DOCUMENT RESUME

ED 139 618 SE 022 308

AUTHOR Bady, Richard J.
TITLE The Development of Hypothesis Testing and Correlational Reasoning.
PUB DATE Mar 77
NOTE 14p.; Paper presented at the annual meeting of the National Association for Research in Science Teaching (50th, Cincinnati, Ohio, March 22-24, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *College Students; *Correlation; Developmental Tasks; *Educational Research; Higher Education; *Hypothesis Testing; Learning Theories; Science Education; Secondary Education; *Secondary School Students
IDENTIFIERS *Piaget (Jean); Research Reports

ABSTRACT

Several tasks were developed to investigate adolescents' ability to see correlations in data and to test hypotheses. Emphasis was placed on methodology: do students clearly understand the problem and do the tasks tap the skills they claim to? The 20 ninth-grade and 20 eleventh-grade boys and 15 college freshmen were tested. Most students did rather poorly and many seemed to lack the abilities that may be necessary for meaningful learning in science. (Author/MH)

The Development of Hypothesis Testing and Correlational Reasoning.

Presented at N.A.R.S.T., Cinncinnati, Ohio, March 22, 1977

Richard J. Bady

Science Education Center
Graduate School of Education
Rutgers University
New Brunswick, N.J. 08903

ABSTRACT

Several tasks were developed to investigate adolescents' ability to see correlations in data and to test hypotheses. As much of the previous evidence is surprising and/or contradictory, emphasis is placed on methodology: Do students clearly understand the problem and do the tasks in fact tap the skills they claim to? Twenty ninth- and twenty eleventh-grade boys, and fifteen college freshman were tested. Most students did rather poorly, though better than in some previous studies, and many seemed to lack the abilities that may be necessary for meaningful learning in science.

Thanks are expressed to Janie Cheu for helping to develop and administer the tasks, and to Dave Kutliroff and Elliot Spack as well as the science faculty of New Brunswick High School.

INTRODUCTION

Understanding the logic of hypothesis testing and seeing correlations in data are two abilities we might all hope high school science students have achieved. Certainly no one could have any sense of what science is, or understand the concept of an hypothesis without understanding the logic of hypothesis testing--i.e. knowing that no single instance can prove the hypothesis no matter haw often found, and that in fact hopotheses of the form if. . .then. . . are not even proved at all but rather are accepted, provided they are never falsified. Likewise, the ability to detect a probabilistic relationship between two variables (i.e. correlation) is an ability which is surely needed for understaning phenomena in the biological and social sciences. These simple logical insights are straightforward enough for the science teacher, but are they for students of science?

The ability of students to deal with these problems has been explored in the psychological literature, but with ambiguous results. Wason and Johnson-Laird (1972) found that college students had a great deal of difficulty with his "four card task" in which they were asked to test an hypothesis of the form "if. . .then. . ." by selecting from various available data. He later found that certain improvements on the content and method of the task, making it more comprehensible to the subject, yielded better results. Our own pilot studies revealed that the task still is a bit unclear to many subjects, and poor performance may have been due in part to the difficulties in the task and not the logic.

Likewise, the results of studies on children's ability to deal with correlations have yielded no clear results. Inhelder and Piaget (1958) designed a task to tap this ability and found that by 15 years most children were able to see correlations and make judgements about the strength of the relationship. Smedslund (1963), however, found that adults (student nurses) could not solve his correlations problem. But it seems likely to this author that the task, like Wason's, may be a bit obscure, and it could have been a failure to understand what was asked. Other investigators, using tasks more like Piaget's, found high school age children did well on the tasks (Martarano, 1974, Kuhn 1976), while still others obtained results somewhere in between (Seggie 1975, Neimark, 1975). These discrepancies may be due to differences in the task situation and questions used. Little attention has been paid to the logical aspects of their task requirements and to how the subjects perceive the task.

In the present study various changes were made to increase the likelihood that the questions used tap the logical structures being tested, and that the students understood what they were being asked.

