. In contrast, the tasks used by Moshman and Franks (1986) constitute a stringent test of children's *metalogical knowledge* (knowledge about the nature of logic; cf. Moshman, 1990; Moshman & Franks, 1986; Moshman & Timmons, 1982), specifically in this case, their understanding of the kinds of inferences allowable under different conditions. Thus, it is unsurprising that younger children showed success with the former, whereas older children still performed poorly on the latter. However, the results of these studies taken together pose an interesting question: How (and when) does knowledge about reasoning develop?

One aspect of knowledge about reasoning is the recognition of different types of inference. Presumably, metalogical knowledge includes some notion about the strength of different kinds of inferences: Deductively valid inferences are, by definition, more certain than inferences of whatever degree of inductive strength (Moshman, 1990; Murray, 1990). We examined behavioral recognition of this distinction in children of different ages. Our primary question was, do children who draw different kinds of inferences for deductive and inductive problems also indicate different amounts of confidence in those answers? Children who do show performance differences as a function of inductive versus deductive problems may not understand the inductive-deductive distinction explicitly, but at least are making a distinction in behavior. We believe this behavioral distinction is an implicit step on the road to full, mature, explicit understanding of the inductive-deductive distinction.

Experiment 1

Method

Participants. The sample consisted of 66 elementary school students with an approximately equal number of boys and girls taken from four age groups: 16 kindergartners (8 boys, 8 girls; mean age = 67.6 months, SD = 4.1), 17 second graders (9 boys, 8 girls; mean age = 92.3 months, SD = 6.8), 16 fourth graders (7 boys, 9 girls; mean age = 117.2 months, SD = 4.1), and 17 sixth graders (9 boys, 8 girls; mean age = 141 months, SD = 4.1). Data from one additional first-grade boy and one additional sixth-grade girl were excluded because of procedural error while running the interviews. Students were recruited from three elementary schools and a middle school in a small, college town in southeastern Minnesota and received parental permission to participate in the study. Students participated on a voluntary basis and received no monetary compensation.

Materials and procedure. Stimuli consisted of 16 traditional syllogisms. For each syllogism, we constructed a deductive and an inductive version that were matched for content but differed in the certainty with which an answer could be given. For example, the following was the deductive version of one syllogism: "All shakdees have three eyes. Myro is a shakdee. Does Myro have three eyes?" This question requires a deductive inference to be drawn, and the correct and certain answer to the question is "yes." The following was the inductive version of the same item: "Myro is a shakdee. Myro has three eyes. Do all shakdees have three eyes?" This question requires an inductive inference to be drawn and has no certainly correct answer (although it invites a "yes" response).

Of the 16 syllogisms, 4 were *affirmative* (i.e., the correct answer for

Table 1
Examples of Different Categories of Stimuli Used, Experiment 1

Syllogism version	
Deductive	Inductive
Affirmative and particular	
All poggops wear blue boots. Tombor is a poggop. Does Tombor wear blue boots?	Tombor is a poggop. Tombor wears blue boots. Do all poggops wear blue boots?
Affirmative and universal	
All daxlets are squishy. All squishy animals like to yell. Do all daxlets like to yell?	All squishy animals like to yell. All daxlets like to yell. Are all daxlets squishy?
Negative and particular	
No risomes play checkers. Zapp is a risome. Does Zapp play checkers?	Zapp is a risome. Zapp does not play checkers. Do any risomes play checkers?
Negative and universal	
All berbers wiggle. No wiggly animals wear hats. Do all berbers wear hats?	No wiggly animals wear hats. No berbers wear hats. Are all berbers wiggly?

deductive versions was "yes") and *particular* (i.e., a premise or question mentioned an individual); 4 were *affirmative* and *universal* (i.e., all premises and the question referred to sets, rather than to individuals); 4 were *negative* (i.e., the correct answer for deductive versions was "no") and *particular;* and 4 were *negative* and *universal.* Table 1 presents examples of these.

To heighten students' interest in the task, we created colorful drawings of each of the 16 syllogisms (each of which had a deductive and an inductive version), and presented the drawings along with the questions. Drawings were carefully constructed so as not to give away answers to any of the questions to be asked. Finally, to shorten the length of the task, we used only fantasy content in all of the items.

We arranged the stimuli into two sets of 16 questions; each set had only one version of each syllogism. The inductive version of a syllogism was always assigned to a different set than the deductive version of that syllogism. Both sets had an equal number of deductive and inductive questions (eight of each), an equal number of universal and particular syllogisms (eight of each), and an equal number of affirmative and negative syllogisms (eight of each). More specifically, each set of 16 questions contained 2 deductive-universal-affirmative syllogisms, 2 deductive-universal-negative syllogisms, 2 deductive-particular-affirmative syllogisms, 2 deductive-particular-negative syllogisms, 2 inductive-universal-affirmative syllogisms, 2 inductive-universal-negative syllogisms, 2 inductive-particular-affirmative syllogisms, and 2 inductive-particular-negative syllogisms. Each student received one set of 16 questions, thus no student heard the same set of premises more than once. The presentation of question sets was counterbalanced across age and gender.

