DOCUMENT SUMMARY
This document is a verbatim transcription of P. C. Wason's 1960 experimental paper "On the Failure to Eliminate Hypotheses in a Conceptual Task." It reports a concept-attainment task in which undergraduates tested rules by generating number triads, revealing a strong tendency to rely on confirming evidence (enumerative induction) rather than seeking falsifying evidence (eliminative induction). Key findings include that only subjects who generated sufficient negative instances attained the correct rule, underscoring the importance of hypothesis-falsification in scientific reasoning.

FILENAME

wason_1960_research_paper_failure_to_eliminate_hypotheses_conceptual_task

METADATA
Category: research
Type: report
Relevance: Reference
Update Frequency: Static
Tags: #eliminative_induction #enumerative_induction #concept_attainment #hypothesis_testing #experimental_psychology #wason_1960 #scientific_reasoning #disconfirming_evidence #falsification #cognitive_experiments
Related Docs: []
Supersedes: []

FORMATTED CONTENT

On the Failure to Eliminate Hypotheses in a Conceptual Task

P. C. Wason
From the Medical Research Council Industrial Psychology Research Group, University College, London

This investigation examines the extent to which intelligent young adults seek (i) confirming evidence alone (enumerative induction) or (ii) confirming and disconfirming evidence (eliminative induction), in order to draw conclusions in a simple conceptual task. The experiment is designed so that use of confirming evidence alone will almost certainly lead to erroneous conclusions because (i) the correct concept is entailed by many more obvious ones, and (ii) the universe of possible instances (numbers) is infinite.

Six out of 29 subjects reached the correct conclusion without previous incorrect ones, 13 reached one incorrect conclusion, nine reached two or more incorrect conclusions, and one reached no conclusion. The results showed that those subjects who reached two or more incorrect conclusions were unable, or unwilling to test their hypotheses. The implications are discussed in relation to scientific thinking.

INTRODUCTION

Inferences from confirming evidence (Bacon's "induction by simple enumeration") can obviously lead to wrong conclusions because different hypotheses may be compatible with the same data. In their crudest form such inferences are apparent in the selection of facts to justify prejudices. In research merely confirming evidence is clearly of limited value. For example, suppose that it is suggested that a deficit (x) of a particular substance in the blood is uniquely related to a distinctive symptom (y). (In logical terms, x is a "necessary-and-sufficient" condition for y.) And suppose that this hypothesis had been supported by confirming evidence alone, i.e. whenever the deficit had been induced, the symptom had appeared (ignoring statistical issues for the sake of simplification). It might then be assumed that the hypothesis was tenable. But the evidence only allows the inference that x is a sufficient condition for y, that the deficit always leads to the symptom. To establish the postulated relation, there must also be no disconfirming evidence—no case of the symptom without the deficit. For if such a case were obtained, the deficit would not be a necessary condition for the symptom, and the symptom would not be a sufficient condition for the deficit. The hypothesis that either was a necessary-and-sufficient condition for the other could be eliminated. The symptom would not be a reliable sign. Its absence would rule out the possibility of the deficit, but its presence would be ambiguous.

In general, scientific inferences are based on the principle of eliminating hypotheses, while provisionally accepting only those which remain. Methodologically, such eliminative induction implies adequate controls so that both positive and negative experimental results give information about the possible determinants of a phenomenon.

This investigation seeks to determine the extent to which intelligent young adults make rational inferences about abstract material which does not obviously suggest a conventional scientific approach. A concept attainment task would seem to be most suitable for this purpose, and one such typical task will now be considered in detail.

Bruner, Goodnow and Austin (1956) report some ingenious experiments of this kind. The material of one experiment consisted of an array of 81 instances made up from all the combinations of four attributes, each exhibiting one of three values, i.e. shape (cross, square, circle), colour (green, red, black), number of figures (one, two, three) and number of borders (one, two, three). The subject's task was to attain a concept, defined in terms of the values of these attributes, by choosing successive instances in order to find out whether they exemplified the concept. The authors describe a number of possible strategies which can be used. One of these, "successive scanning," is equivalent to induction by simple enumeration—it involves testing successive hypotheses by using information derived from positive instances alone. For example, if "green squares" is the subject's hypothesis, instances exhibiting these features are selected. If these turn out to be positive, "green squares" is considered to be necessary-and-sufficient for the concept. But if any of these instances are negative, a different hypothesis is similarly tried out, and accepted so long as it leads to positive instances.

