February 1986

Development of the Concept of Inferential Validity

David Moshman
University of Nebraska - Lincoln, dmoshman1@unl.edu

Bridget A. Franks
University of Nebraska - Lincoln

Published in *Child Development*, *57*:1 (February 1986), pp. 153–165.

The authors are grateful to the numerous students who participated in these studies; to the many teachers and administrators—especially Sherry Kaup, Beth Briney, Del Emerson, and Frank Masek—who graciously facilitated the work in their schools; and to Jeffrey Bisanz, Edith Neimark, David O'Brien, Willis Overton, Robert Sternberg, and several reviewers for detailed and helpful comments on earlier drafts of the manuscript. Portions of this research were presented at meetings of the Jean Piaget Society in Philadelphia in 1981, 1982, 1983, and 1984.

Development of the Concept of Inferential Validity

David Moshman and Bridget A. Franks

University of Nebraska—Lincoln

Abstract

An argument is valid if its conclusion necessarily follows from its premises, regardless of whether the premises and conclusion are empirically true or false. This research tested the hypothesis that understanding validity of inference (including its differentiation from empirical truth) is a relatively late development. Students in Experiment 1 were asked to sort sets of deductive arguments. None of the fourth graders used validity as a basis for distinguishing arguments, while 45% of the seventh graders and 85% of the college students did so. Experiments 2 and 3 explored whether the dramatic age difference could be narrowed by *(a)* varying the types of arguments used, *(b)* explaining the concept of validity and instructing students to use it, and/or (c) providing feedback after each trial. Fourth-grade performance remained poor, while seventh-grade performance increased to nearly the level of the college students. It was concluded that the concept of validity typically develops between ages 10 and 12 but that application of that competence continues to increase over a much longer age span. Students not understanding validity commonly evaluated arguments on the basis of empirical truth of component propositions, though even fourth graders revealed an implicit awareness of logical form.

Research on deductive reasoning and its development has included work on *(a)* conclusions reached or preferred by individuals who are presented with various combinations of premises (e.g., Hawkins, Pea, Glick, & Scribner, 1984), (*b*) the way these premises and conclusions are mentally represented (e.g., Neimark & Chapman, 1975), (*c*) real-time mental processes involved in reaching or evaluating these conclusions (e.g., Braine, 1978), and *(d)* the more abstract cognitive structures underlying these processes at various levels of development (e.g., Moshman, 1977). Another line of investigation increasingly represented in the literature involves the development of metalogical knowledge—knowledge about the nature of logic. The major locus of such research has been on the development of concepts of logical necessity.

Moshman and Timmons (1982) proposed a three-stage model of the development of logical necessity. The model is based on a conception of development in which each new level of understanding is constructed via active coordination of and metacognitive reflection on earlier concepts, such that knowledge implicit in an earlier structure achieves explicit representation (Bickhard, 1978). The child in stage 1 may act in accord with logical norms (e.g., in seriating blocks or deducing the conclusion to a transitive argument) but does not grasp the concept of logical necessity. The stage 2 child, by contrast, distinguishes conclusions that are logically necessary from those that are merely empirically likely or conventionally accepted. Reasoning at this level, however, is always within the context of premises accepted as true or reasonable. The child does not focus on the abstract form of the argument as a whole. Thus, although stage 1 and stage 2 children can often distinguish valid arguments (in which the conclusion follows necessarily from the premises) from invalid arguments (in which this is not the case), the concept of a necessary, content-independent relation between premises and conclusion is only implicit in their thinking, rather than being an object of explicit awareness.

The stage 3 individual, by contrast, not only distinguishes valid from invalid arguments but can think

explicitly about the form of an argument, thus differentiating the necessity of the relation between premises and conclusion from the empirical truth or falsity of each proposition. An individual who has attained this level of metalogical understanding may be said to have a mature concept of *inferential validity.*

Available evidence is consistent with the first two stages of the model in that preschool children can make correct deductions from a wide variety of premises (e.g., Braine & Rumain, 1983; Hawkins et al., 1984) while the (stage 2) ability to distinguish conclusions *required* by given premises from conclusions merely made plausible by those premises develops during middle childhood (e.g., Bereiter, Hidi, & Dimitroff, 1979; Pieraut-Le Bonniec, 1980), probably beginning about age 6 (Somerville, Hadkinson, & Greenberg, 1979), Unfortunately, there is almost no research on the development of the stage 3 concept of inferential validity. Some indirect indications consistent with the three-stage model are provided by studies showing relatively late development of concepts of tautology and logical contradiction (Cummins, 1978; Osherson & Markman, 1975) and by Piagetian theory, which associates hypotheticodeductive reasoning (such as deducing a conclusion from premises known to be false) with the emergence of formal operations (Inhelder & Piaget, 1958),

The primary purpose of the present research was to look more directly at children's and adults' understanding of the concept of validity. It was hypothesized that understanding and use of this concept would be a relatively late development, rare in 9–10-year-olds and increasingly prevalent beginning in early adolescence.

Experiment 1 focused on spontaneous use of the concept of validity. This was studied via an adaptation of the "Rep Test," a procedure developed by personality theorist George Kelly (1955) to assess the constructs or categories spontaneously applied by an individual in construing his or her experience.

Experiment 1

Method

Subjects.—The participants were 20 fourth graders (13 boys and seven girls), ranging in age from 9-0 to 10-6 (mean = 9-8); 20 seventh graders (10 boys and 10 girls), ranging in age from 12-0 to 14-3 (mean = 13-0); and 41 undergraduates (31 females and 10 males), ranging in age from 18 to 37 years (mean = 21.3). The children were all volunteers from public schools in Lincoln, Nebraska, and the undergraduates were recruited from a course required of all education majors at the University of Nebraska–Lincoln.