To assess confidence, we constructed a graphic 5-point confidence scale, which remained in view throughout the interview. On the left end of the scale was a drawing of a child (of the same sex as the student) with hands raised in celebration (depicting high confidence); on the right end was a drawing of a child shrugging shoulders (depicting low confidence). Between the two pictures were five dots that represented varying degrees of confidence. The children were asked to indicate their confidence by pointing to whichever dot best represented their level of confidence in their answer. The experimenter (Sara Voelz) explained that the second figure looked confused because someone asked him a question that he had to guess the answer to. The experimenter explained that some questions were like that. For instance, if the child had to tell the experimenter what the experimenter's birthday was, he or she would have to guess. She then instructed the child that when the child had to guess the answer to questions, he or she should point to the circle closest to the shrugging figure. Other questions, the experimenter explained, were ones that one could know the answer to, such as "What's your birthday?" In those cases, the child should point to the circle closest to the smiling figure. Intermediate circles indicated when the child felt "a little bit" sure that they knew the answer. The experimenter then had the child show what circle they should point to if they knew an answer, guessed an answer, and "sort of knew" an answer. All of the children responded correctly to these instructions.

Children were then told that they were going to hear some stories about make-believe creatures that came from another planet. They were told that they were also going to see pictures of the creatures and then answer questions about them. During the instructions, and again throughout the interview, children were reminded that they would not be able to answer the questions just by looking at the pictures but, instead, would have to listen very carefully to the stories and questions. Children were also told that, after answering the questions, they would be asked how sure they felt about their answers to the questions. Children were instructed to point to the dot on the confidence scale that best indicated how sure they were of their answer. Children were discouraged from answering "maybe" to any question. If a child did answer "maybe," she or he was encouraged to guess what the answer was. Such children always chose the figure on the confidence scale who looked perplexed, indicating that they were able to use the confidence scale correctly. The experimenter made sure the children understood the instructions before proceeding to read the questionnaires to them. Children's answers to the questions and their confidence ratings were recorded, and the interviews were also tape-recorded. Each student was interviewed individually at his or her school in a single, 10- to 15-min interview.

Table 2
Mean Responses to Questions as a Function of Age–Grade, Inference, and Status, Experiment 1

	Inference type and status			
	Deductive		Inductive	
Grade	Affirmative	Negative	Affirmative	Negative
K	1.18	0.88	0.97	0.81
2	1.27_a	0.56_b	1.00	0.74
4	1.59_a	$0.31_{b,c}$	0.84_c	$0.97_{c,d}$
6	1.59_a	0.29_b	$0.76_{b,c}$	1.00_c

Note. Responses were scored 1 if the student answered "yes." Maximum possible score = 2. Within rows, numbers with different subscripts differ significantly ($p < .01$) by a post hoc Tukey's honestly significant difference test. K = kindergarten.

Results and Discussion

Responses, confidence ratings in those responses, and time to respond were analyzed separately.

Responses. Responses were coded as 1 for *yes*. Any uncertain response (e.g., "maybe," "not sure," or "probably") was excluded from the analysis. The mean proportion of uncertain responses was .02, .02, .06, and .02 for the kindergartners, second graders, fourth graders, and sixth graders, respectively.

Every child saw two problems of each combination of the three factors (type of inference—deductive vs. inductive; quantification of premises—universal or particular, and status of premises—affirmative or negative). Therefore, their total scores for each type of problem could range from 0 to 2, with a score of 1 indicating inconsistent responses to the two problems. Preliminary analysis indicated no differences in the patterns of responses for boys and girls, so the data were collapsed over gender in subsequent analysis. Proportions of "yes" responses were subjected to a 4 (age–grade: kindergarten, 2, 4, 6) × 2 (inference: deductive or inductive) × 2 (quantification of premises: universal or particular) × 2 (status of premises: affirmative or negative) mixed analysis of variance (ANOVA), with repeated measures on the last three variables. A corresponding ANOVA on the proportion of "no" responses yielded the inverse pattern of results.

The ANOVA revealed main effects of status, with more "yes" responses given to affirmative items ($M = 1.15$ for affirmative items, .69 for negative items), $F(1, 62) = 47.88$, $p < .001$, $MSE = 0.57$, as expected. There were no overall main effects for any of the other factors. However, several interactions emerged.