However, Bruner et al. are not concerned with the fact that successive scanning can lead to merely sufficient, as opposed to necessary-and-sufficient, concepts. For example, if the correct concept is "red figures" (of which there are 27 positive instances), it is possible to attain the incorrect concept, "red circles" (of which there are 9 positive instances), by consistent use of confirming evidence, since all positive instances of this latter concept are also positive instances of the correct one. The incorrect concept, in this case, entails the correct one but is not entailed by it, i.e. "red circles" is a sufficient, but not a necessary-and-sufficient, condition for "red

figures." Thus it is logically possible to arrive at incorrect concepts by successive scanning because only confirming evidence is utilized. Such a result, however, may not occur often because other instances, which might act as reminders of alternative possibilities, are displayed in front of the subject. In the above example, the subject's attention might be directed to the possibility of "circles" being necessary, owing to the presence of instances exhibiting red squares and crosses.

The present investigation is designed to compel the subject to encounter plausible sufficient conditions of the concept which must ultimately be eliminated in order to attain the necessary-and-sufficient conditions. The logical mechanism underlying efficient performance in the task rests upon what Von Wright (1951) terms "the fundamental, though trivial, fact that no confirming instance of a law is a verifying instance, but that any disconfirming instance is a falsifying instance."

In the experiment the concept to be attained is "three numbers in increasing order of magnitude."

Subjects were told that the three numbers 2, 4, 6 conformed to a simple relational rule and that their task was to discover it by making up successive sets of three numbers, using information given after each set to the effect that the numbers conformed, or did not conform, to the rule.

It will be seen that this task, as a whole, differs from previous studies of concept attainment. In the first place, the attributes are rules referring to relations between numbers, e.g. "consecutive even numbers." In this respect, the possible concepts are more like those found in the Goldstein-Scheerer (1941) test of abstractive ability, than those which can be attained in Bruner's experiment, in which attributes cannot be combined into classes designated by names but have to be enumerated, e.g. "three red circles."

Secondly, the possible instances (triads of numbers) are, in principle, infinite. In previous studies a finite universe of instances has been used. Such a universe, however, places great constraint on the scope of inductive thinking because the number of instances which exemplify any sufficient concept will always be less than the number which exemplify any necessary-and-sufficient one. Hence the number of instances of a given kind which can be tested is limited. But in the present task an endless series of instances exemplifying a sufficient rule can be generated without forcing the subject to encounter an instance which would not exemplify it.

Thirdly, the instances are not presented as stimuli, but have to be generated by the subject. In this way he is completely free to decide on the kind and amount of evidence which he considers adequate. If, on the other hand, all the possible instances are displayed simultaneously, the subject will know that all available evidence for the solution of the problem is already present.

Finally, no memory of previous instances is involved. The subject keeps a record of successive instances, the reasons why he generated them and their outcome. The introduction of memory into concept attainment studies is an interesting, but essentially gratuitous variable which may distort inductive reasoning by placing an additional burden on the subject.

FIGURE I

Record sheet

PROCEDURE

The subjects, 29 psychology undergraduates (17 men and 12 women), were examined individually and instructed as follows:

> "You will be given three numbers which conform to a simple rule that I have in mind. This rule is concerned with a relation between any three numbers and not with their absolute magnitude, i.e. it is not a rule like all numbers above (or below) 50, etc.
>
> Your aim is to discover this rule by writing down sets of three numbers, together with reasons for your choice of them. After you have written down each set, I shall tell you whether your numbers conform to the rule or not, and you can make a note of this outcome on the record sheet provided. There is no time limit but you should try to discover this rule by citing the minimum sets of numbers.
>
> Remember that your aim is not simply to find numbers which conform to the rule, but to discover the rule itself. When you feel highly confident that you have discovered it, and not before, you are to write it down and tell me what it is. Have you any questions?"