All fourth and seventh graders were asked at the conclusion of their participation whether they had any previous experience that seemed related to and/or helped them with the tasks they had just done. None cited any formal training in logic. College students were asked whether they had taken a course in logic. The few that had done so were not included in the experiment. No students at any age gave any indication during the research (e.g., through use of technical terminology) that formal training in logic was responsible for their performance. If any of the participants had studied any logic (e.g., syllogistic forms), apparently the present tasks were too different from the content of their course work for them to see any connection.

Materials: seven arguments.—Seven arguments were constructed by systematically varying *(a)* truth of premises, *(b)* truth of conclusion, and *(c)* validity of argument form (see Table 1). To highlight the logical nature of the task, each argument was placed in an "If . . . then" format. Since the research was intended to focus on conceptions about the nature of logical arguments rather than facility with particular inference patterns, simple logical forms were used throughout. The intent was to maximize the likelihood that, if students understood the *idea* of distinguishing valid from invalid arguments, they would easily be able to determine which of the present arguments belonged in which category. Six of the arguments involved transitive inference, which Piaget associates with concrete operations (beginning about age 7) and researchers in the information processing tradition view as understood by children as young as 4 (see Breslow, 1981). For the seventh (and third valid) argument, a simple disjunctive inference was used to insure that the three valid arguments could not be distinguished from the others simply on the basis of identical logical form.

Materials: test booklets.—Test booklets (for the college students) were constructed using the seven arguments of Table 1. Each booklet presented a series of tasks intended to provide students with multiple opportunities to spontaneously distinguish arguments on the basis of validity. Page 1 of each booklet presented general instructions and a sample exercise involving three geometric figures. Each of the next three pages presented three arguments and asked the student to "find as many ways as you can that two of

Table 1
Seven Arguments in Experiment 1

Premises:	Both True		At Least One False	
Form of argument:	Valid	Invalid	Valid	Invalid
Conclusion:				
True	1	2	4	5
False	*	3	7	6

1. If elephants are bigger than dogs
   And dogs are bigger than mice
   Then elephants are bigger than mice

2. If adults are older than babies
   And children are older than babies
   Then adults are older than children

3. If dogs are bigger than mice
   And elephants are bigger than mice
   Then dogs are bigger than elephants

4. If dogs are bigger than elephants
   And elephants are bigger than mice
   Then dogs are bigger than mice

5. If babies are older than adults
   And babies are older than children
   Then adults are older than children

6. If mice are bigger than dogs
   And mice are bigger than elephants
   Then dogs are bigger than elephants

7. If elephants are either animals or plants
   And elephants arc not animals
   Then elephants are plants

* No such argument is possible since a valid argument with true premises cannot have a false conclusion.

the following are similar and the other is different" and to provide a written explanation for each sorting. One of these three pages presented arguments 1, 2, and 3; one presented 4, 5, and 6; and one presented 1, 2, and 7 (see Table 1). The particular groupings of arguments were designed to provide several distinct potential sortings on each page. Thus, for example, arguments 1, 2, and 3 can be sorted on the basis of validity (1 vs. 2 and 3), truth of conclusion (1 and 2 vs. 3), or content (1 and 3 vs. 2). Order of the three pages was systematically varied across test booklets.

Page 5 was identical with pages 2, 3, and 4, except that all seven arguments were presented and the instructions specified that each sorting should divide the seven arguments into two groups (again along any dimension that the student could think of and again with an explanation of the basis for each sorting). Finally, page 6 presented the same seven arguments with instructions to order them from "most logical" to "least logical" (with ties allowed) and to explain this ranking.

Procedure.—College students were tested in groups of about eight or 10. The experimenter handed out the test booklets, carefully went through the instructions and sample exercise, solicited questions, and remained present in case of difficulties.

The same arguments and the same sequence of tasks were used for the younger students. There were, however, some major procedural differences designed to avoid false negatives by making it as easy as possible for children to demonstrate any inclination to think about arguments in terms of their validity.

First, the arguments appeared on index cards that could be sorted and ranked physically. Second, students were individually interviewed by the first author and their responses were tape-recorded so that it was unnecessary for them to deal with a test booklet or express themselves in writing. Third, each fourth grader was asked to read the first few cards aloud to ascertain that reading was not a source of difficulty (it was not). Fourth, in sorting the set of all seven arguments, children were permitted, if they wished, to use more than two categories in a single sorting. Fifth, the interview procedure allowed for follow-up questions by the interviewer to ascertain whether validity might be involved in what initially appeared to be an insufficient explanation. Finally, on each set of three arguments, if a given child did not spontaneously produce all three possible ways of dividing them, the experimenter would demonstrate each missing division of the cards, including the valid versus invalid division, and ask whether that division made any sense and, if it did, why.

Scoring.—Students' sortings of each set of arguments were classified on the basis of which cards they placed together and their explanations. In the case of fourth and seventh graders, full credit for understanding validity (or any other basis for sorting) was given for post hoc explanation of sortings suggested by the interviewer. Seven mutually exclusive categories were used:

1. *Validity*.—Student separates valid from invalid argument(s), indicating that in the case of the valid argument(s) the conclusion must be true provided the premises were true (because of the form of the argument), while for invalid argument(s) one could not tell from the premises whether or not the conclusion were true (though one might be certain of its truth or falsity on the basis of other empirical knowledge). Formal explication of the distinction between validity and truth is not required, but the explanation must be sufficient to rule out sortings based on form, truth, content, or partial matching (see below).