The first interaction was between status and inference. It showed that the difference in "yes" answers given to affirmative versus negative items was evident on deductive ($Ms = 1.41$ vs. 0.51) but not on inductive items ($Ms = 0.89$ vs. 0.88), $F(1, 62) = 77.52$, $p < .001$, $MSE = 0.33$). Post hoc Tukey honestly significant difference (HSD) tests indicated that the former pair of means differed reliably ($p < .01$), but the latter pair did not.

A three-way interaction among age–grade, status, and inference was also statistically significant, $F(3, 62) = 11.68$, $p < .001$, $MSE = 0.33$. Means for the interaction are presented in Table 2. Post hoc Tukey HSD tests indicated that kindergartners showed no significant differences in the proportion of "yes" answers they gave as a function of inference type or status. Second graders responded differentially between affirmative and negative items, but only with deductive versions. The two oldest groups showed an even more differentiated pattern of responding.

The inductive items had no strictly "correct" answers, but the deductive items did, and it was always "yes" for affirmative items and "no" for negative items. The pattern of results indicates that older children are more sensitive than younger children to these distinctions, but that the sensitivity begins by at least second grade.

There was also a significant interaction between inference and quantification, $F(1, 62) = 17.01$, $p < .001$, $MSE = 0.48$. Universally quantified items had means of 0.86 versus 1.04 for deductive versus inductive versions, whereas particularly quantified items had corresponding means of 1.05 versus 0.74. Post hoc Tukey HSD tests indicated that the difference for the latter, but not former, pair of means was statistically reliable ($p < .01$).

Finally, a significant three-way interaction emerged between age–grade, status, and quantification, $F(3, 62) = 4.04$, $p < .05$, $MSE = 0.34$. Means for this interaction are shown in Table 3. Once again, post hoc Tukey tests indicated that kindergartners showed no significant differences in the proportion of "yes" answers they gave as a function of status (affirmative vs. negative) or quantification. Second graders responded differentially between affirmative and negative items, but only with particular versions. The two oldest groups, in contrast, showed a somewhat more differentiated pattern of responding.

Confidence ratings. Confidence ratings were coded 1–5, with numbers corresponding to the five dots on the scale shown to children. Higher numbers designated greater confidence. Confidence ratings were subjected to a mixed ANOVA using the same four variables as above. The analysis yielded a main effect for inference, $F(1, 62) = 11.29$, $p < .001$, $MSE = 4.46$. Mean confidence rating on deductive inferences was 4.17, whereas the corresponding mean for confidence on inductive items was 3.87. There was also a significant two-way interaction between inference and quantification, $F(1, 62) = 13.66$, $p < .001$, $MSE = 2.78$. Mean confidence rating was 4.05 for deductive-universal items, 4.01 for inductive-universal items, 4.30 for deductive-particular items, and 3.72 for inductive-particular items. Post

Table 3
Mean Responses to Questions as a Function of Age–Grade, Quantification, and Status, Experiment 1

	Quantification and status			
	Universal		Particular	
Grade	Affirmative	Negative	Affirmative	Negative
K	1.03	0.91	1.13	0.78
2	1.06	0.80	1.21_a	0.50_b
4	1.38_a	0.66_b	1.06	0.63_b
6	1.27_a	0.53_b	1.09_a	0.77

Note. Responses were scored 1 if the student answered *yes*. Maximum possible score = 2. Within rows, numbers with different subscripts differ significantly ($p < .01$) by a post hoc Tukey's honestly significant difference test. K = kindergarten.

hoc Tukey HSD tests indicated that the difference between deductive and inductive items was significant ($p < .01$) only for particular items.

Finally, there was a four-way interaction among age–grade, inference, status, and quantification, $F(3, 62) = 3.62, p < .05$, $MSE = 2.00$. No clear pattern of means emerged from this interaction.

Reaction times. Response latencies were measured in seconds and subjected to a natural log transformation because of the well-known problems of skewness with reaction time (RT) distributions. All analyses were performed on logs, but for ease of exposition, the antilogs of all means will be reported. Log reaction times were again subjected to a mixed ANOVA, with the same four variables as before.

The analysis yielded a main effect for inference, $F(1, 60) = 19.76, p < .001$, $MSE = 0.37$. Students responded more quickly to deductive than inductive items ($M = 4.57$ vs. 5.81 s). There was also a main effect for status, $F(1, 60) = 5.20, p < .05$, $MSE = 0.22$, with students responding more quickly to affirmative ($M = 4.91$ s) than to negative items ($M = 5.40$ s). No other effects or interactions were significant.

Analyses of individual children's responses and confidence. Our last analyses of these data focused on the pattern of responses of individual children. If a child really understood deduction, we reasoned, she or he should answer all affirmative-deductive items as "yes" and all negative-deductive items as "no." Using this stringent criterion, we found that 0 out of 16 kindergartners, 3 out of 17 second graders, 5 out of 16 fourth graders, and 5 out of 17 sixth graders showed clear evidence of understanding deduction.