Subjects then wrote down their first set of numbers, under the numbers 2, 4, 6, on the record sheet, together with the reasons why they had chosen them. The Experimenter then said "those numbers do conform to the rule," or "those numbers do not conform to the rule," according to whether they were in increasing order of magnitude. The second set of numbers was then written down by the subject, the Experimenter giving the appropriate information as before. This procedure continued until the subject wrote down a rule. If the rule was incorrect, the subject was told so and instructed to carry on as before. The experiment continued until the correct rule was announced, or the time for the session exceeded 45 minutes, or the subject expressed a wish to give up. The time was recorded and the implications of the results discussed. Finally, subjects were warned not to talk about the experiment.

RESULTS

Quantitative

Results will be classified as follows:

1. **Frequency of correct and incorrect rules**
2. **Extent of enumerative and eliminative thinking**
3. **Frequency of negative instances**
4. **Immediate response to incorrect rules**
5. **Types of incorrect rule**

(a) Frequency of correct and incorrect rules

Table I shows the frequency of successive announcements of rules. The first announcement can either be correct (defined as the “immediate correct announcement”) or incorrect. (In nearly all cases incorrect rules were sufficient ones, e.g. “increasing intervals of two.”) A third category, “none,” covers those subjects who made no announcement of a rule of any kind throughout the experiment. Those cases which fall into the “incorrect” category at the first announcement are redistributed within the second announcement categories, according to whether the second rule announced is correct or incorrect, with “none” included for those who made no subsequent announcement. Similarly for third announcements. Thus, the table provides a running record of the behaviour of the subjects.

In the sample there were 11 undergraduates in their first year, 12 in their second year and six in their third year. Of the six subjects who made the immediate correct announcement, four were in their second year and two in their third year. There were no sex differences and no differences between Arts and Science background.

Table I. Frequency of Announcements (n = 29)

Announcement Number	Correct	Incorrect	None
1st	6	22	1
2nd	10	12	7
3rd	4	5	20
4th	1	0	28
5th	0	0	29

(b) Extent of enumerative and eliminative thinking

An index of the kind of thinking used was constructed by considering the nature of the instances in relation to the reasons given by the subject for choosing them. Each “reason for choice” was classified as either compatible or incompatible with all subsequent instances (including that instance for which it was given). The total number of instances, compatible and incompatible with reasons, was computed for each subject (i) up to the first rule announced, and independently (ii) from the first rule to the second one (if any). It was assumed that thinking was enumerative if there were a high proportion of compatible instances, and eliminative if there were a high proportion of incompatible ones.

The ratio of the number of incompatible to compatible instances provided an eliminative/enumerative index for each subject. The mean ratio was 1.79 (n = 6) for those who made the immediate correct announcement, and 0.24 (n = 22) for those who made a first incorrect announcement. A significant difference was obtained between these two means ($p = 0.0002$, one-tail test). The statistic used was Whitfield’s (1947) extension of Kendall’s S (1948) to a dichotomous variable.

The mean number of instances generated before making the immediate correct announcement was 8.0 (range 5 to 9), and the mean number before making a first incorrect announcement was 3.68 (range 1 to 7). This difference is highly significant. Thus, subjects who arrived at merely

sufficient rules did so on the basis of relatively few confirming instances, while those who attained the necessary-and-sufficient rule tended to eliminate sufficient ones.

For those subjects whose second announcement was (i) correct, and (ii) incorrect, the mean ratios were 0.50 and 0.19 respectively. The difference between them was just short of statistical significance ($p = 0.08$, one-tail test).

It should be noted that in 18 out of the 22 cases of a first incorrect rule, the reason given for the subject's first instance was subsequently announced as at least a part of their first rule.

Figure 2 shows Vincent curves (Hilgard, 1938) of the mean number of incompatible instances (per fifths of the total number of instances for each subject) for (i) the immediate correct announcement, (ii) the first incorrect announcement, and (iii) the first incorrect announcement omitting 15 subjects without any incompatible instances.

(c) Frequency of negative instances

The ratio of the number of negative instances of the correct rule to the total number of instances generated was computed for each subject, (i) up to the first rule announced, and (ii) from the first rule to the second one (if any). The mean ratio was 0.21 ($n = 6$) for those who made the immediate correct announcement and 0.04 ($n = 22$) for those who made a first incorrect announcement. A significant difference was obtained between these two means ($p = 0.00002$, one-tail test).