2. *Form*.—Student notes similarity and/or difference in abstract form of argument but does not indicate that one form is superior to another in terms of the logicality of the connection between premises and conclusion.

3. *Truth*.—Student distinguishes arguments on the basis of whether the premises and/or conclusions are empirically true or false.

4. *Content*.—Student differentiates arguments on the basis of the content of the terms (e.g., animals vs. people) and/or relations (e.g., size vs. age).

5. *Partial matching*.—Arguments are matched on the basis of having the same or comparable words in a particular location (e.g., two arguments both use the word "dog" as the first term in the conclusion).

6. *Mixed*.—Student sorts the set of seven arguments into three or more categories based on an idiosyncratic combination of factors (e.g., true conclusions vs. animal content vs. human content).

7. *Miscellaneous*.—Explanation missing, not fitting above categories, insufficient to classify, or not in accord with student's own grouping of cards.

Reliability was assessed by having a second coder classify 50 randomly chosen written responses. Agreement with the original rater was 82% (with 98% agreement on the critical question of whether or not each sorting reflected an understanding of validity).

The major analysis of each student's ranking of the seven arguments consisted of determining whether she or he placed the three valid arguments (1, 4, and 7) as the three most logical (regardless of whether those were perceived as tied with each other or were placed in some order). Because there seemed to be no systematic basis for such rankings other than validity and since the probability of such a ranking being produced by chance is less than 3%, no explanation was required (though adequate validity explanations were usually provided).

Results

Validity.—Not a single fourth grader ever used the concept of validity in sorting or ranking the cards. Of the seventh graders, 35% sorted one or more of the trios on the basis of validity (with explanation), 20% sorted the set of all seven this way (with explanation), and 35% chose the three valid arguments as "most logical" in ranking the set of seven. Corresponding figures for college students were 78%, 46%, and 61%. Overall, the percentages of fourth graders, seventh graders, and college students respectively using the concept of validity at some point in the research (i.e., in at least one of the four sortings—with explanation—or in the ranking) were 0%, 45%, and 85%, $\chi^2(2, N = 81) = 40.4$, $p < .001$. Fourth-grade performance was significantly below that of seventh graders (binomial $p < .01$), which in turn was significantly below that of college students, $\chi^2(1, N = 61) = 10.9$, $p < .001$. Performance

of males and females did not differ significantly at any age. As individuals, the seventh graders spanned the entire range of possible performance, with many showing behavior indistinguishable from that of the fourth graders and others providing sortings, rankings, and explanations indistinguishable from those of the most sophisticated college students.

Other concepts used.—Table 2 shows the percentage of students at each grade using each of various bases for sorting. A series of χ^2 tests showed no age differences in sortings on the basis of content, truth, or partial matching ($p > .10$ in each case). There were significant age differences in (a) sortings on the basis of validity, for trios, $\chi^2(2, N = 81) = 34.6$, $p < .001$; for set of seven, $\chi^2(2, N = 81) = 15.1$, $p < .001$; (b) sortings on the basis of form, for trios, $\chi^2(2, N = 81) = 19.6$, $p < .001$; for set of seven, $\chi^2(2, N = 81) = 13.6$, $p < .01$; and (c) miscellaneous sortings, for trios, $\chi^2(2, N = 81) = 13.2$, $p < .01$; for set of seven, $\chi^2(2,N = 81) = 19.4$, $p < .001$.

Students in all three grades were remarkably similar in the nearly universal use of content as a basis for distinguishing arguments, the very common use of empirical truth, and the occasional use of partial matching. The major age difference is that sortings on the basis of form and validity were much less common in the seventh graders than in the college students and were entirely absent in the fourth graders. The relative absence of miscellaneous responses in the children confirms the success of the interview procedure in making it possible to assign vague or incomplete initial explanations to nonmiscellaneous categories on the basis of further questioning.

Turning to the ranking data, the percentages of students selecting the three valid arguments as the three most logical for grades four, seven, and college respectively were, as noted earlier, 0%, 35%, and 61%, $\chi^2(2,N = 81) = 21.1$, $p < .001$. Most of the students who did not select the three valid arguments as most logical instead selected the two arguments (1 and 2) in which each of the three component propositions was empirically true. This pattern accounted for the final ranking of 55% of the fourth graders, 62% of those seventh graders who did not use a validity ranking, and 63% of those college students who did not use a validity ranking. Similarly, students not using validity commonly chose as least logical either the three arguments (5, 6, and 7) containing two or more false propositions (25%, 46%, and 31% for grades 4, 7, and college, respectively) or the three arguments (3, 6, and 7) with false conclusions (20%, 23%, and 6% for the three groups, respectively). In explaining their rankings, students not ranking by validity typically gave empirical truth (or some combination of validity and truth) as their primary rationale (80%, 85%, and 87% for the three groups, respectively). Overall, the results of the rankings suggest a developmental trend from (a) a very strong tendency among fourth graders to interpret "most logical" as meaning most empirically true to (b) a modal tendency among college students to interpret the same expression, at least in the context of the present set of tasks, as meaning most valid.