SUBJECTS

The subjects were 55 male students: twenty from 9th- and twenty from 11th-grade scuence classes at a large urban high school, and fifteen freshman from a psychology course at a nearby community college. The mean California Achievement Test total score for the ninth grade students was 85.3 (SD=11.8), and for the eleventh graders (who took the test in ninth grade), 87.3 (SD=16.4). Thus each group was about average, 90.0 being the expected score at the end of ninth grade, and the two groups were comparable in standardized skills achievement.

TASKS

Three tasks were given: (a) the "turtles task", a modification of Wason's hypothesis testing task, (b) the "eye color-hair color problem", a modification of Inhelder and Piaget's correlations problem, and (c) the "bats task," a newly constructed variant of the eye color-hair color problem, designed to tap abilities similar to those needed in the correlations problem.

In the turtles task the student told a story and asked to make a decision about a hypothesis. The task is very similar in structure to Wason's original task, but several changes have been made to facilitate performance: (a) the content is more relaistic and meaningful, (b) the possibility that the hypothesis might be wrong is made obvious, and (c) the subject is presented with a example of a falsifying instance, among others before he makes his final choice, to be sure he has not simply overlooked it. It is possible to determine from the subject's responses. as well as from his explanations, whether he tests the hypothesis by trying to find confirming instance or by trying to find disconfirming instances, the latter being the only logically adequate way to test the hypothesis.

The eye color-hair color problem was adapted from Inhelder and Piaget's original correlations problem. The most important change made was the introduction of a game in which the child was asked to solve a problem, rather than simply asking him to detect a "relation". It was found in pilot work that most children were unable to respond to the notion of a relation, but were quite able to respond, though not necessarily adequately, to the game. The task involved seeing that for a sample of people, hair color and eye color are correlated, that is, each of the two possible hair colors tends to be associated with one of the two possible eye colors and vice-Versa. The notion is slightly different, of course, from the statistical notion of correlation, but is not unrelated. The relations is better described as a probabilistic biconditional (see appendix). It is possible to tell from the subjects responses and explanations whether he is capable of grasping this kind of relationship, and how he deals with this type of data.

The bats task was developed to provide a situation which is similar, from a logical standpoint, to the eye color--hair color problem, but with a content that tends to focus attention on certain aspects of the data. In addition the problem is highly pragmatic, and the student can easily make a clear dicision from the data. There is no need to play a hypothetical game or use the troublesome word " "relation". The subject is presented with won-loss records for two different kinds of baseball bats used by a team in order to determine if one bat is better than the other. Thus, as before, two binary variables are correlated.

It was seens as a result of pilot work that the manner in which the data is presented and the questions that are asked are crucial. Complete details of how to present the task and the problems with the earlier versions are found in The Appendix.

METHOD

All three tasks were given, in different orders, and by two experimenters, in a clinical (one-to-one) situation. Analyses of variance showed no significant effect of order of presentation of tasks or for experimenter. The testing lasted about 40 minutes per subject and was tape recorded for later help in scoring.

For the two correlations tasks (eye color-hair color, and the bats), the criterion of whether the subject used all the data in a logical fashion to draw his conclusion was used to give a success/failure classification. It was not required that he demonstrated an understanding of the biconditional, and indeed few subjects managed this. In addition, points were given based on several of the questions asked to give a possible score of 0-5 for the eye-color-hair color task, and 0-4 for the bats task. The score gives a rough indication of the extent to which the subject can deal with the various logical requirements of the tasks.

The hypothesis testing task allowed most of the students to be labeled as seeking to verify only, falsify only, or both verify and falsify. In addition points were given for the various aspects of understanding of the problem, yielding a score of 0-3.

Details of the tasks and scoring procedures are to be found in the Appendix.

RESULTS

Analyses of variance on the scores given for each task across age showed significant differences, but relatively low scores in general (see table 1).

