1979

# Development of Formal Hypothesis-Testing Ability

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

Published in *Developmental Psychology, 15*:2 (1979), pp. 104-112. 

Submitted February 27, 1978.

# Development of Formal Hypothesis-Testing Ability

David Moshman

*University of Nebraska–Lincoln*

It was postulated that formal operational hypothesis-testing ability includes at least three cognitive capacities: (a) implication comprehension, the ability to understand conditional relationships; (b) falsification strategy, the realization that to test a hypothesis, one must seek information that would falsify it; and(c) nonverification insight, the realization that hypotheses are not conclusively verified by supporting data. A total of 24 males in each of Grades 7, 10, and college evaluated data descriptions with respect to each of four hypothesized implication relationships and chose an experiment to test each hypothesis. Results suggested three sequences of qualitative change in hypothesis-testing ability: (a) from no systematic interpretation of conditionals to an implication interpretation, (b) from content-based information seeking to a falsification strategy, and (c) from a symmetrical conception of truth and falsity to a non-verification insight. However, formal operational performance was far from universal, even in college students.

The ability to test hypotheses, which has been postulated to play a central role in a variety of cognitive processes (e. g. , Inhelder & Piaget, 1958; Neimark & Santa, 1975; Wason & Johnson-Laird, 1972), is one of those very general aspects of cognition that can be studied in some form at any level of development. At the level of formal operations, one might expect the development of relatively conscious hypothesis testing based on conceptualization of hypotheses as such. Both Piagetian theory and available evidence (see below) suggest the involvement of at least three cognitive capacities: (a) comprehension of implication relationships (if *p* then *q*; all *p* are *q*; or any logically isomorphic expression), since this is the form hypotheses usually take; (b) falsification strategy, the realization that to test a hypothesis, one must seek information that would falsify it; and (c) nonverification insight, the realization that hypotheses are not conclusively verified by supporting data. The aim of this study was to trace the development of these three aspects of hypothesis-testing ability.

Beginning with the first, one would expect implication comprehension to develop during adolescence, since implication is one of the 16 bivariate relationships composing the complete combinatorial system—the lattice structure underlying the stage of formal operations (Inhelder & Piaget, 1958). Of course, young children routinely succeed on various reasoning tasks apparently involving implications (e.g., Bourne & O'Banion, 1971; Kodroff & Roberge, 1975; Kuhn, 1977); the question is whether they conceptualize the putative implications as such or, rather, assimilate them to a more primitive cognitive structure. Studies analyzing in-

This article is based on a thesis submitted in partial fulfillment of the requirements for a PhD degree at Rutgers University, and was presented in an earlier version at the meeting of the Eastern Psychological Association, Boston, April 1977. I am grateful to my advisor, Edith Neimark, for her valuable assistance throughout the project, and to the other members of my dissertation committee: David Brodzinsky, George Pallrand, John Santa, and Juliet Vogel. For assistance in obtaining subjects, I am indebted to Mike Novello and William Williams, both of Polytechnic Preparatory Country Day School, and also to Harold Rubenstein, Jean Sindhikara, and Mike Lieberman. Finally, I would like to thank Rick Bady, Marilyn Edmonds, and Royce Ronning for comments on earlier drafts of the manuscript.

trasubject patterns across a range of deductive reasoning problems suggest that implication comprehension does not emerge until adolescence (Staudenmayer & Bourne, 1977; Taplin, Staudenmayer, & Taddonio, 1974).

Perhaps the most direct way to assess comprehension of a logical relationship between variables is to present a broad range of data descriptions and to see if people can correctly evaluate these as consistent or inconsistent with the relationship in question. Ward (1972) and Paris (1973) both used this evaluation procedure to assess implication comprehension and found very poor performance even by adolescents, with no developmental trends. One explanation is that the implications in both of these studies involved arbitrary relationships(e. g., in the Paris study, "If the bird is in the nest, then the shoe is on the foot"), which seem to be a hindrance to mature reasoning (Bracewell & Hidi, 1974). In the present study, I attempted to show developing comprehension of implication during adolescence, using an evaluation procedure with implications expressing a meaningful (causal) connection between meaningful terms.