Remaining questions.—The results so far show important changes beyond age 10 in how people think about arguments. Fourth graders failed to dis-

Table 2
Percentage of Students Using Various Bases for Sorting Arguments in Experiment 1

	Grade 4		Grade 7		College	
Basis for sorting	Trios[a]	Allt[b]	Trios[a]	Allt[b]	Trios[a]	Allt[b]
Validity	0	0	35	20	78	46
Form	0	0	15	5	51	34
Truth	60	45	70	30	66	34
Content	100	85	90	95	95	85
Partial matching	30	15	35	25	17	15
Miscellaneous	10	0	10	0	46	39
Mixed[c]	–	25	–	40	–	–

[a] Percent using the indicated concept or strategy in sorting at least one of the three sets of three arguments.
[b] Percent using the indicated concept or strategy in sorting the set of all seven arguments.
[c] Mixed sortings were possible only for fourth and seventh graders sorting the set of all seven arguments.

tinguish arguments on the basis of validity despite a number of opportunities to do so. Their failure cannot be explained in terms of limited ability to verbalize their understanding since the age trend on the ranking task (where no verbalization was required) paralleled that on the sorting task. These data are consistent with what we may call the competence deficit hypothesis: fourth graders do not understand the concept of validity.

There are, however, at least two other plausible explanations. One possibility is that the results reflect not a competence deficit but rather a general performance deficit: fourth graders may understand the concept of validity but do not spontaneously apply it in distinguishing and evaluating arguments. A third possibility is that the fourth graders have a specific performance deficit: they understand the concept of validity but have trouble applying it to particular forms of argument such as the transitive arguments that were predominant in Experiment 1.

The main purpose of Experiment 2 was to decide among these three alternatives with respect to the fourth graders. A secondary purpose was to consider these same three possibilities with respect to the difference in performance between the seventh graders and the college students. The possibility of a specific performance deficit was explored by (a) presenting five types of argument, all known to be simple for young children, and (b) including a separate "control" task designed to assess facility with each argument type without requiring students to think about the validity of the argument as a whole. The possibility of a general performance deficit was investigated by defining validity, discussing examples of valid and invalid arguments, and specifically asking students to use this concept in evaluating the arguments on the main task. The intent was not to teach the concept of validity but to elicit it if it was already within the student's competence.

Experiment 2

Method

Subjects.—The participants, all new volunteers, were 18 fourth graders (13 boys and five girls), ranging in age from 9-1 to 11-4 (mean = 9-9); 18 seventh graders (13 boys and five girls), ranging in age from 12-7 to 13-11 (mean = 13-0); and 20 college undergraduates (five males and 15 females), ranging in age from 19 to 23 (mean = 20.9), none of whom had taken a logic course. Children were recruited from Lincoln public schools and the college students from a class required of all education majors at the University of Nebraska–Lincoln.

Materials.—Each participant completed a six-page test booklet and a two-page control task. The first page of the booklet defined what it means for an argument to be (a) valid ("the last line follows from the earlier information. In other words, the last line would have to be true if the earlier information were true") and (b) invalid ("the last line does not follow from the earlier information. In other words, the last line could be false even if the earlier information were all true"). It then provided simple examples of (a) a valid argument with true premises and conclusion, (b) a valid argument with false premises and conclusion, (c) an invalid argument with true premises and conclusion, and (d) an invalid argument with false premises and conclusion. In each case the explanation indicated whether the argument was valid, explained why, and highlighted the difference between validity and truth. The remaining pages of the booklet presented 40 arguments, with instructions to label each as valid or invalid.

Five forms of valid argument were selected: *Transitivity* (e.g., "Abe is taller than Bob; Bob is taller than Chuck; therefore, Abe is taller than Chuck"); *Class instantiation*, corresponding to Braine and Rumain's (1983) PL12 (e.g., "All horses are fish; Blacky is a horse; therefore, Blacky is a fish"); *Disjunction*, corresponding to Braine and Rumain's N8 (e.g., "Either bears fly or birds fly; bears do not fly; therefore, birds fly"); *Conjunction*, corresponding to Braine and Rumain's N1 (e.g., "Cars have motors; trucks have motors; therefore, cars and trucks have motors"); and *Reverse conjunction*, corresponding to Braine and Rumain's N2 (e.g., "Nickels and dimes are plants; therefore, dimes are plants"). The transitivity form had been used in Experiment 1 on the basis of evidence of its simplicity. The other four forms are all among those proposed by Braine and Rumain (1983) as fundamental inference schemas of natural logic. Evidence summarized by Braine and Rumain suggests that all are well understood by age 5 or 6.

For each of the five argument forms, an analogous but invalid form was constructed (e.g., for conjunction, "Cars have motors; trucks have motors; therefore, trucks and buses have motors"), thus yielding 10 argument forms (one valid and one invalid for each type of argument). Next, two variants of each of the 10 forms were constructed in such a way as to maintain the validity or invalidity of the form (i.e.,

by reversing the order of the premises or the order of two conjoined or disjoined terms), thus yielding 20 argument forms (two valid and two invalid for each type of argument). Three arguments were then constructed for each of the 20 forms by filling in true content (e.g., "Cars have motors," as in the example above), false content (e.g., "Cars have wings"), or neutral content (e.g., "Blorks have tails"), thus yielding 60 arguments. Because that seemed an excessive number, 40 were systematically selected such that (a) half of the arguments in each content category (true, false, neutral) were valid and half invalid, and (b) true and false content were equally represented in both the valid and invalid variants of each of the five types of argument. A single random order of the 40 arguments was used for half of the test booklets at each age and the reverse of that order for the other half.

The control task included ten written arguments with neutral content, one for each of the ten valid argument forms included in the test booklet. For each of the ten arguments, the student was asked to choose the better of two conclusions (e.g., "Stan is older than Bob; David is older than Stan; (A) Therefore, David is older than Bob; (B) Therefore, Bob is older than David"). In each case, one conclusion followed necessarily from the premises and the other was inconsistent with them.