TABLE 1

MEAN SCORES for each task as a function of age

	Eye color-hair color	Bats	Turtles
Comm. Col. Freshman (n=15)	3.53	2.33	2.20
11th grade (n=20)	2.85	1.85	1.84
9th grade (n=20)	1.45	1.50	0.75
F(df=2/52)	11.2	3.49	7.60
P	.01	.05	.01

Relatively few students were able to demonstrate ability to deal logically with the data in the correlations tasks, as show in table 2.

TABLE 2

PERCENT-SUCCESS ON CORRELATIONS TASKS AS A FUNCTION OF AGE

	Eye color-hair color	Bats
Comm. col. Freshman (n=15)	40%	27%
11th grade (n=20)	15%	30%
9th grade	5%	10%
Total	18%	22%

Performance on the turtles task was also surprisingly poor, as can be seen in table 3.

TABLE 3

PERCENT AGE OF SS IN EACH RESPONSE CATEGORY ON TURLES TASK AS A FUNCTION OF AGE

	Falsifying	Verifying & Falsifying	Verifying	Other
Comm. Col. Freshman (n=15)	27%	27%	27%	19%
11th grade (n=20)	20%	10%	40%	30%
9th grade (n=20)	0%	10%	45%	45%
Total	15%	15%	38%	32%

DISCUSSION

The above results indicate that the abilities to deal with these problems are developing over the age range tested, but it is clear that high school science teachers cannot assume that their students have these abilities. The inability of the student to deal with correlated data even in the case of the bats, where the confirming and disconfirming cases might seem more obvious indicates that the students may be far from having developed these schemes and that their laboratory experiences, for instance, may be quite different from what we might have imagined. Most subjects found some way to deal with the data and arrive at an answer - sometimes, fortuitously, the right answer- but for the wrong reasons. Here is the beauty of the clinical method: Without the interaction with the student, one really doesn't know much about how the student is thinking. Answers alone don't tell the whole story.

It may not be surprising that performance on these tasks was so poor; after all, correlation is a difficult concept, and one that is probably not encountered even in science classes until late in high school (if at all). Of course, every science student, in one of his first science courses, learns (at least by rote) what a hypothesis is and how science ptoceeds by empirical test. But how often is the student really given a hypothesis to test, as in our task? Probably not often, except perhaps in a well-run inquiry-oriented program.

The concept learning literature in psychology indicates that in some cases people can learn concepts more easily from positive (confirming) instances than from negative (disconfirming) ones. (Bourne 1970). One can conclude from this and the results of this study that the basic notions involved in testing scientific theories is more difficult for many students than many of us may have suspected. What kind of an understanding of science could a student have who thinks that you test a hypothesis by looking for confirming cases only?

The typical high school science course focuses on content, perhaps at the expense of the development of the logical thought processes that are required for a real understanding of the meaning of any data. A rote knowledge of scientific "facts" is probably not very useful without the logical apparatus to understand how they were arrived at, and how they may be empirically tested.

APPENDIX

THE TURTLES TASK

The subject is told to pretend that he is an expert on turtles and that the particular kind of turtle he is studying has either diamonds or circles on its back, and that its stomach is either red or green. He is shown several cards depicting turtles- on the side of each card is the top view of a turtle (with either diamond or circle markings) and on the other side of the card is bottom view of the turtle (the bottom of the shell is either red or green). Thus the content is (almost) realistic, not symbolic and arbitrary as in Wason's task. The subject is them told that these turtles are being studies on a particular island and that one biologist on the island has claimed "All the turtles with diamonds on their backs, have green bottoms." This replaced the more troublesome "if . . .then . . ." formulation that Wason originally used. The hypothesis, like Wason's, asserts a connection that is purely empirically determined. There is no (apparent) casual connection to be inferred. This type of statement was used to avoid the complication of the subjects' reasoning about some casual or theoretical link, rather than focussing on the data only. The subject is also told that a second biologist has asserted "That is not true," in response to the statement of the first biologist. This is to assure that the subject always remembers that he is not to assume that the hypothesis is true. The subject is given a card on, which the two statements are written and is told that he is to decide who is right from examining the turtles on the island.