The second aspect of hypothesis-testing ability—the strategy of testing hypotheses by seeking disconfirmatory cases—has also been postulated to involve formal operations (Beth & Piaget, 1966, pp. 181–182). The crucial importance of this falsification strategy to the complete combinatorial system is apparent when one considers the latter not as a juxtaposition of implication and 15 other logical relationships but rather as a structure d'ensemble, that is, a tightly knit cognitive structure involving not only the component operations but the complex net of interrelationships linking them. Suppose, for example, that a formal operational thinker, having encountered several people who use fluoridated toothpaste ($p$) and have healthy teeth ($q$), as well as some who don't use fluoridated toothpaste ($\bar{p}$) and don't have healthy teeth ($\bar{q}$), becomes interested in the relationship between the use of fluoridated toothpaste and the health of one's teeth. The available evidence already rules out 12 possible relationships, leaving four: (a) implication (if $p$ then $q$), (b) converse implication (if $q$ then $p$), (c) biconditionality ($q$ if and only if $p$), and (d) complete affirmation ($p$ and $q$ mutually independent). To decide among these, the formal thinker considers their interrelationships and thus seeks information to efficiently narrow down the possibilities. For example, seeking someone who does not use fluoridated toothpaste and does have healthy teeth ($\bar{p} \cdot q$) would be a useful experiment, since such evidence would rule out converse implication and biconditionality, leaving only implication and complete affirmation. One could decide between these latter by seeking people who use fluoridated toothpaste and don't have healthy teeth ($p \cdot \bar{q}$), since they would rule out the implication relationship.

The formal thinker thus reasons within the complete combinatorial system not by confirming particular relationships but by systematically excluding others, and this is accomplished by seeking falsifying information. Over the last decade, experiments involving Wason's "four-card task" have generated considerable evidence regarding use of the falsification strategy by college students. Although earlier studies showed surprisingly poor performance (Wason & Johnson-Laird, 1972), recent data show greater achievement on versions of the four-card task with meaningful content (e.g., Bracewell & Hidi, 1974; Gilhooley & Falconer, 1974), which is consistent with Piaget's(1972) suggestion that formal operations may only be applied to familiar materials. There seems to be no published evidence regarding development of the falsification strategy, though unpublished work by Lunzer (Note 1) and Bady (Note 2) does suggest development trends during adolescence, as Piaget's theory predicts. The present study used a variant of the four-card task to assess development of the falsification strategy.

The third aspect of hypothesis-testing ability investigated was the realization that hypotheses cannot be conclusively verified. The importance of this nonverification insight in formal operational thinking is implicit in the preceding discussion of the complete combinatorial system as a structure d'ensemble: The formal thinker does not attempt to verify directly

a particular relationship, realizing that no finite number of consistent instances would establish its truth. No matter how many people use fluoridated toothpaste and have healthy teeth, the hypothesis that all people who use fluoridated toothpaste have healthy teeth remains unproven, though it can easily be disconfirmed by a single person who uses fluoridated toothpaste and does not have healthy teeth. The nonverification insight reflects this crucial asymmetry between verification and falsification.

There appears to be no published evidence directly relating to the nonverification insight. Several studies (Evans, 1972; Johnson-Laird & Tagart, 1969; Wason, 1968) have shown that college students typically evaluate certain consistent information as making an implication rule true, but these subjects were not explicitly asked to evaluate the truth or falsity of the rule as a hypothesis of potentially universal applicability. In the present study, subjects were asked to determine whether each of various pieces of information proved conclusively that a given hypothesis was always correct. Since the nonverification insight, like implication comprehension and the falsification strategy, is involved in the complete combinatorial system, it too was expected to develop during adolescence.

In order to assess generality of application of the three hypothesis-testing schemes, several logical forms and content areas were used in constructing the hypotheses presented to the subjects. On the basis of available data showing the nonuniversality of formal operations in adults (Neimark, 1975; Piaget, 1972) and major effects of form and content on reasoning (e. g. , Wason & Johnson-Laird, 1972), it was expected that consistent application of these hypothesis-testing schemes would be far from universal even in college students.