If children really make a deductive-inductive distinction, we reasoned further, then there should be a less clear pattern of responses on the eight inductive items. Examining these patterns for deductive answers (i.e., clear "yes" responses to affirmative items, clear "no" responses to negative items), we found that no child showed this pattern on inductive items. We concluded that analysis of individual children's responses showed the same pattern of results as did the analysis of the group data, albeit more weakly: At least by fourth grade, children respond differently to deductive and inductive inferences.

A similar analysis on confidence ratings was also performed. Using the criterion of giving the highest possible confidence rating (5) to all deductive items, we found that 2 out of 16 kindergartners, 4 out of 17 second graders, 4 out of 16 fourth graders, and 2 out of 17 sixth graders showed this pattern. This level of maximum confidence would be misplaced for inductive items. We found that 2 out of 16 kindergartners, 3 out of 17 second graders, 0 out of 16 fourth graders, and 1 out of 17 sixth graders gave the maximum confidence ratings to all inductive items. Thus, analyses of individuals' confidence ratings were not as consistent with the group analyses: Children (at all ages studied) did not show as clear a confidence distinction between the two items as is warranted. In particular, children appeared to be underconfident on the deductive items.

Experiment 2

We did not ask the children in Experiment 1 to explain how they came to their answers. In Experiment 2, we followed the same procedures as described above except that, after children responded to the question, we asked them to say why they arrived at that response. Then, we again asked for a confidence rating.

Asking for explanations served two purposes. First, a child's ability to give different kinds of explanations for deductive and inductive inferences would provide further evidence for him or her distinguishing between the two. Second, we wondered what effect requiring explanations would have on children's responses and confidence. In particular, we speculated that being asked to provide explanations might make children more careful and thoughtful about their answers and their confidence ratings.

Because the kindergartners showed little understanding of the distinction, and second graders only partial understanding, we also focused the age range of children studied. Here, we included only children from Grades 2 through 5.

Method

Participants. The sample consisted of 64 elementary school students with an approximately equal number of boys and girls taken from four age groups: 16 second graders (9 boys, 7 girls; mean age = 97.1 months, $SD = 7.6$), 16 third graders (8 boys, 8 girls; mean age = 110.4 months, $SD = 5.7$), 16 fourth graders (7 boys, 9 girls; mean age = 123.3 months, $SD = 5.05$), and 16 fifth graders (9 boys, 7 girls; mean age = 137.3 months, $SD = 5.2$). No students who participated in Experiment 1 were included in this study. Data from one additional third grader, one additional fourth grader, and two fifth graders were excluded because of equipment malfunction or procedural error. Students were recruited from three elementary schools and received parental permission to participate in the study. Students participated on a voluntary basis and received no monetary compensation.

Materials and procedure. We used the materials and procedures from Experiment 1. The only difference was that, after responding to the question, children were asked to say why they answered that way. Confidence ratings were collected after these explanations.

Results and Discussion

Once again, responses, confidence in those responses, and response latencies, were analyzed separately. We also coded the explanations (described below) and analyzed those separately.

Responses. Responses were coded as 1 for "yes" responses. Any other uncertain response (e.g., "maybe," "not sure," or "probably") was excluded from the analysis. The mean proportion of uncertain responses was .05, .03, .02, and .02 for the second graders, third graders, fourth graders, and fifth graders, respectively.

Every child saw two problems of each combination of the three factors (type of inference—deductive vs. inductive; quantification of premises—universal or particular; and status of premises—affirmative or negative). Therefore, their total scores for each type of problem could range from 0 to 2, with a score of 1 indicating inconsistent responses to the two problems. Preliminary analysis indicated no differences in the patterns of responses for boys and girls, so the data were collapsed over gender in subsequent analysis. Proportion of "yes" responses were subjected to a 4 (age–grade: 2, 3, 4, 5) × 2 (inference: deductive or inductive) × 2 (quantification of premises: universal or particular) × 2 (status of premises: affirmative or negative) mixed ANOVA, with repeated measures on the last three

Table 4
Mean Responses to Questions as a Function of Inference, Quantification, and Status, Experiment 2

	Quantification and status			
	Universal		Particular	
Inference	Affirmative	Negative	Affirmative	Negative
Deductive	1.54_a	0.25_b	1.66_a	0.16_b
Inductive	1.16_a	0.66_b	0.75_b	0.48_b

Note. Responses were scored 1 if the student answered *yes*. Maximum possible score = 2. Within rows, numbers with different subscripts differ significantly ($p < .01$) by a post hoc Tukey's honestly significant difference.

factors. A corresponding ANOVA on the proportion of "no" responses yielded the inverse pattern of results.