For those subjects whose second announcement was (i) correct, and (ii) incorrect, the mean ratios were 0.38 and 0.09 respectively, and the difference between them was significant ($p = 0.003$, one-tail test).

A highly significant correlation was obtained between the eliminative/enumerative index and the negative instance index ($\tau = +0.72$, $p < 0.00003$, one-tail test). It should be noted that a negative instance of the correct rule does not suffice to eliminate a sufficient rule. For example, a negative instance such as 6, 4, 2 is compatible with the sufficient rule, "intervals of two in increasing magnitude." A sufficient rule can only be eliminated by a positive instance which is incompatible with it.

(d) Immediate response to incorrect rules

The instance which is generated immediately after an incorrect announcement can either be compatible or incompatible with the rule announced, and it can be either a positive or a negative instance of the correct rule.

Table II. Frequency of Types of Instance Immediately Following Incorrect Rules

	Compatible with Incorrect Rule	Incompatible with Incorrect Rule	Total
Positive instance of correct rule	5	11	16
Negative instance of	2	13	15

correct rule			
Total	7	24	31

On a purely logical criterion it would be expected that, when subjects knew that their rule was incorrect (by being told so), they would depart from it and try a new one. But it will be seen from the table that in more than half the cases the rule is maintained, even though some other attribute, e.g. order, may be tested.

(e) Types of incorrect rule

Four different kinds of incorrect rule were announced most frequently:

1. Numbers increasing in intervals of 2, i.e. a, (a + d), (a + 2d) where d = 2
2. Increasing multiples of the first number, i.e. a, (a + d), (a + 2d) where a = d
3. Consecutive even numbers, i.e. 2a, (2a + d), (2a + 2d) where d = 2
4. Arithmetic progression, i.e. a, (a + d), (a + 2d)

In addition, five rules were announced only once: (i) “the first number added to the second number gives the third”; (ii) “numbers which add up to twelve”; (iii) “ascending progression formed by adding or multiplying by a constant”; (iv) “arithmetic or geometric progression”; and (v) “the second number is the first number plus one and the third is the first number plus four.” All these rules, except the last one, are consistent with the initially given instance.

Table III. Frequency of Incorrect Rules at Successive Announcements

Rule	1st Announcement	2nd Announcement	3rd Announcement	4th	5th
Increasing intervals of two	14	1	0	0	0
Multiples of first number	8	2	0	0	0
Consecutive even numbers	7	0	0	0	0
Arithmetic progression	6	12	1	0	0
Others	4	1	0	0	0
Correct rule	6	10	4	1	0

Qualitative

Six protocols are given below, the first three from subjects who made the immediate correct announcement, and the second three from those who made incorrect announcements.

The most interesting qualitative feature of the results is that the successive announcements of three subjects consisted in the repetition of a single rule in different terms from those used on their previous announcements. In addition, three subjects reformulated their rules immediately after being told that they were incorrect, i.e. before generating any further instances. (This latter type of repetition was not counted as a new announcement in the quantitative results.)

In protocol No. 4 it will be seen that the second rule reads, “the middle number is the arithmetic mean of the other two,” and the third reads, “the difference between two numbers next to each other is the same.” These rules do not mean exactly the same thing. The first can be fulfilled by instances such as 5, 7, 9 and 9, 7, 5; the second by 5, 7, 9; 9, 7, 5 and 7, 7, 7. But the fourth rule, “adding a number, always the same one to form the next number,” which follows the negative instance 12, 8, 4, is restricted in scope so that it could not be fulfilled by some instances which could fulfil the two previous rules. Its connotation is now exactly that of arithmetic progression. It is not clear whether this subject (i) appreciated the fine difference between the second and third rules, in spite of the fact that their instances conformed to arithmetic progression, or (ii) thought that there were no differences between the rules, but assumed that their expression was relevant, or (iii) thought that the rules were completely different, failing to realize that arithmetic progression was their common factor.

Below are examples of protocols:

A. Immediate correct announcement

No. 1. Female, aged 25, 3rd year undergraduate
12 24 36: even figures are even and increase in twos
8 10 12: even numbers increasing in twos
2 6 10: even numbers increasing in fours
6 4 2: even numbers decreasing in twos
2 6 8: even numbers ascending
8 14 20: even numbers ascending
1 17 23: ascending numbers
1 18 23: ascending numbers
1 2 3: ascending numbers

The rule is ascending numbers (9 minutes).