## Method

### *Subjects*

A total of 24 male volunteers in each of Grades 7, 10, and college participated in the experiment. The seventh and tenth graders attended a highly selective, all-male, private school, and the college students lived in two Rutgers University dormitories. Mean ages (and standard deviations) for Grades 7, 10, and college, respectively were 12 years 10 months (4. 6 months), 15 years 9 months (4. 4 months) and 19 years 4 months (7. 3 months). Five of the college students had taken a course in logic, but they performed no better than the others.

### Materials

A five-page test booklet was prepared for each subject. The first page presented general instructions, including the following:

> In each of the following problems you will have a *theory*[1] that you want to test. *Do not* assume that the theory is true. A theory is a rule that *might be true* and *might be false*. Testing a theory means trying to find out whether it is true or false. Proving a theory is true means that you find *out for sure* that it is *always right.* Proving a theory is false means that you find out *for sure* that it is *sometimes wrong.*

Each of the next four pages presented a hypothesis (e.g., "If a person uses fluoridated toothpaste he will have healthy teeth" is a hypothesis of the form If $p$ then $q$), followed by these instructions:

> For each of the following people, what conclusion would you reach about the theory? Check *Proves True* if you think this person shows *for sure* that the theory is right. Check *Proves False* if you think this person shows *for sure* that the theory is wrong. Check *No Proof* if you think this person does not show *for sure* whether the theory is right or wrong.

Next came eight data descriptions of the form $p \cdot q$ ($p$ and $q$ ), $p \cdot \bar{q}$ ($p$ and not $q$), $\bar{p} \cdot q$, $\bar{p} \cdot \bar{q}$ , $q \cdot p$, $q \cdot \bar{p}$ , $\bar{q} \cdot p$, and $\bar{q} \cdot \bar{p}$ , with space next to each for the subject to check proves true, proves false, or no proof. For example, given the hypothesis above, the eight data descriptions would read: Albert uses fluoridated toothpaste and has healthy teeth, Bertie uses fluoridated toothpaste and does not have healthy teeth, and so on. Following this part of each page, the subject was asked to decide whether it would be better to test the hypothesis by studying qs or by studying qs (in each case to see whether they are $p$ or $\bar{p}$ ). In the above example, the choice would be between (a) asking patients with healthy teeth whether they use fluoridated toothpaste and (b) asking patients who don't have healthy teeth whether they use fluoridated toothpaste (which is the correct choice, since only people who use fluoridated toothpaste and don't have healthy teeth disprove the hypothesis). Finally, at the bottom of each page, the subject was asked to explain the basis for this choice between two experiments.

### *Design*

The four test pages in each booklet were selected from a set of 16, identical in format but differing in the hypothesis

[1] For the benefit of the younger subjects, hypotheses were referred to as theories throughout the experiment.

("theory") presented. The 16 hypotheses were produced by factorially combining 4 logical forms (if $p$ then q; all $p$ are q; if $p$ then $\bar{q}$; no $p$ are $q$) with 4 content areas (toothpaste fluoridation, student performance, canine character, and automotive maintenance). Thus, the 4 hypotheses for the content area toothpaste were: (a) "If a person uses fluoridated toothpaste he will have healthy teeth, (b) all people who use fluoridated toothpaste have healthy teeth, (c) if a person uses fluoridated toothpaste, he will not get cavities, and (d) no one who uses fluoridated toothpaste gets cavities."[2]

Since only 4 of the 16 factorial combinations of form and content were used in each test booklet and since they had to be arranged in some order, the variables form, content, and serial position were necessarily confounded within any test booklet. By constructing the test booklets in groups of four, however, it was possible to use a Greco-Latin square design in which (a) each form, each content area, and of course each serial position appeared exactly once in each test booklet, and (b) the variables form, content, and position were mutually orthogonal within sets of four booklets. Three Greco-Latin squares were randomly selected and used to construct 12 variants of the test booklet, each of which was used by two subjects at each grade level. It was thus possible to analyze the main effects of age, form, content, and position, as well as the interaction of age with form, with content, and with position. Planned comparisons within the variable form were also undertaken, involving (a) positive (if $p$ then q; all $p$ are $q$) versus negative (if $p$ then $\bar{q}$; no $p$ are q) forms, (b) connective (if . . . then) versus quantified (all; no) wordings, (c) the interaction of a and b, and (d) the interaction of each of the above with age (in the case of c, a three-way interaction).