The ANOVA revealed main effects of status, with more "yes" responses given to affirmative items ($M = 1.28$ for affirmative items, .39 for negative items), $F(1, 60) = 164.44$, $p < .001$, $MSE = 0.62$, as expected, and replicating an effect of Experiment 1.

Two other main effects emerged that had not occurred in Experiment 1. The first was a main effect for inference, with significantly more "yes" responses given to deductive ($M = .93$) than to inductive questions, $M = .78$; $F(1, 60) = 5.58$, $p < .05$, $MSE = 0.45$. Second, more "yes" responses were given to universal ($M = .90$) than to particular items ($M = .76$), $F(1, 60) = 6.72$, $p < .05$, $MSE = 0.38$. Several interactions were statistically significant as well.

The first interaction was between status and inference. It showed that the difference in "yes" answers given to affirmative versus negative items was especially evident on deductive (*M*s = 1.50 vs. 0.20) relative to inductive items (*M*s = 0.95 vs. 0.57), $F(1, 60) = 73.31$, $p < .001$, $MSE = 0.45$). Post hoc Tukey HSD tests indicated that both pairs of means differed reliably ($p < .01$). This effect again replicated Experiment 1.

No interactions with age/grade as a variable emerged in this analysis (in contrast with the two that did emerge in Experiment 1). One plausible explanation is that the age range studied in Experiment 2 was more constrained.

The significant interaction between inference and quantification found in Experiment 1 was also replicated, $F(1, 60) = 7.58$, $p < .05$, $MSE = 0.37$. Universally quantified items had means of 0.90 versus 0.91 for deductive versus inductive versions, whereas particularly quantified items had corresponding means of 0.91 versus 0.62. Post hoc Tukey HSD tests indicated that the difference for the latter, but not former, pair of means was statistically reliable ($p < .01$).

Finally, a significant three-way interaction emerged in Experiment 2 that did not occur in Experiment 1 among inference, status, and quantification, $F(1, 60) = 4.82$, $p < .05$, $MSE = 0.32$. Means for this interaction are displayed in Table 4, which indicates a more pronounced difference between affirmative and negative scores for deductive items.

Confidence ratings. Confidence ratings were again coded 1–5, with higher numbers designating greater confidence. Confidence ratings were subjected to a mixed ANOVA using the same four variables as above.

The analysis yielded a main effect for inference, $F(1, 60) = 42.78$, $p < .001$, $MSE = 3.06$. Mean confidence on deductive inferences was 4.44, whereas the corresponding mean for confidence on inductive items was 3.93. There was no significant interaction between age–grade and inference, suggesting that the effect (greater confidence on deductive than inductive items) held for all age groups. Again, this pattern replicates Experiment 1.

There was also a marginally significant two-way interaction between inference and quantification, $F(1, 60) = 3.98$, $p < .06$, $MSE = 1.83$, another effect that almost replicates one for Experiment 1. Mean confidence was 4.32 for deductive-universal items, 3.94 for inductive-universal items, 4.54 for deductive-particular items, and 3.92 for inductive-particular items. Post hoc Tukey HSD tests indicated that the difference between deductive and inductive items was significant ($p < .01$) only for particular items.

The four-way interaction among age–grade, inference, status, and quantification was not significant in Experiment 2 (as it was in Experiment 1). However, a two-way interaction between age/grade and status did emerge, $F(1, 60) = 2.75$, $p < .05$, $MSE = 2.20$. Mean confidence for affirmative items was 4.00, 5.00, 4.09, and 4.21, for second, third, fourth, and fifth graders, respectively, whereas the corresponding figures for negative items were 4.23, 4.34, 4.14, and 3.96. No clearly interpretable pattern emerged from this interaction.

Reaction times. Response latencies were measured in seconds and subjected to a natural log transformation. All analyses were performed on logs, but for ease of exposition, the antilogs of all means are reported. Log reaction times were subjected to a mixed ANOVA using the same four variables as above.

The analysis yielded a main effect for inference, $F(1, 63) = 54.86$, $p < .001$, $MSE = 0.41$. As in Experiment 1, students responded more quickly to deductive than inductive items ($M = 4.17$ vs. 6.36 s). Unlike Experiment 1, there was no main effect for status. However, there was a significant two-way interaction between inference and quantification, $F(1, 63) = 7.59$, $p < .01$, $MSE = 0.27$. For inductive items, mean antilog reaction time was 6.11 versus 6.55 seconds for universally versus particularly quantified items; for deductive items, the figures are 4.53 versus 3.82. Post hoc Tukey HSD tests indicated that the difference between differently quantified items was significant ($p < .01$) only for deductive items. No other effects or interactions were significant in this analysis.