Procedure.—Students were tested in groups of no more than four so that the experimenter could monitor their attention to the definitions and examples and their understanding of the instructions. The first author presented each student with a test booklet. He then read the definitions, examples, and instructions aloud while students read along. After soliciting questions, he remained present while students worked on the test booklets. None had any difficulty following the directions. After completing the test booklet, each student then completed the control task. The entire session typically took 20–30 min.

Results

Performance was analyzed with respect to each of two criteria—one fairly stringent and one quite lax. The stringent criterion for success was correct evaluation of at least 90% of the 40 arguments as valid or invalid. The percentages of students meeting this criterion for grades four, seven, and college, respectively, were 11%, 67%, and 80%, $\chi^2(2, N = 56) = 19.9$, $p < .001$. The proportion of males and females meeting the criterion did not differ significantly at any age.

The lax criterion was correct evaluation of at least 70% of the arguments, representing a performance which, though quite inconsistent, was significantly ($p < .01$, one-tailed) above chance level. The percentages of students meeting this criterion for grades four, seven, and college, respectively, were 56%, 89%, and 100%. Most of the fourth graders thus met the lax criterion, though their performance as a group remained significantly inferior to that of the seventh graders, $\chi^2(1,N = 36) = 4.98$, $p < .05$, and college students (binomial $p < .01$). Again, there were no sex differences.

Specific performance deficit.—One possible explanation for the observed age differences is that younger children have trouble understanding some of the specific arguments used. It is certainly plausible that an individual might understand the distinction between valid and invalid arguments but be unable to determine, in the case of a particular type of argument, which variants fit in which category.

To explore this possibility, the present study included a control task involving the same five types of argument as the validity task but merely requiring subjects to select the best conclusion for a given set of premises rather than to evaluate arguments as valid or invalid. Performance on the control task was excellent at each age. No subject scored less than eight out of 10. Mean scores out of 10 (and % scoring 10 out of 10) for grades four, seven, and college, respectively, were 9.6 (67%), 9.7 (78%), and 10.0 (100%). These results provide no support for the hypothesis that the younger subjects were simply confused by the types of argument used.

Further analysis showed that the overall age trend on the main task held for each of the five types of argument (see Table 3). Differences among argument types in frequency of consistently correct performance for the fourth graders were not significant (binomial $p > .10$ for each pair). Again, the specific performance deficit hypothesis is not supported.

General performance deficit versus competence deficit.—Although there appear to be strong grounds for ruling out the specific performance deficit hypothesis, it is not so easy to choose between the other two alternatives: general performance deficit versus competence deficit. The crux of the matter is the basis for the partial success of the substantial number of fourth graders who met the lax (70%) criterion for success on the present task but not the more stringent (90%) criterion. One possibility is that these children understand the concept of validity

Table 3
Percentage of Students at Each Grade Showing Perfect Performance (8/8) on Each Type of Argument in Experiment 2

Type of Argument	Grade 4	Grade 7	College
Transitivity	17	50	85
Class instantiation	6	44	50
Conjunction	28	72	70
Reverse conjunction	28	67	90
Disjunction	22	61	85

but were somehow unable to apply it with reasonable consistency. The other possibility is that the marginal fourth graders have heuristics for correctly classifying many arguments but that their inconsistency reveals their lack of a genuine grasp of validity.

Further analysis aimed at addressing more directly just how children were reasoning about the arguments presented. The 40 arguments in the test booklet included (a) 16 arguments with neutral content, (6) 12 arguments in which truth and validity corresponded (six valid arguments with true content and six invalid arguments with false content), and (c) 12 arguments in which truth and validity conflicted (six valid arguments with false content and six invalid arguments with true content). This final set of 12 may be considered the core arguments on the test since, by definition, a genuine grasp of the concept of validity involves recognition of the distinction between validity and empirical truth. Moreover, the results of Experiment 1 suggest that fourth graders understand the concept of empirical truth and use it as a basis for distinguishing arguments and evaluating their logicality.

How can total score on these 12 critical arguments be interpreted? A student who was consistently responding on the basis of validity would obviously obtain a score of 12. A student who was consistently evaluating items with true content as valid and those with false content as invalid (despite explicit instructions to the contrary) would miss every one of these items and thus score 0. Finally, a student responding randomly would be expected to obtain a score of about 6. A binomial test showed that scores of 0, 1, 11, and 12 deviate significantly ($p < .01$) from chance performance. Accordingly, all students were divided into three categories: (a) score of 0 or 1(indicating evaluation on basis of empirical truth), (b) score between 2 and 10 (not significantly different from chance), and (c) score of 11 or 12 (indicating evaluation on the basis of validity).

The percentages of students responding on the basis of validity at a level significantly above chance for grades 4, 7, and college, respectively, were 11%, 61%, and 75%, $\chi^2(2, N = 56) = 16.8$, $p < .001$. Seventh-grade performance was significantly better than that of the fourth graders, $\chi^2(1, N = 36) = 9.75$, $p < .01$, but did not differ significantly from that of the college students, $\chi^2(1, N = 38) = 0.85$, $p > .10$. Thus, on these critical items, very few fourth graders showed a grasp of validity and there was a sharp increase with age. The results, however, do not support the view that those students not responding on the basis of validity were systematically responding on the basis of truth. Only three students—two fourth graders and a seventh grader—selected the arguments with true content as valid at a level significantly greater than chance. What then were the remaining students, including 78% of the fourth graders, doing?