### *Procedure*

Subjects were tested in groups of up to four students. The experimenter handed out test booklets and read the cover sheet aloud, emphasizing the italicized words, while the students read along in their own booklets. After soliciting questions, the experimenter asked the subjects to turn to the next page and pointed out that they each had a "theory" written on top of the page. He then read aloud the instructions following the theory, still emphasizing the italicized words. Questions were again solicited, after which each subject continued on his own. The experimenter remained present in case of difficulties, but even the seventh graders apparently had little trouble understanding the instructions.

## Results and Discussion

### *Implication*

Comprehension of implication was inferred from subjects' evaluations of the eight data descriptions for each hypothesis. To be credited with an implication interpretation, a subject had to produce an evaluation pattern meeting three criteria: (a) each of the last four data descriptions was evaluated identically with its logically identical counterpart among the first four; (b) the subject correctly evaluated as disconfirmatory those data descriptions—and only those data descriptions—that indeed disconfirmed the hypothesis; and (c) any data descriptions evaluated as verifying were not such as to indicate a biconditional interpretation (e.g., identical evaluation of $p \cdot q$ and $\bar{p} \cdot \bar{q}$).[3] The mean numbers of implication interpretations (out of 4 possible) for subjects in Grades 7, 10, and college, respectively, were .75, 1.71, and 2.42 (see Table 1). Trend analysis revealed a significant linear trend, $F(1, 69) = 18.36$, $p < .001$, and insignificant deviation from linearity, $F(1, 69) < 1$.

The most common alternative evaluation patterns reflected biconditionality (interpretation of if as if and only if). The criteria for biconditionality were that (a) each of the last four data descriptions was evaluated identically with its counterpart among the first four; and also ($b_1$) data descriptions matching both antecedent and consequent of the conditional (e.g.,

[2] The meaning of $q$ changes from "healthy teeth" in the first two hypotheses to "getting cavities" in the last two. If $q$ had been held constant (referring to healthy teeth in all four forms), these last two hypotheses would have expressed a negative rather than positive relationship between fluoridation and health of teeth and thus differed from the first two hypotheses in content as well as form of expression. Analogous changes were made in the meaning of $q$ in the other three content areas such that the hypothetical relationship expressed remained essentially unchanged across the four logical forms. This change in $q$ across the four forms within each content area was not a source of confusion for subjects, since each subject was exposed to only one form in any content area.

[3] Inconsistent evaluations were disallowed on the grounds that a formal approach to data evaluation would recognize when superficially different data descriptions are formally identical. Evaluations suggesting a biconditional interpretation were disallowed on the basis of previous data (cited below) that biconditionality is a common misinterpretation of conditional statements. Reanalysis of the data using a very lax criterion (only the first four data descriptions were considered, and these simply had to meet criterion b) resulted in the same pattern of results with respect to age differences and the effects of task variables, though overall level of performance was naturally somewhat higher.

**Table 1.** *Frequency of Implication, Defended Falsification, and Nonverification Responses at Each Grade*

| Grade | No. responses | | | | | Total |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | |
| | | | Implication | | | |
| 7 | 14 | 6 | 2 | 0 | 2 | 24 |
| 10 | 4 | 8 | 6 | 3 | 3 | 24 |
| College | 3 | 6 | 3 | 2 | 10 | 24 |
| Total | 21 | 20 | 11 | 5 | 15 | 72 |
| | | | Falsification | | | |
| 7 | 18 | 4 | 1 | 1 | 0 | 24 |
| 10 | 9 | 3 | 6 | 0 | 6 | 24 |
| College | 10 | 2 | 2 | 2 | 8 | 24 |
| Total | 37 | 9 | 9 | 3 | 14 | 72 |
| | | | Nonverification | | | |
| 7 | 23 | 1 | 0 | 0 | 0 | 24 |
| 10 | 20 | 2 | 0 | 1 | 1 | 24 |
| College | 15 | 1 | 0 | 0 | 8 | 24 |
| Total | 58 | 4 | 0 | 1 | 9 | 72 |

Note. Each subject could give up to four correct responses for each of the three aspects of hypothesis testing.