Verbal explanations. Explanations given by the participants were transcribed from the tape recording. Kathleen M. Galotti and two other research assistants coded these explanations, blind to the child's age or gender and to whether the item presented had been deductive or inductive. The categories used to code the explanations are presented in Table 5, along with examples used by coders, and the interrater reliabilities, computed with coefficient alpha. This coding taxonomy was adapted from the one used by Hawkins et al. (1984).

The two most frequently used categories were A (refers to premises or story) and C (refers to conjecture or uncertainty). The former code was used for approximately 54% of the explanations; the latter for 21%. Category E was never used, and all

Table 5
Coding Categories for Verbal Explanations, Experiment 2

Category	Interrater reliability[a]	Verbal explanation
A (premises, story)	.92	"Because you just told me that. The story said that. The story said all ____ are ____, and ____ is a ____, so ____ must be a ____." The child refers to the given information in the problem, mentioning both premises to derive a definite conclusion.
B (real world knowledge)	.66	"Because all animals have ____. Because animals don't play tennis, only people do." The child refers to her or his knowledge of the world, independent of the premises.
C (conjecture or uncertainty)	.79	"You only told me about one ____, but the others could be different. I don't know about all of the ____. Maybe all the ____ are alike but maybe not." The child refers to why he or she can only guess at something but cannot be certain.
D (picture)	.90	"It looks like he has ____. Because he looks blue in the picture." The child refers to the picture that was shown to her or him during the time the question was asked.
E (authority)[b]	—	"Because my daddy told me that." The child makes reference to another authority (e.g., parent, teacher, hero, friend).
F (no justification)	.90	"Just because." The child makes no attept to offer an explanation.
G (other or irrelevant justification)	.61	The child refers to irrelevant information or gives an explanation that does not make sense or does not fall into one of the above categories.

[a] Computed over four raters with coefficient alpha. [b] This category was never used by any of the coders.

others were used infrequently. Thus, we present analyses only for Categories A and C.

As with the other dependent measures, we conducted a mixed ANOVA using the same four variables as above, first on the number of explanations coded with Category A (refers to premises).

The analysis yielded a main effect for inference, $F(1, 60) = 82.43$, $p < .001$, $MSE = 0.77$. Mean number of explanations coded with this category (maximum 2, because there were two problems of each type) was 1.43 for deductive items, and 0.72 for inductive items. There was also a main effect for quantification, $F(1, 60) = 13.30$, $p < .001$, $MSE = 0.46$. Mean number of explanations coded with this category (maximum 2) was 1.18 on universally quantified items and 0.96 on particularly quantified items.

There were also two significant two-way interactions. The first was between inference and quantification, $F(1, 60) = 34.45$, $p < .001$, $MSE = 0.42$. Mean number of explanations coded with this category (maximum 2) for deductive items was 1.37 for universally quantified items, and 1.48 for the particularly quantified items. The corresponding figures for the inductive items are 1.00 versus 0.45. Post hoc Tukey HSD tests indicated that the difference between universal and particular items was significant ($p < .01$) only for inductive items.

Similarly, there was a significant interaction between inference and status, $F(1, 60) = 4.60$, $p < .05$, $MSE = 0.17$. Mean usage of this category (maximum 2) for deductive items was 1.42 for affirmative items, and 1.43 for the negative items. The corresponding figures for the inductive items are 0.80 versus 0.65. Post hoc Tukey HSD tests indicated that the difference between affirmative and negative items was significant ($p < .01$) only for inductive items.

Finally, a significant three-way interaction emerged among inference, status, and quantification, $F(1, 60) = 10.97$, $p < .01$, $MSE = 0.21$. Means for this interaction are presented in the top half of Table 6. Post hoc Tukey HSD tests indicated that Category A explanations were given more often ($p < .05$) for all deductive problems than any kind of inductive problems. Affirmative universal inductive items were given Category A explanations more often than negative universal inductive problems, which in turn received more such explanations than did any items with particular quantification.

