

CONFIRMATION BIAS AND THE USE OF EVIDENCE IN THE 2-4-6 TASK

ALBERTO GRECO

University of Genoa

Abstract. *This paper discusses whether studies using the 2-4-6 task demonstrate that the inductive discovery of a rule is hindered by confirmation bias. It is argued that, since such studies treat as “confirming” any example compatible with the subject's current hypothesis, it is not possible to say whether an example is put forward to confirm or disconfirm a hypothesis when that hypothesis cannot be clearly identified. Second, a subject's strategy cannot be evaluated without considering his or her expectations, since it is possible to choose examples that contrast with one's hypothesis in order to confirm it, or conversely, examples that are congruent with it in order to falsify it. Further problems stem from the artificiality of the task: the instructions do not favor the construction of a clear representation of the problem, and no account is taken of subjects' conceptual expectations regarding tasks of this kind. An experiment is presented in which an attempt was made to overcome these methodological limitations. The results indicate that the use of confirmatory or falsificatory strategies does not affect success; instead, the use of an overall focusing strategy and the manipulation of subjects' expectations through preliminary instructions appear to be relevant. In conclusion, it is argued that the impact of confirmatory or disconfirmatory strategies on inductive discovery cannot be examined without an overall analysis of the process of evidence use.*

Introduction

The 2-4-6 task, originally proposed by Wason (1960), continues to be used in research on the strategies subjects adopt when trying to discover a rule inductively by choosing the examples to be tested themselves (Tweney, Doherty, Worner, Pliske, & Mynatt, 1980; Gorman & Gorman, 1984; Tukey, 1986). As is well known, since the target rule is very general (any three numbers in ascending order), subjects typically hypothesize rules that are more specific than the target, testing examples that always turn out to follow it and thus obtaining many confirmations for their hypotheses. The interest of the task therefore lies in the fact that it reproduces a situation in which explanations of empirical data are easily confirmed yet wrong — a situation that is common in scientific inquiry.

The central point of studies adopting the 2-4-6 task is to show that discovery of the rule is hindered by the adoption of an inadequate strategy — that is, by the tendency to confirm one's own hypotheses (confirmation bias) or by the failure to use disconfirmation where it would be preferable. This interpretation is suggested by the fact that subjects who solve the task at their first stated hypothesis turn out to be those who show the greatest use of a disconfirmation strategy. Both Wason and other authors have pointed out that this finding appears consistent with Karl Popper's well-known epistemological position (1959), according to which scientific research essentially progresses when one is able to falsify hypotheses rather than confirm them.

However, the available experimental results do not clearly and unambiguously support the view that subjects' difficulties with the task are attributable to the confirmation-disconfirmation dimension. For example, Tweney et al. (1980) attempted to modify subjects' strategies using instructions that made them aware of the possibility of using a disconfirmation strategy; although they succeeded in encouraging that strategy, they failed to obtain any improvement in performance. Gorman and Gorman (1984), by contrast, following a similar methodology with explicit disconfirmation instructions but avoiding giving feedback on the subject's provisional hypotheses (as had already been done in the Wason and Katzman experiment cited in Wason & Johnson-Laird, 1968), obtained results supporting a decisive role for disconfirmation in performance. According to Gorman and Gorman, when subjects are required to decide for themselves

whether a hypothesis is correct, instructions that induce disconfirmation are also effective in improving performance in discovering the target rule. In contrast, our own earlier experiment (Greco, 1989) found no link between the use of disconfirmation and rule discovery (although in that experiment feedback was normally given after each hypothesis was stated).

These contradictory results may stem from the fact that studies using the 2-4-6 task have often introduced minor modifications to one aspect or another of it without reconsidering its overall context. Moreover, each of these replications has presented the task in a different way, and these ways are not always equivalent in terms of how subjects represent the problem.

The present study therefore aims to discuss the assumption that inductive discovery is necessarily hindered by confirmation bias, and to highlight some problems and inconsistencies in the experimental design adopted in the various replications of Wason's 2-4-6 task. Even though the task is clearly inadequate as an emulation of the scientific research context (Klayman & Ha, 1987; Moneta, 1987) and of everyday inductive reasoning itself (Wetherick, 1962), we believe it is still worth reflecting on some of the crude simplifications this paradigm has involved.

In the next section we will examine two issues that, in our view, constitute important sources of difficulty for this empirical paradigm. The first derives from the operational definition of the concepts of confirmation and disconfirmation usually adopted; the second concerns the artificiality of the task. In the following section we will present an empirical study in which the experimental conditions were modified so as to avoid these problems.