$p \bullet \bar{q}$ and $q \bullet p$ for if $p$ then $q$) and those matching neither (e.g., $p \bullet q$ and $q \bullet p$ for if $p$ then $q$) were all evaluated as verifying, and/or ($b_2$) descriptions matching the antecedent or consequent but not both (e.g., $p \bullet \bar{q}$, $q \bullet \bar{p}$, $\bar{p} \bullet q$, and $\bar{q} \bullet p$ for if $p$ then $q$) were all evaluated as falsifying. For Grades 7, 10, and college, respectively, the mean numbers of biconditional interpretations (out of four possible) were .96, .58, and .58 (age trend nonsignificant). Many previous studies have also found the biconditional to be a prevalent misinterpretation of conditional statements (e.g., Paris, 1973; Taplin et al., 1974). Perhaps the symmetry of the biconditional makes it easier to comprehend, so people unable to grasp implication instead assimilate the conditional to the simpler biconditionality strategy.

Most response patterns other than implication and biconditionality were inconsistent in that logically identical data descriptions were evaluated differently. Mean numbers of inconsistent interpretations (out of 4 possible) for Grades 7, 10, and college, respectively, were 1.92, 1.50, and .71. This decline, linear trend $F(1, 69) = 13.32$, $p < .001$, considered together with the lack of an age trend for biconditionality, suggests a developmental sequence from inconsistent data evaluations to implication interpretation, with biconditionality very likely an intermediate step in at least some cases.

All three task variables showed significant effects. Planned comparisons revealed that the effect of logical form was entirely due to a higher proportion of implication interpretations for negative hypothesis forms(47%) than for positive forms (34%), $F(1, 189) = 11.15$, $p < .01$. This superficially conflicts with extensive evidence that negation hinders mature reasoning (e.g., Moshman, 1977; Wason & Johnson-Laird, 1972), but it is consistent with evidence that negative universals (no $p$ are $q$) are better comprehended than positive universals (all $p$ are $q$) (Neimark & Chapman, 1975) and that negation of the consequent ($q$) of an implication may facilitate reasoning (Roberge, 1971; Wildman & Fletcher, 1977), or at least not hinder it (Roberge, 1974). This may be because such forms as If $p$ then $\bar{q}$ and No $p$ are $q$ emphasize the single falsifying instance ($p \bullet q$).

The other two main effects were (a) a substantial difference in difficulty among the four content areas, $F(3, 189) = 12.47$, $p < .001$, with proportions of implication interpretation ranging from 28% to 33% to 42% to 60% for dog, student, toothpaste, and engine, respectively, and (b) an increase in implication interpretations from 36% to 46% across the four pages of the test(linear trend $F(1, 189) = 4.52$, $p < .05$). None of these task variables interacted significantly with age: The developmental trend in implication interpretation held for all four logical forms, all four content areas, and all four serial positions.

### *Falsification Strategy*

Use of a falsification strategy was inferred for each subject on each hypothesis from (a) his choice between the two experiments and (b) his explanation of this choice. Choice alone was not a useful criterion because the prob-

ability of a correct guess was 50%. Explanations were categorized (by a coder unaware of the subject's age) as reflecting falsification provided the subject either (a) explicitly indicated that one should test the hypothesis by trying to prove it false or (b) cited as the crucial consideration a potential datum which, according to his own data evaluations, would prove the hypothesis false. Recoding of 50 randomly selected explanations by an independent coder showed 94% agreement.