A bit of logical symbology is appropriate here. The statement can be symbolized p - q, where "p" represents having diamonds on the back, and "q" represents having a green bottom. Having circles on the back is symbolized "p", or "not p", and having a red bottom is written "q". The four possible kinds of turtles and the conclusion that could be drawn about the biologist#1's statement after seeing just one of those turtles, is given in table 1.

TABLE 1

TURTLE			
Back	Bottom	Symbolically	Conclusion about rule after seeing one.
Diamonds	Green	$p.q$	Can't tell
Diamonds	Red	$p.\overline{q}$	Proves #1 wrong
Circles	Green	$\overline{p}.q$	Can't tell
Circles	Red	$\overline{p}.\overline{q}$	Can't tell

Notice that although the p.q instance might be viewed as confirming the hypothesis, seeing only one does not prove anything. However, seeing just once instance on $p.\bar{q}$ proves that biolgist #1 is wrong and #2 is right.

After having been presented with the statements of the two biologists, the subject is handed one at a time, the four possible kinds of turtles, and asked, for each on separately, what he would conclude, if anything, as to which biologist is right given that this turtle was the first one he happened to pick up upon arriving at the island. He is asked if he could decide for sure which biologist is right, It is made clear to the subject that each turtle is to be considered separately, and that there are other turtles on the island. This procedure (the instance evaluation) allows the subjects interpetation of the statement to be seen, as well as whether the subject realizes that no instance by itself can prove #1 right. For instance, some subjects, especially younger ones, think that since all turtles with diamond must have green bottoms that all turtles with green bottoms must have diamonds as well. This is the fallacy of assuming the converse. Since it is well known that this is a common fallacy and since it was thought that it might be an unrelated interference to the task, it was decided that during the instance evaluation, subjects who made this error would be corrected by simply asking -- "But what does biologist #1 say about turtles with green bottoms?" Indeed this helped a few subjects, but interestingly, there were also some who a few seconds later would resume committing this fallacy.

At this point the subject has seen the hypothesis, been told he should test it, and been exposed to the four possible turtles, including the one that could prove the hypothesis false. He is then presented with a group of turtles and told that fortunately all the turtles on the island liked to swim in a big pond and that they are all here. Thus the subject is assured that all the turtles are available to be checked so that there is no problem about deciding about the hypothesis for sure. The subject is also told to forget the previously seen turtles., as it was noticed in pilot work that some subjects tried to reason on the basis of the earlier instances.

It is pointed out to the subject that some of the turtles in the pond are floating on their backs, and thus it can be seen that the bottoms are either red or green but that they might have diamonds or circles on their backs, and one can't tell without turning them over. Further, the rest of the turtles are floating on their backs, and thus it can be seen that they have either diamonds or circles on their backs but that one cannot tell if they have red or green stomachs. The question then posed to the subject is: "Which of the turtles will you have to turn over, in order to find out for sure which biologist is right?" The subject makes his selection and is asked to explain it. The subject with insight into hypothesis testing will want to turn over the diamond backed turtles and the red bottomed turtles since these are the ones that could reveal a falsifying instance ($p.\bar{q}$). The most popular incorrect answer is to try to verify the rule by turning over the diamond backed turtles and the green bottomed turtles in search of verifying (p.q) instances.

From the subject's choice and explanation it is possible to determine his hypothesis testing strategy. Besides the verifying and falsifying strategies just mentioned, some subjects wanted to do both, turning over the diamond-backed, the red and the green bottomed turtles in search of both verifying and falsifying instances. Some subjects will focus on the content and not the logic, and base their choice on some knowledge of turtles. Others were not able to make us understand what their strategy was. These two responses were classified as a separate group.