Although the performance of these remaining students did not differ significantly from chance, it seems unlikely that they were actually responding randomly. As we have already seen, most of the fourth graders and nearly all of the older students scored significantly above chance level on the test as a whole. A more plausible alternative is suggested by the excellent performance of all the students on the "control" task. Although the control task was purposely designed so as not to require students to think about the validity of entire arguments, it is difficult to see how one could consistently select proper deductive conclusions for the 10 sets of premises on this task without at least an implicit grasp of logical form. It is likely that even the fourth graders were also responding to the main task on the basis of logical form but that, because they do not really understand validity, they tended to incorporate elements of truth and falsity in making their judgments.

One source of support for the view that students have an implicit understanding of form prior to understanding inferential validity is performance on the 16 arguments with neutral content, in which empirical truth is not an issue. A score of 14 or better on these arguments is significantly above chance level (binomial $p < .01$, one-tailed). This criterion was met by 39% of the fourth graders, 78% of the seventh graders, and 90% of the college students, $\chi^2(2,N = 56) = 12.5$, $p < .01$. Moreover, 94% of the fourth graders (as well as 94% of the seventh graders and 100% of the college students) scored above the chance score of 8, even if not significantly so. It is difficult to account for the fact that nearly all the fourth graders scored above chance on the neutral arguments, and many significantly so, without postulating some awareness of argument form.

If it is true, however, that younger children also take truth into account when this is possible, then they should be less likely to respond correctly to arguments where validity conflicts with truth than to arguments where they correspond. To test this, each student's score on the 12 critical items where truth and validity conflicted was compared with his or her score on the 12 items where they corresponded. Of the 18 fourth graders, 10 scored higher on arguments where truth and validity corresponded than on arguments where they conflicted and only one showed the reverse pattern (binomial $p < .01$, one-tailed). For eight of the 10, the difference was fairly substantial (3 or more points). Overall, whereas 33% of the fourth graders scored significantly above chance level in distinguishing valid from invalid arguments when validity corresponded to truth, only 11% did so when truth and validity conflicted.

Of the 18 seventh graders, only six scored higher on arguments where truth and validity corresponded than on arguments where they conflicted, and seven showed the reverse pattern (binomial $p > .10$). Only four of the 13 differences were substantial (3 or more points). Similarly, only three college students scored higher on arguments where truth and validity corresponded, and seven showed the reverse pattern (binomial $p > .10$). None of these differences exceeded 2 points. Thus, differences in performance on the two categories of arguments tended to be small for the older students and did not favor one category over the other, whereas for the fourth graders the differences were more often substantial and strongly tended to favor arguments where validity corresponded to truth. Despite explicit instructions to evaluate arguments on the basis of validity and to ignore truth, fourth graders were less likely to evaluate an argument as valid if its content was false than if its content was true.

Although these data support the earlier conclusion that empirical truth played an important role in the reasoning of many fourth graders, they also support the conclusion that it was rarely, if ever, the sole consideration. Thus these analyses further support the view that subjects not reasoning on the basis of an explicit understanding of validity do consider the form of arguments but are often inconsistent about this and, in particular, commonly incorporate considerations of empirical truth. This is what one would expect from an individual whose understanding of logical form is only implicit in the unconscious operation of basic inference schemata (Braine, 1978) rather than an object of explicit awareness.

On the whole, the results tend to support the competence deficit hypothesis with respect to the fourth graders. Even after careful definition of validity, detailed examples, and explicit instructions to use this concept in evaluating arguments, only two fourth graders were fairly consistent in distinguishing valid from invalid arguments for the entire set of 40 and only two (the same two students) showed a significant tendency to distinguish validity from truth on the critical set of 12. The difference in Experiment 1 between seventh graders and college students, by contrast, appears to have been due to a general performance deficit in that seventh graders performed at the level of college students under the favorable conditions of Experiment 2. Seventh graders are less likely than college students to spontaneously use the concept of validity (a performance deficit revealed in Experiment l), but most do grasp that concept (a competence revealed in Experiment 2).

Most 9–10-year-olds apparently do not grasp the concept of inferential validity, and a genuine grasp of this concept and its distinction from empirical truth develops rapidly beyond that age. It might still be argued, however, that many of the fourth graders really do grasp the concept of validity but were not sufficiently consistent in applying it for their competence to become apparent in the present experiment. The aim of Experiment 3 was to decide between these possibilities by giving students feedback after each judgment of validity. If the competence deficit explanation is correct, the fourth graders are really responding on a variety of bases other than validity and would find it difficult to profit systematically from the feedback. If, instead, they do

grasp validity and are simply inconsistent in applying it, then regular feedback should improve the consistency of their performance and thus reveal their underlying competence.

Experiment 3

Method

Subjects.—The participants, all new volunteers, were 20 fourth graders (three boys and 17 girls), ranging in age from 9-0 to 10-5 (mean = 9-7); 20 seventh graders (12 boys and eight girls), ranging in age from 12-5 to 13-5 (mean = 13-0); and 20 college undergraduates (six males and 14 females), ranging in age from 19 to 43 (mean = 22.0), none of whom had taken a logic course. Once again, the children were recruited from Lincoln public schools and the college students from a course required of all education majors at the University of Nebraska–Lincoln.

Materials.—Each of the 40 arguments from Experiment 2 was typed on a separate index card. A single random order of the cards was used for half of the subjects in each age × explanation group (see below) and the reverse of that order for the other half.