To be credited with use of a falsification strategy, a subject had to choose the experiment that could falsify the hypothesis in question and defend this choice with a falsification explanation. As Table 1 indicates, the mean numbers of defended falsification choices (out of 4 possible) for Grades 7, 10, and college, respectively, were .38, 1.63, and 1.83, a significant linear trend, $F(1, 69) = 11.85$, $p < .001$, with insignificant deviation from linearity, $F(1, 69) = 2.02$.

Further examination of subjects' choices revealed an interesting pattern. Of the 42 subjects erring on at least one of the positive hypothesis forms (i.e., failing to choose *q* in both cases), 57% chose *q* in both cases, which significantly exceeds chance expectation of 33%, $\chi^2(1) = 10.71$, $p < .01$. This is consistent with Johnson-Laird and Wason's (1970; Wason & Johnson-Laird, 1972) thesis that many people attempt to verify a given hypothesis rather than falsify it. Given the hypothesis If a person uses fluoridated toothpaste, he/she will have healthy teeth, for example, the verification strategy would be to study people with healthy teeth, on the assumption that if such people use fluoridated toothpaste it verifies the hypothesis. The present evidence suggests that this strategy is limited to positive forms (e.g. , if *p* then *q*). Perhaps people are more likely to see such hypotheses as inductive generalizations of certain types of evidence ($p \bullet q$), and thus are more likely to test them by accumulating more such evidence, not realizing that accumulation of positive instances does not test a hypothesis unless it is accomplished in such a way that negative instances could have turned up.

The number of subjects (out of 24) who showed a verification pattern in their responses to positive forms was 5, 10, and 9 for Grades 7, 10, and college, respectively. This constituted only 29% of all error patterns for positive forms in Grade 7, but 71% in Grade 10 and 82% for college students, $\chi^2(2) = 9.22$, $p < .01$, suggesting that a verification orientation toward positive forms was an increasingly important source of error with increasing age.

However, inferring strategies from choice patterns is not without its dangers. In order to obtain further information on use of the verification strategy, as well as to investigate strategies used for negative hypothesis forms and by younger subjects, an examination of nonfalsification explanations was undertaken. Verification explanations—in which the subject either (a) explicitly indicated that one should test the hypothesis by trying to prove it true or (b) cited as the crucial consideration a potential datum which, according to his own data evaluations, would prove the hypothesis true—accounted for 16% of all nonfalsification explanations, though in contrast to the choice pattern results above, this did not vary as a function of age or hypothesis form.

The major source of error revealed in the explanations was a tendency to ignore the form of the hypothesis and the structure of the hypothesis-testing task and to respond instead to idiosyncratic aspects of the content area (e.g., one should study people with healthy teeth "because a patient with healthy teeth would be most likely to explain why his teeth were healthy and what measures he took to keep them in that condition"). Such explanations, resembling the sort of reasoning reported in cross-cultural work by Cole and Scribner (1973), accounted for 71% of all explanations in Grade 7, 18% in Grade 10, and 9% for all college students. One could thus postulate a developmental sequence from content-based hypothesis testing to falsification strategy, with verification strategy a possible intermediate step, at least in the case of positive hypotheses. Since verification strategy, like falsification strategy, involves consideration of hypothesis form and task structure, it is a plausible transitional pattern, though the pres-

ent cross-sectional data cannot show that it is a necessary prerequisite for falsification.

The only significant main effect of a task variable was a difference between content areas. Proportions of defended falsification choices were 26%, 28%, 35%, and 39% for dog, student, engine, and toothpaste, respectively, $F(3, 189) = 3.05$, $p < .05$. No simple explanation for the effect of content seems feasible, since the relative difficulty of the four content areas with respect to falsification differs from their relative difficulty with respect to implication comprehension.

There was only one significant interaction—Age × Positive versus Negative Forms, $F(2, 189) = 4.11$, $p < .05$—and even this interaction did not seriously comprise the main effect for age: Post hoc Scheffe tests revealed significantly better performance by tenth graders than by seventh graders for both positive, $F(5, 282) = 15.39$, $p < .05$, and negative, $F(5, 282) = 45.91$, $p < .001$, forms, and nonsignificant differences between tenth graders and college students in each case. Thus, the existence and direction of the age trend in falsification strategy was not a function of form, content, or serial position.