At this point, in order to be sure the subject has not made a simple mistake, he is given a second change (assuming his first answer was incorrect). He is asked "Suppose we turned this turtle and found ____________". He is given several possibilities including turning over the turtles with red bottoms and finding diamonds on their backs. After several instances to determine that he has considered the various possibilities and their implications, the subject is asked if he would like to make a different selection.

Thus any subject who was classified as a verifier, did so after being shown an instance that would falsify the hypothesis. Much to our surprise quite a few subjects, seconds after telling us that turning over a red bottomed turtle and seeing diamonds would prove #1 wrong, still insisted that to test the hypothesis you need only look for verifying instances.

It is our feeling that these various manipulations of the task present a fair indicator of the subject's reasoning. It is interesting to note that, despite the poor performance, the subjects did better than in some of Wason's work.

The subjects' interpretation of the statement and his understanding as to whether one instance can prove a rule true or not, is interesting but was not specifically studied here. Also the relationship of these notions to the hypothesis testing strategy deserve further study.

To obtain a more useful measure of the subject task performance, besides the classification of the hypothesis testing strategy, a score was given to each subject. One point was given for realizing (on the instance evaluation) that no instance could prove the rule; one point was given for using a falsification strategy in testing the hypothesis (even if verification was used as well) and an extra point given if falsification only was used. Thus the hypothesis testing score ranged from 0 to 3.

THE HAIR COLOR-EYE COLOR TASK

The subject is presented with a deck of cards, and on each card is a simply drawn face with either orange or green hair and either purple or yellow eyes. Thus making four "kinds" of people. There are four decks used throughout the task, and the number of each kind of person in each deck is given in Table II

TABLE II

Hair color: Eye color:	Orange Purple	Orange Yellow	Green Purple	Green Yellow
Deck I	5	2	2	5

TABLE II (cont'd).

Hair color: Eye color:	Orange Purple	Orange Yellow	Green Purple	Green Yellow
Deck II	6	1	1	6
Deck III	6	1	3	4
Deck IV	10	4	4	10

The subject is told to imagine that he has arrived on a distant planet and is going to be asked questions about the people there. The eye and hair colors are pointed out to him and he asked to find how many different kinds of people there are. An extraterrestrial world and four different and usual hair and eye colors were used to insure that the subject does not apply any preconceived notions to the task and that not matching strategies (brown hair goes with brown eyes) are available. Pilot testing showed these to be occasionally troublesome.

The subject is first asked if he can see a relationship between hair color and eyecolor. Pilot testing showed most people had trouble with understanding what this meant, so the following story line was developed. The subject is told that we are going to play a game and for a randomly selected person from the deck, he is to predict the hair color (or eye color), having seen only the eye color (or hair color).. All four examples are done, e.g.,: suppose you saw someone with orange hair, what color eyes would you predict that he has? The subject is questioned until (if he is able) he can make the best prediction in each case. What is of interest is whether the subject can see the relationship as a biconditional, that is, if you have orange hair, you have purple eyes and vice versa, and that therefore if you have green hair, you have yellow eyes and vice versa. However, this relationship holds only probabilistically. Thus eye color and hair color are correlated.

Several lines of questioning were adopted from Inhelder and Piaget to tap the subjects ability to see this relationship. First the subject is asked to make a deck so that he could always predict correctly and a deck for which he would not be able to predict at all. It was noted whether the subject responded by making a deck with a perfect correlation (e.g.: 3,0,0,3) and one with a zero correlation (e.g.: 2,2,2,2).