Procedure.—Each student was assessed individually by the second author. Half of the students at each age were randomly assigned to the explanation condition. The experimenter first explained the concept of validity to them, using the definitions and examples from Experiment 2, and solicited questions. She then presented the index cards one at a time and asked them to sort each argument into one of two piles—valid versus invalid. After each argument, the experimenter indicated whether the student's choice was correct or incorrect and allowed the student, if necessary, to move the card to the correct pile. Students were permitted at any time to look back at cards already sorted for guidance.

The other half of the students at each age performed the same task but without an initial explanation of validity. They were simply told that a rule existed for sorting the cards and that their task was to use the feedback from the experimenter to discover the rule.

Scoring.—Students were credited with understanding validity if they correctly placed 90% of the 40 arguments or 90% of the last 20 or 90% of the last 10.

Results

Age differences were substantial: The criterion of success was met by only 10% of the fourth graders but by 75% of the seventh graders and 90% of the college students, $\chi^2(2, N = 60) = 30.6$, $p < .001$. As in Experiment 2, seventh-grade performance significantly exceeded that of the fourth graders, $\chi^2(1, N = 40) = 17.3$, $p < .001$, but did not differ significantly from that of the college students (binomial $p > .10$). Differences between the explanation and no-explanation groups were not significant. Most seventh graders and college students achieved criterion regardless of whether they received prior explanation of validity, while few fourth graders (one out of 10 in each condition) did so. The mean numbers of correct choices for the explanation and no explanation groups, respectively, were 27.9 and 23.8 for grade 4, 34.9 and 30.1 for grade 7, and 37.9 and 36.1 for college students. Once again, there were no sex differences.

These results provide further support for a competence deficit explanation with respect to fourth graders and a general performance deficit explanation with respect to seventh graders. Even with regular feedback, most fourth graders were unable to achieve reasonable consistency in sorting simple arguments on the basis of validity. Seventh graders, in contrast, were able to use the feedback to attain a level of performance comparable to that of college students, thus revealing their underlying competence.

General Discussion

Over the course of three studies of the concept of validity, we have varied (a) the nature of the task (free sorting of several arguments, explanation of sortings provided by experimenter, ranking from most to least logical, evaluation of single arguments as valid or invalid); (b) the forms of argument (transitive, disjunctive, etc.); (c) whether subjects were interviewed individually or tested in groups using test booklets; (d) whether or not subjects were provided with an initial explanation of validity; (e) whether or not they were provided with regular feedback on their responses; and (f) whether or not they were required to explain the basis for their responses.

The consistent difficulty of the fourth graders across this wide range of tasks, arguments, testing conditions, and criteria has led us to conclude that they do not understand the concept of validity (a competence deficit) rather than that they were failing to apply it in particular circumstances (a performance deficit). The hypothesis that fourth graders merely have trouble with particular types of ar-

gument (e.g., the transitive arguments of Experiment 1) was rejected on the basis of poor performance in Experiments 2 and 3 on a wide variety of arguments that were purposely selected (and independently shown) to be very simple. The possibility that fourth graders do not spontaneously think about arguments in terms of validity but can do so if asked to was ruled out on the basis of their failure to profit from the explicit directions of Experiment 2. The possibility that they are merely inconsistent and need systematic feedback to keep them on track is countered by their inability to profit from such feedback in Experiment 3. Finally, the possibility that they do think in terms of validity but simply fail to provide adequate verbal explanation of their understanding is ruled out by the fact that no verbal explanation was required in Experiments 2 or 3 or in the ranking task of Experiment 1.

Nevertheless, no claim of lack of competence can ever be conclusively proved. It remains possible that some new task and/or conditions can be devised that will reveal a grasp of validity in children younger than the present results indicate. A key issue in attempting to devise such a task is ascertaining that, although it removes any inappropriate sources of difficulty, it still constitutes a genuine test of the concept of validity.

It might be noted, for example, that all of the present experiments involved arguments varying not only in validity but also in argument type, content, and/or truth status of premises and conclusions. It could reasonably be hypothesized that, after brief guidance, fourth graders might fare quite well in distinguishing valid from invalid variants of arguments if all the arguments were of the same general type (e.g., conjunctive), similar in content (e.g., animals), and neutral in truth value. But it is questionable whether such evidence would demonstrate a genuine grasp of validity. As we have used that term, understanding the concept of validity by definition includes differentiating it from empirical truth and generalizing across some range of content areas and argument types. Further research aimed at demonstrating greater competence at earlier ages should involve tasks that, though simplified, still require a sufficiently differentiated and generalizable concept of validity to support the claim that what is being assessed really is validity and not a more primitive heuristic.

Although it remains possible that fourth graders understand validity, the current evidence strongly suggests that they do not. A useful direction for future research would be to further explore what they do understand about the nature of logic.

We have seen that the fourth graders did not spontaneously evaluate arguments on the basis of their validity, a finding consistent with previous evidence that children of this age do not think about logical form independent of empirical truth (Cummins, 1978; Osherson & Markman, 1975). Even after (a) careful definition, examples and instructions and/or (b) systematic trial-by-trial feedback, very few consistently distinguished valid from invalid variants of even the simplest arguments. The fourth graders did, however, show an impressive ability to select the correct conclusions for sets of premises reflecting these same types of arguments, a finding consistent with extensive evidence on deductive reasoning in young children (Braine & Rumain, 1983; Hawkins et al., 1984). Moreover, most scored significantly above chance level in distinguishing valid from invalid arguments in Experiment 2.