*Nonverification*

A subject demonstrated nonverification insight for any hypothesis by (a) not choosing any of the eight data descriptions as conclusively verifying the hypothesis and (b) choosing at least one data description(any one) as conclusively falsifying it. Occasional response patterns recognizing the impossibility of conclusive verification (thus meeting criterion a) but considering conclusive falsification equally impossible (thus not meeting criterion b) were not considered to reflect nonverification, since they seem to be based not on insight into the asymmetrical nature of truth and falsity but rather on a generally skeptical attitude concerning the possibility of reaching conclusions. As Table 1 indicates, the mean numbers of nonverification patterns (out of 4 possible) for Grades 7, 10, and college, respectively, were .04, .38, and 1.38, a significant (though modest) linear trend, $F(1, 69) = 13.60$, $p < .001$, with insignificant deviation from linearity, $F(1, 69) = 1.13$).[4] A tendency to cite data descriptions "fitting" the hypothesis (e.g., $p \cdot q$ for hypotheses of the form If $p$ then q) as conclusively verifying it accounted for 53% of all errors, and is consistent with the verification strategy discussed above. The resulting evaluation patterns typically resembled the modal "defective truth table" (e.g., for If $p$ then $q$, $p \cdot q$ is confirming, $p \cdot \bar{q}$ disconfirming, and $\bar{p} \cdot q$ and $\bar{p} \cdot \bar{q}$ irrelevant) found in previous research (Evans, 1972; Johnson-Laird & Tagart, 1969; Wason, 1968).

One explanation for the scarcity of the nonverification insight is that most people do not sufficiently distinguish data and hypotheses. Rather, the two are seen as co-existing in the same general realm: If there is a fit between hypothesis and datum, this is seen as verifying both; if there is a mismatch, it is seen as falsifying. With the development of formal operations, there is an increasingly habitual and systematic orientation toward possibilities and increasing grasp of the subtle but crucial relationship between possibility and reality (Inhelder & Piaget, 1958). In terms of hypothesis testing, the formal thinker would coordinate hypotheses to the level of possibility and data to the level of reality. Perhaps only late in the development of formal operations (in those who progress this far) is the subordination of reality to possibility sufficiently advanced for the individual to grasp the fundamental asymmetry of truth and falsity—the fact that hypotheses of the sort used in this experiment can be proven false by a single disconfirming instance, but cannot, in an infinite world, be conclusively proven true.

As Table 1 shows, subjects were fairly consistent in applying or not applying the

[4] The bimodal nature of the response distribution in this case probably violates the assumption of normality more than is allowable even for a test as robust as analysis of variance, and thus the level of significance for the linear age trend may be inflated. However, the reality of the age trend is hard to doubt: It is unlikely that eight of the nine consistently correct subjects would have been at the highest grade level due to chance alone ($p < .001$, binomial test).

nonverification insight. There were no main effects for task variables, nor were there any interactions with age.

## Conclusions

The results suggest three sequences of qualitative change in hypothesis-testing ability: (a) from no systematic interpretation of conditionals to implication interpretation, (b) from content-based information seeking to falsification strategy, and (c) from a symmetrical conception of truth and falsity to nonverification insight. Several promising lines for further investigation may be noted. First, longitudinal and clinical investigations would be useful in further exploring the details of these developmental trends and their interrelations. Second, from a practical point of view, it would be useful to study the relation of hypothesis-testing ability to students' level of comprehension of courses in the natural and social sciences. For example, lack of the nonverification insight might be associated with a tendency toward uncritical acceptance of scientific (and pseudoscientific) theories as fact. Third, it would be of both practical and theoretical import to see how and to what extent the development of hypothesis-testing ability may be facilitated. Finally, from a theoretical point of view, we need empirical and conceptual work relating hypothesis-testing ability to the emergence of such currently studied abilities as the control of variables, the design of factorial experiments, the abstraction and formulation of generalizations, and the comprehension of causal relationships, with the goal of ultimately generating a broad and powerful conception of the development of scientific reasoning.