A similar ANOVA was run on the number of explanations

Table 6
Mean Number of Explanations in Category A or C as a Function of Inference, Quantification, and Status, Experiment 2

	Quantification and status			
	Universal		Particular	
Inference	Affirmative	Negative	Affirmative	Negative
	Category A			
Deductive	1.31	1.42	1.53	1.44
Inductive	1.16_a	0.84_b	0.44_c	0.45_c
	Category C			
Deductive	0.25	0.19	0.11	0.11
Inductive	0.36	0.36	1.05	0.95

Note. Maximum possible score = 2. Categories A and C are mutually exclusive. Within the first two rows, numbers with different subscripts differ significantly ($p < .01$) by a post hoc Tukey's honestly significant difference test. Tukey tests were not run on the Category C measures, as the 3-way interaction was not statistically reliable. Category A refers to premises, story. Category B refers to conjecture or uncertainty (see Table 5).

coded with Category C (refers to conjecture or uncertainty). It yielded a main effect for inference, $F(1, 60) = 65.83$, $p < .001$, $MSE = 0.52$. Mean usage for this category (maximum 2) was 0.16 on deductive items and 0.68 on inductive items. There was also a main effect for quantification, $F(1, 60) = 33.19$, $p < .001$, $MSE = 0.27$). Mean number of explanations coded with this category was 0.29 for universally quantified items, and 0.55 for particularly quantified items. Finally, there was a significant two-way interaction between inference and quantification, $F(1, 60) = 56.70$, $p < .001$, $MSE = 0.32$. Mean number of explanations coded with this category for deductive items was 0.22 for universally quantified items and 0.11 for the particularly quantified items. The corresponding figures for the inductive items are 0.36 versus 1.00. Post hoc Tukey HSD tests indicated, once again, that the difference between universal and particular items was significant ($p < .01$) only for inductive items. Mean number of explanations coded by Categories A and C are presented together for purposes of comparison in Table 6, broken down by the variables of inference, status, and quantification.

Taken together, these analyses indicate that children are sensitive to the deductive-inductive distinction, in that they are much more likely to explain their answers by referring to premises with deductive items. Conversely, children are much more likely to explain answers to inductive items by referring to reasons for uncertainty.

Analyses of individual children's responses and confidence. In our last analyses of these data, we focused on the pattern of responses of individual children. Once again, the responses replicated the patterns found in Experiment 1. To recap, if a child had a firm grasp of the inductive-deductive distinction, we reasoned, she or he should answer all affirmative-deductive items as "yes" and all negative-deductive items as "no." Using this stringent criterion, we found that 7 out of 16 second graders, 8 out of 16 third graders, 5 out of 16 fourth graders, and 7 out of 16 fifth graders showed clear evidence of understanding deduction.

Similarly, if children really make a deductive-inductive distinction, we reasoned, then we should see a less clear pattern of responses on the eight inductive items. Examining these items for deductive answers (i.e., clear "yes" responses to affirmative items, clear "no" responses to negative items), we found that 0 out of 16 second graders, 1 out of 16 third graders, 0 out of 16 fourth graders, and 1 out of 16 fifth graders gave these responses. We conclude that the analysis of individual answers parallels the analysis of group data: Children at all ages studied are responding differently to deductive than to inductive items. Indeed, the second graders here seem to be performing much more differentially than did the second graders studied in Experiment 1.

A similar analysis on confidence ratings was performed. Using the criterion of giving the highest possible confidence rating ("5") to all deductive items, we found 4 out of 16 second graders, 5 out of 16 third graders, 2 out of 16 fourth graders, and 3 out of 16 fifth graders, respectively, showed this pattern. On inductive items, in which this level of maximum confidence would be inappropriate, we found that 1 out of 16 second graders, 3 out of 16 third graders, 4 out of 16 fourth graders, and 0 out of 16 fifth graders showed this pattern. Thus, this analysis on confidence ratings provided weaker support for the inductive-deductive distinction than did either the analysis of group data, or the analysis of individual answers. Once again, it seems clear that children do not have as high confidence in their deductive answers as is warranted.

General Discussion

Together, these studies suggest two things. First, young children can do more than draw deductive and inductive inferences. Even by second grade, they show the beginnings of implicit recognition that these two types of inferences are different and, as is appropriate, show more consistent answering, and higher confidence, in deductive inferences than in inductive inferences (although confidence in deductive answers is not as high as it should be).

What shows development across the ages studied, then, is a refinement of sensitivity to different aspects of reasoning tasks. In Experiment 1, kindergartners show little difference in the proportion of "yes" responses they provide as a function of inference type or status (affirmative vs. negative), although their pattern does follow the trend of older students. Fourth and sixth graders show clear sensitivity in their responses to inference type as well as to status, and second graders show an intermediate pattern of differentiation. Experiment 2 replicated the findings of Experiment 1 with children in second through fifth grade. Children in the second grade or older made a performance distinction between types of inference, expressing overall greater confidence in responding to deductive inference items. The RT measure confirmed the distinction shown with other measures: Students take less time to answer deductive items.

Moreover, the types of explanations children gave to inductive versus deductive items differed in a predictable way and again strongly supported our point: That children throughout most of the elementary grades make a behavioral distinction between these two types of inference. In contrast, kindergartners showed little understanding of the distinction (they were not asked to provide explanations, but their pattern of responses, response latencies, and confidence ratings did not show the distinction). Some development of understanding of the distinction, then, seems to be occurring during the early school years.