No. 2. Female, aged 21, 2nd year undergraduate
3 6 9: three goes into the second figure twice and into the third figure three times
2 4 8: perhaps the figures have to have an L.C.D.
2 4 10: same reason
2 1 10: the second number does not have to be divided by the first one
10 6 4: the highest number must go last
4 6 10: the first number must be the lowest
2 3 4: it is only the order that counts
4 5 6: same reason
1 7 13: same reason

The rule is that the figures must be in numerical order (16 minutes).

No. 3. Male, aged 25, 2nd year undergraduate
8 10 12: continuous series of even numbers
14 16 18: continuous series of even numbers
20 22 24: continuous series of even numbers
3 5 7: continuous series of odd numbers
1 2 3: continuous series but with smaller intervals
3 2 1: reverse
2 4 8: doubling series
2 2 4: two numbers the same
6 4 2: reverse of original numbers
1 9 13: simple ascending numbers

The rule is any ascending series of different numbers. (10 minutes).

B. Incorrect announcements

No. 4. Female, aged 19, 1st year undergraduate
8 10 12: two added each time
14 16 18: even numbers in order of magnitude
20 22 24: same reason
1 3 5: two added to preceding number

The rule is that by starting with any number two is added each time to form the next number.
2 6 10: middle number is the arithmetic mean of the other two
1 50 99: same reason

The rule is that the middle number is the arithmetic mean of the other two.
3 10 17: same number, seven, added each time
0 3 6: three added each time

The rule is that the difference between two numbers next to each other is the same.
12 8 4: the same number is subtracted each time to form the next number

The rule is adding a number, always the same one to form the next number.
1 4 9: any three numbers in order of magnitude

The rule is any three numbers in order of magnitude. (17 minutes.)

No. 5. Female, aged 19, 1st year undergraduate
1 3 5: add two to each number to give the following one
16 18 20: to test the progression of two theory, using even numbers
99 101 103: to test the progression of two theory, using odd numbers

As these numbers can hardly have any other connection, unless it is very remote, the rule is a progression of adding two, in other words either all even or all odd numbers.
1 5 9: the average of the two numbers on the outside is the number between them

The rule is that the central figure is the mean of the two external ones.
6 10 14: the difference between the first two numbers, added to the second number gives the third

7 11 15: to test this theory
2 25 48: to test this theory

The rule is that the difference between the first two figures added to the second figure gives the third.
7 9 11, 11 12 13, 12 9 8, 77 7.5 71

Subject gives up. (45 minutes.)

No. 6. Male, aged 23, 2nd year undergraduate
8 10 12: step interval of two
7 9 11: with numbers not divisible by two
1 3 5: to see if rule may apply to numbers starting at two and upwards
3 5 7: rule does not necessarily require ascending order
5 3 1: could be descending order

The rule is that the three numbers must be in ascending order separated by intervals of two.
11 13 15: must have one number below ten in the series
1 6 11: ascending series with regular step interval

The rule is that the three numbers must be in an ascending series and separated by regular step intervals.
The rule is that the first number can be arbitrarily chosen; the second number must be greater than the first and can be arbitrarily chosen; the third number is larger than the second by the same amount as the second is larger than the first.
1 3 13: any three numbers in ascending order

The rule is that the three numbers need have no relationship with each other, except that the second is larger than the first, and the third larger than the second. (38 minutes.)

DISCUSSION

Only six of the 29 subjects gave the correct rule at their first announcement. These subjects tended both to eliminate more possibilities, and to generate more negative instances than did those who announced a first incorrect rule. Significant differences were obtained between these two groups on both criteria at the 0.0001 level of confidence.

On the other hand, the 13 subjects who announced only one incorrect rule presumably did so on the basis of simple enumeration, i.e. they assumed that confirming evidence alone justified their conclusions. But there are two other possible explanations of their behaviour.

Firstly, these subjects may have announced a rule in the hope that doing so would remove them from the experimental situation. This seems unlikely, however, because they all appeared highly motivated, and they expressed considerable surprise when told that their rule was not the one which the experimenter had in mind.