## Reference Notes

1. Lunzer, E. A. *The Development of formal reasoning: Some recent experiments and their implications.* Paper presented at the second symposium of the IPN, Kiel, West Germany, March 1972.
2. Bady, R. *Logical reasoning abilities in male high-school science students.* Paper presented at the meeting of the National Association for Research in Science Teaching, Cincinnati, Ohio, March 1977.

## References

Beth, E. W. , & Piaget, J. *Mathematical epistemology and psychology.* Dordrecht, Netherlands: Reidel, 1966.

Bourne, L. E. , Jr. , & O'Banion, K. Conceptual rule learning and chronological age. *Developmental Psychology,* 1971, *5*, 525–534.

Bracewell, R. J. , & Hidi, S. E. The solution of an inferential problem as a function of stimulus materials. *Quarterly Journal of Experimental Psychology*, 1974, *26*, 480–488.

Cole, M. , & Scribner, S. *Culture and thought: A psychological introduction*. New York: Wiley, 1973.

Evans, J. St. B. T. Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology,* 1972, *24*, 193–199.

Gilhooley, K. J. , & Falconer, W. A. Concrete and abstract terms and relations in testing a rule. *Quarterly Journal of Experimental Psychology,* 1974, *26*, 355–359.

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

Johnson-Laird, P. N. , & Tagart, J. How implication is understood. *American Journal of Psychology,* 1969, *82*, 367–373.

Johnson-Laird, P. N. , & Wason, P. C. A theoretical analysis of insight into a reasoning task. *Cognitive Psychology,* 1970, *1*, 134–148.

Kodroff, J. K. , & Roberge, J. J. Developmental analysis of the conditional reasoning abilities of primary-grade children. *Developmental Psychology,* 1975, *11*, 21–28.

Kuhn, D. Conditional reasoning in children. *Developmental Psychology,* 1977, *13*, 342–353.

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

Neimark, E. D. Intellectual development during adolescence. In F. D. Horowitz (Ed. ), *Review of child development research* (Vol. 4). Chicago: University of Chicago Press, 1975.

Neimark, E. D. , & Chapman, R. H. Development of the comprehension of logical quantifiers. In R. J. Falmagne (Ed.), *Reasoning: Representation and process.* New York: Erlbaum, 1975.

Neimark, E. D. , & Santa, J. L. Thinking and concept attainment. In M. R. Rosenzweig & L. W. Porter (Eds. ), *Annual review of psychology* (Vol. 26). Palo Alto, Calif.: Annual Reviews, 1975.

Paris, S. G. Comprehension of language connectives and prepositional logical relationships. *Journal of Experimental Child Psychology,* 1973, *16*, 278–291.

Piaget, J. Intellectual development from adolescence to adulthood. *Human Development,* 1972, *15*, 1–12.

Roberge, J. J. Some effects of negation on adults' conditional reasoning abilities. *Psychological Reports,* 1971, *29*, 839–844.

Roberge, J. J. Effects of negation on adults' comprehension of fallacious conditional and disjunctive arguments. *Journal of General Psychology,* 1974, *91*, 287–293.

Staudenmayer, H. , & Bourne, L. E. , Jr. Learning to interpret conditional sentences: A developmental study. *Developmental Psychology,* 1977, *13*, 616–623.

Taplin, J. E. , Staudenmayer, H. , & Taddonio, J. L. Developmental changes in conditional reasoning: Linguistic or logical? *Journal of Experimental Child Psychology,* 1974, *17*, 360–373.

Ward, J. The saga of Butch and Slim. *British Journal of Educational Psychology*, 1972, *42*, 267–289.

Wason, P. C. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 1968, *20*, 273–281.

Wason, P. C. , & Johnson-Laird, P. N. *Psychology of reasoning.* Cambridge, Mass. : Harvard University Press, 1972.

Wildman, T. M. , & Fletcher, H. J. Developmental increases and decreases in solutions of conditional syllogism problems. *Developmental Psychology,* 1977, *13*, 630–636.