Although Experiment 2 replicated almost all of the findings of Experiment 1, an interesting difference occurred between the two experiments. The inductive-deductive distinction appeared to be much clearer in the second experiment. It may be simply that the kindergarteners did not participate. A more interesting possibility is that the requirement that children explain their responses sharpened their thinking. This issue must be left for further investigation.

Of all the dependent measures used (answers, confidence ratings, reaction times, and explanations), the confidence ratings in both studies showed the weakest discrimination between deductive and inductive items. In particular, children at all ages studied appeared underconfident on deductive items. Although deductively valid conclusions are necessarily true and therefore warrant the highest possible confidence, only some of even the fifth and sixth graders consistently displayed this level of confidence. It appears that this facet of understanding the inductive-deductive distinction develops relatively late.

In future work, researchers must explore a variety of issues. One is how well children of the ages in our samples can respond

to more complicated reasoning tasks. Will they make the deductive-inductive distinction (in responses, in confidence, and in latency to respond) when more demands are made of them in the course of actually drawing an inference, for example, with harder problems? A second issue is how salient the distinction is between deductive and inductive inferences. Are children explicitly aware of the difference and, if so, at what age? When and how are children able to assess inductive strength, making distinctions between inferences that are highly likely to be true versus inferences that are only slightly likely to be true?

References

Bucci, W. (1978). The interpretation of universal affirmative propositions. *Cognition, 6,* 55–77.

Dias, M. G., & Harris, P. L. (1988). The effect of make-believe play on deductive reasoning. *British Journal of Developmental Psychology, 6,* 207–221.

Dias, M. G., & Harris, P. L. (1990). The influence of the imagination on reasoning by young children. *British Journal of Developmental Psychology, 8,* 305–318.

Freeman, N. H., & Sepahzad, M. (1987). Competence of young children who fail to make a correct deduction. *British Journal of Developmental Psychology, 5,* 275–286.

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin, 105,* 331–351.

Galotti, K. M., & Komatsu, L. K. (1989). Correlates of syllogistic reasoning skills in middle childhood and early adolescence. *Journal of Youth and Adolescence, 18,* 85–96.

Gellatly, A. R. H. (1987). Acquisition of a concept of logical necessity. *Human Development, 30,* 32–47.

Girotto, V., Blaye, A., & Farioli, F. (1989). A reason to reason: Pragmatic basis of children's search for counterexamples. *European Bulletin of Cognitive Psychology, 9,* 297–321.

Hawkins, J., Pea, R. D., Glick, J., & Scribner, S. (1984). "Merds that laugh don't like mushrooms": Evidence for deductive reasoning by preschoolers. *Developmental Psychology, 20,* 584–594.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96,* 674–689.

Light, P., Blaye, A., Gilly, M., & Girotto, V. (1989). Pragmatic schemas and logical reasoning in 6- to 8-year-old children. *Cognitive Development, 4,* 49–64.

Markovits, H. (1993). The development of conditional reasoning: A Piagetian reformulation of mental models theory. *Merrill-Palmer Quarterly, 39,* 131–158.

Markovits, H., Schleifer, M., & Fortier, L. (1989). Development of elementary deductive reasoning in young children. *Developmental Psychology, 25,* 787–793.

Moshman, D. (1990). The development of metalogical understanding. In W. F. Overton (Ed.), *Reasoning, necessity, and logic: Developmental perspectives* (pp. 205–226). Hillsdale, NJ: Erlbaum.

Moshman, D., & Franks, B. A. (1986). Development of the concept of inferential validity. *Child Development, 57,* 153–165.

Moshman, D., & Timmons, M. (1982). The construction of logical necessity. *Human Development, 25,* 309–323.

Murray, F. B. (1990). The conversion of truth into necessity. In W. F. Overton (Ed.), *Reasoning, necessity, and logic: Developmental perspectives* (pp. 183–204). Hillsdale, NJ: Erlbaum.

Overton, W. F. (1990). Constraints on logical reasoning development. In W. F. Overton (Ed.), *Reasoning, necessity, and logic: Developmental perspectives* (pp. 1–31). Hillsdale, NJ: Erlbaum.

Overton, W. F., Ward, S. L., Black, J., Noveck, I. A., & O'Brien, D. P. (1987). Form and content in the development of deductive reasoning. *Developmental Psychology, 23,* 22–30.

Rips, L. J. (1990). Reasoning. *Annual Review of Psychology, 41,* 321–353.

Ward, S. L., & Overton, W. F. (1990). Semantic familiarity, relevance, and the development of deductive reasoning. *Developmental Psychology, 26,* 488–493.

Received January 2, 1995
Revision received March 25, 1996
Accepted March 25, 1996 ■