Secondly, they may have assumed that there could be only one rule to which the initially given instance could conform. Familiarity with the “number series” type of problem in which there is supposed to be only one correct continuation might have induced a set for the one “right” answer. It could also be argued that the correct rule (increasing magnitude) was so trivial that

students would have been reluctant to entertain it. However, the point is not that most subjects failed to give the correct rule at their first announcement, but that they adopted a strategy which tended to preclude its attainment.

But after they had announced one incorrect rule, 10 of these 13 subjects (three made no further announcement) gave the correct rule at their next announcement. And the results suggest that they did this by eliminating more possibilities, and generating more negative instances than those subjects who announced a second incorrect rule. The difference between these two groups was just short of significance ($p = 0.08$) on the eliminative/enumerative index, and was significant at the 0.003 level on the negative instance index. Thus, it is possible that, at the beginning of the experiment, the reinforcement of these subjects' rules by their confirming instances blocked the notion that there might be any alternative. But after this set had been broken by an incorrect announcement, they were able to eliminate any remaining alternative which occurred to them.

These possibilities can hardly apply to those nine subjects who announced two or more incorrect rules. For after their first announcement they had evidence to show that stating rules would not remove them from the situation, and that there could be more than one possible rule. Hence, it appears as if from then on they were reasoning by simple enumeration, and announcing sufficient rules, either from an inability to do otherwise, or from a preference for what Bruner has called a "direct test." In other words, they might not have known how to attempt to falsify a rule by themselves; or they might have known how to do it, but still found it simpler, more certain or more reassuring to get a straight answer from the experimenter about the correctness of their rules. This second possibility, however, seems rather remote because the method of elimination is not difficult to apply. The attempted falsification of a rule in no way depends on discovering a suitable alternative to substitute for it. All that the subject had to do is to generate an instance which is similar to previous positive ones, but does not conform to the tested rule. If the outcome is positive, the rule can be decisively eliminated. Thus, the subject has to reason that a rule will be false if it does not cover a positive instance of the correct rule. He must, at some stage, relinquish a rule which may have been confirmed, and adopt that eliminative strategy which Bruner calls "conservative focusing" and which Mill called the method of difference.

The announcement of a sufficient rule is, in fact, the frequent result of enumerative thinking. In the present investigation, in contrast to Bruner's experiment, the use of confirming evidence alone will compel the announcement of a rule, as the only way of finding out whether or not it is the correct one. Here there are no ready-made instances displayed which might be used to correct a sufficient rule. On the contrary, instances exemplifying such a rule can never be exhausted. Thus, the experiment demonstrates the dangers of induction by simple enumeration as a means of discovering truth. In real life there is no authority to pronounce judgment on inferences: the inferences can only be checked against the evidence.

The results show that very few intelligent young adults spontaneously test their beliefs in a situation which does not appear to be of a "scientific" nature. The task simulates a miniature scientific problem, in which the variables are unknown, and in which evidence has to be systematically adduced to refute or support hypotheses. Generating an instance corresponds to doing an experiment, knowledge that the instance conforms, or does not conform, corresponds to its result, and an incorrect announcement corresponds to an inference from uncontrolled data. The kind of attitude which this task demands is that implicit in the formal analysis of scientific procedure proposed by Popper (1959). It consists in a willingness to attempt to falsify hypotheses, and thus to test those intuitive ideas which so often carry the feeling of certitude.

The methodological analogue of this attitude consists in the use of increasingly stringent controls. Obviously scientific method can be taught and cultivated. But the readiness (as opposed to the capacity) to think and argue rationally in an unsystematized area of knowledge is presumably related to other factors besides intelligence, in so far as it implies a disposition to refute, rather than vindicate assertions, and to tolerate the disenchantment of negative instances. And certainly these qualities are no less important for thinking in general than the more obvious cognitive functions associated with purely deductive reasoning.

I am primarily indebted to Dr. A. K. Jonckheere. Frequent arguments with him about the logical issues involved in this research have greatly helped to clarify my ideas, and his constructive criticism of the manuscript has been invaluable. I should also like to thank Professor G. C. Drew, Dr. R. P. Kelvin and Mr. R. D. Shepherd for valuable comments. Finally, I am indebted to my subjects for the great interest and enthusiasm which they expressed in the experiment.