We can account for the above pattern of findings by postulating (in accord with Moshman & Timmons, 1982) that 9–10-year-olds do not explicitly think about logical form but are nonetheless implicitly aware of it. They can use their basic inference schemata (including an implicit awareness of form) to reach or recognize logically necessary conclusions and can associate this with the idea of validity. However, because they do not think explicitly about form, they do not make the abstract differentiation between validity of form and truth of content. For this reason, they do not spontaneously sort arguments on the basis of form or validity (Experiment l), they incorporate consideration of empirical truth in judging the validity of arguments even after instructions and counterexamples (Experiment 2), and they fail to profit from systematic feedback (Experiment 3). Many seventh graders, by contrast, spontaneously distinguished arguments on the basis of form and validity, though evaluations based on empirical truth remained common. With appropriate definition, examples, instructions, and/or feedback, most were reasonably consistent in evaluating arguments on the basis of validity. Thus most 12–13-year-olds, unlike most 9–10-year-olds, apparently do have a sufficiently explicit concept of logical form to differentiate validity of form from empirical truth, though many failed to spontaneously apply this competence in Experiment 1.

Finally, almost all college students spontaneously distinguished arguments on the basis of validity and, after definition, examples, instructions, and/or feedback, were consistent in doing so. This impressive performance stands in sharp contrast to the well-documented difficulties they have on a variety of complex logical reasoning tasks (e.g., Evans, 1983). Although college students often have genuine difficulty with abstract deductive reasoning, they do have a clear "sense of the game."

It is noteworthy that, whereas fourth graders showed little grasp of validity in any of the present experiments, and college students showed excellent performance in all of them, seventh graders did substantially better in Experiments 2 and 3, involving definition, examples, instructions, and/or feedback, than in Experiment 1, involving spontaneous application of the concept. Seventh-grade performance was about midway between fourth-grade and college performance in Experiment 1 but was statistically indistinguishable from college performance in Experiments 2 and 3. These findings suggest that, whereas application of the concept of validity (performance) increases over a wide age range (Experiment l), the underlying competence develops fairly suddenly between ages 10 and 12 (Experiments 2 and 3).

Development of the concept of validity can thus be divided into two phases (cf. Moshman, 1977; Overton & Newman, 1982). The first phase involves a relatively sudden emergence of the concept between ages 10 and 12. This is followed by increasing ability to use the concept under conditions that are not ideal in eliciting it and/or in facilitating its successful application (e.g., Experiment 1). The evidence for relatively sudden development of competence in this area is consistent with Fischer and Pipp's (1984) suggestion that cognitive development in general is marked by relatively sudden changes in "optimal level," that is, in performance under optimal conditions. The specific ages correspond to what Fischer and Pipp predict for the elementary level of abstract competence. The present data directly address what has long been one of the central issues in the study of the development of logical reasoning: Do children's logical abilities undergo qualitative change? It appears to us that the clearest evidence for such change comes not from studies of first-order inferential abilities (how people reach conclusions from premises) but rather from studies of metalogic (how people construe the nature of logic) (e.g., Bereiter et al., 1979; Osherson & Markman, 1975; Pieraut-Le Bonniec, 1980; Somerville et al., 1979; see also Braine & Rumain, 1983). Moshman and Timmons (1982) proposed that, in addition to the relatively early changes related to the concept of logical necessity that have already been documented in the literature, there is a second and later qualitative shift consisting of comprehension of the concept of inferential validity and its differentiation from empirical truth. The results of the present research support that hypothesis.

References

Bereiter, C., Hidi, S., & Dimitroff, C. (1979). Qualitative changes in verbal reasoning during middle and late childhood. *Child Development, 50,* 142-151.

Bickhard, M. H. (1978). The nature of developmental stages. *Human Development, 21,* 217- 233.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85,* 1-21.

Braine, M. D. S., & Rumain, B. (1983). Logical reasoning. In J. H. Flavell & E. M. Markman (Eds.), P. H. Mussen (Series Ed.), *Handbook of child psychology: Vol. 3. Cognitive development* (pp. 263-340). New York: Wiley.

Breslow, L. (1981). Reevaluation of the literature on the development of transitive inferences. *Psychological Bulletin, 89,* 325-351.

Cummins, J . (1978). Language and children's ability to evaluate contradictions and tautologies: A critique of Osherson and Markman's findings. *Child Development, 49,* 895-897.

Evans, J. St. B. T. (Ed.). (1983). *Thinking and reasoning: Psychological approaches.* London: Routledge & Kegan Paul.

Fischer, K. W., & Pipp, S. L. (1984). Processes of cognitive development: Optimal level and skill acquisition. In R. J. Sternberg (Ed.), *Mechanisms of cognitive development* (pp. 45-80). New York: W. H. Freeman.

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.

Inhelder, B., & Piaget, J. (1958).*The growth of logical thinking: From childhood to adolescence.* New York: Basic.

Kelly, G. A. (1955). *The psychology of personal constructs.* New York: Norton.

Moshman, D. (1977). Consolidation and stage formation in the emergence of formal operations. *Developmental Psychology, 13,* 95-100.

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

Neimark, E. D., & Chapman, R. H. (1975). Development of the comprehension of logical quantifiers. In R. J. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 135- 151). Hillsdale, NJ: Erlbaum.

Osherson, D. N., & Markman, E. (1975). Language and the ability to evaluate contradictions and tautologies. *Cognition, 3,* 213-226.

Overton, W. F., & Newman, J. L. (1982). Cognitive development: A competence-activation/utilization approach. In T. M. Field, A. Huston, H. C. Quay, L. Troll, & G. E. Finley (Eds.), *Review of human development* (pp. 217- 241). New York: Wiley.

Pieraut-Le Bonniec, G. (1980). *The development of modal reasoning: Genesis of necessity and possibility notions.* New York: Academic Press.

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