The next line of questioning involved comparing decks to judge which was the higher correlation. The subject is presented with deck II (see Table I), and asked if he would rather play the game on this "planet" or on the first one. Both decks are left for the subject to manipulate and classify as he pleases. Deck II was used simply to assure that the subject understood what is being asked. Most subjects even with the more primitive strategies, are able to see that deck II is better, although for a variety of reasons depending on their own logic. The subject is asked to compare Deck I to Deck III and to Deck IV in turn and to explain his choice. Notice that Deck III has a total of ten cases that fit the rule (6 + 4) and four cases that don't (1+3). Further, notice that Deck I also has ten cases that fit the rule (5+5) and four that don't (2+2). Thus in either case, the chances of being right are 10/14, as long as people are selected randomly and the variable to be predicted (hair or eye color) is also caried, as specified in the game. Both decks have the same correlation. The subject is asked if one deck would be better to play with than the other, of if they're both about the same. Seeing that they are the same requires both an ability to deal with the relationship plus a notion of probability. The fourth deck was also compared to the first. Note that though it has twice as many cases that fit the rule, it also has twice as many that don't, thus making the correlation the same.

It was found that ome subject do not express an understanding of the biconditional, but rather treat the relationship as consisting of two separate rules, eg: one for predicting hair color and one for predicting eyecolor. This method is logically adequate, but makes deck comparisons cumbersome. Thus an adequate method may still allow the subject to make mistakes in comparing decks.

Subjects were scored success/fail on the basis of whether they used all the data in a logical fashion to compare decks. Thus the biconditional solution as well as the two rule mthod described above were scored as successes, whether or not the subject compared decks correctly. A subject who, for instance, compared decks by looking at people with orange hair only, was scored as failing.

A score for the task ranging from 0-5 was determined by giving one point for the correct construction of perfectly correlated and non-correlated decks; one point was given for each of the two latter deck comparison; one point was added if the subject used all the information to compare decks (as in the precious scoring classification), and an additional point was added if he expressed the rule as one rule (the biconditional) rather tha.. two separate rules.

BATS

The bats task was designed to be very similar to the hair color-eye color task. In fact the questions asked and the scoring procedures used are identical except that since there was only one deck comparison made, the score ranged from 0-4. In addition, the subject was rated success/failure depending on whether he used all the data to make his comparison.

In this task the subject is told that someone has designe d a new kind of baseball bat and that in order to test it he has asked a team to play some games with both bats to see what happens. The subject is presented with a deck of cards. Each card contains a picture of a bat and the words "standard bat" or "test bat", plus the word "win" or "lose". Each card represents a game in which the team used the bat indicated and either won or lost that game. Two decks were used (see Table III)

TABLE III

	Test Bat		Standard Bat	
	Won	Lost	Won	Lost
Deck I	[illegible]	[illegible]	[illegible]	[illegible]
Deck II	[illegible]	[illegible]	[illegible]	[illegible]

The subject is given Deck I and asked to decide if he thinks one bat is better than the other. He is asked to make a deck that would definitely show that the bat chosen makes a difference, (analogous to the perfect and zero correlation questions in the eye color-hair color task).

Some subjects were found to try to invoke their knowledge of baseball and were concerned over the fact that other factors may be important as well. They were assured that all other factors (the pitcher, the weather etc.) were equal.

The subject is then asked to compare Deck I (one team) to Deck II (another team), and to decide whether which bat that is used makes a bigger difference for one team, or if the difference is about the same. The task is not identical with the previous one, because the structure of the data does not make the biconditional solution seem as appropriate as comparing the test bat's record directly with that of the standard bat. None the less, many of the logical structures involved are the same. It maybe fruitful to pursue in detail the differences in these two tasks and we have already begun some of this work.

REFERENCES

Inhelder, B. and Piaget, J. *The Growth of Logical Thinking From Childhood to Adolescence.* New York: Basic Books, 1958.

Kuhn, Deanna. Relation of two Piagetian stage transitions to I.Q. *Developmental Psychology*, 1976, 2, 157-161.

Martarano, S. The development of formal operational thought. Unpublished Ph.D. dissertation, Rutgers University, 1974.

Neimark, E.D. Longitudinal Development of formal operational thought. *Genetic Psychology Monographs*, 1975, 91, (2), 171-226.

Seggie, J.L. The empricial observation of the Piagetian concept of correlation. *Canadian Journal of Psychology*, 1975, 29, 32-42.

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