Analysis of the 2-4-6 Paradigm

Operational definition of “confirmation” and “disconfirmation”

For Wason, to confirm means to propose triads of numbers that are compatible with the hypothesis a subject holds in mind at a given moment (and, conversely, to disconfirm means to propose triads that are incompatible with that hypothesis). This definition entails at least two problems.

a) Formulation of hypotheses

The first problem is that this definition presupposes the existence and clear identifiability of a hypothesis behind every example a subject proposes for testing; however, there are cases in which it is impossible to determine what hypothesis the subject has formulated, or even whether he or she has formulated one at all. In such cases it becomes difficult to assess whether an example is disconfirming or not, and to relate the discovery or non-discovery of the rule to a specific strategy.

This problem arises from the fact that Wason's paradigm (and that of subsequent researchers) simply takes for granted that subjects always use the strategy of successively examining hypotheses, which Bruner, Goodnow, Austin, and Brown (1956) termed scanning, and which was of particular interest to Wason when he devised the 2-4-6 task.¹ No consideration was given to the possibility that subjects might use the other strategy (called focusing), which consists of concentrating attention on an attribute (e.g., “even” or “upper limit”) or an operator (e.g., “add 2”) and then testing examples in which that attribute or operator is changed

¹Wason explicitly stated that the task had been devised with the intention of reproducing the situation of “induction by enumeration” that arises when the strategy of successive examination is adopted (cf. Wason & Johnson-Laird, 1972, pp. 245 ff., Italian translation).

or replaced. In fact, empirical research shows that this strategy is very often preferred by subjects (cf. Greco, 1989).

If confirmation bias is assessed and quantified — as has been done in studies following Wason's paradigm (e.g., Wetherick, 1962; Tweney et al., 1980; Gorman & Gorman, 1984) — by considering the number of examples congruent with the current hypothesis, then in cases where there is in fact no hypothesis at all but only a systematic variation of attributes, or a fluctuation between scanning and focusing, or even random exploration, it is difficult to obtain such a measure without distorting the protocols.²

In the classical experimental context, in short, no explicit attempt was made to check whether the instructions induced subjects to formulate hypotheses, how these hypotheses were encoded, or how subjects judged the compatibility of tested examples with those hypotheses. In the studies cited above, the instructions were modified only to get subjects to adopt a disconfirmation strategy, but no attempt was made to facilitate the adoption of the strategy of successive hypothesis examination (scanning), nor to encourage a clear encoding — first of all for the subjects themselves — of their hypotheses, together with a check on how well those hypotheses corresponded to the examples tested.³

b) Falsification from the subjective point of view

A second problem arising from Wason's definition of confirmation-disconfirmation calls into question the assumption implicit in it — namely, that an example chosen by a subject for testing should be regarded as disconfirming only on condition that it contrasts with his or her current hypothesis.

Klayman and Ha (1987) clarified an aspect of the 2-4-6 task that makes it distinctive while also being the source of some confusion about “disconfirmation.” These authors showed that the relationship between the hypothesized rule and the rule to be discovered (the target) is not necessarily the same in every inductive task as it is in the 2-4-6 task. In Wason's task, the hypothesized rules are almost always special cases of the target rule, but in other tasks it is possible to hypothesize rules that are more general than the target, or rules that only partially overlap with it.

This analysis shows that adopting a disconfirmation strategy in Wason's sense is not always and inevitably the most rational choice. Indeed, in cases where the relationship between the hypothesized rule and the target rule differs from that in the 2-4-6 task, proposing examples that contrast with one's own hypothesis may not be a valid strategy.⁴

²For example, when a subject keeps changing the attribute on which the choice of examples to test is based, one is forced to treat each triad proposed as motivated by a new hypothesis, whereas in reality the subject's reasoning process is entirely different. In the case of scanning, a hypothesis is normally abandoned only once it has been found to be wrong, whereas if a focusing strategy is followed, testing of an attribute may be temporarily set aside — even though it has not been eliminated — and possibly taken up again later.

³An earlier experimental study (Greco, 1989) found not only that subjects engaged in this task often prove unable to formulate hypotheses, but also that, in cases where they do, they are unable to do so clearly or to follow a coherent strategy. The empirical evidence shows that subjects frequently propose examples to be tested without being able to state precisely what hypothesis motivates them, or by reporting hypotheses that are vague, confused, contradictory with the evidence already obtained, or that do not motivate the chosen examples at all. These observations, reminiscent of the hypothesis of “tacit knowledge” (Polanyi, 1966), suggest that the relations grasped in the data can be encoded at different levels of explicitness, ranging from a vague impression to a formulation in clear, precise, and complete terms.

⁴Indeed, if the hypothesized rule were more general than the target rule (for example: rule to be discovered = “consecutive even numbers”; hypothesized rule = “numbers increasing by 2”), the only way to falsify the hypothesis would be to propose examples that follow the hypothesized rule and receive the answer that they do not follow the target rule (e.g., numbers increasing by 2 that are not also consecutive even numbers, such as 3-5-7). In other words, an eliminative line of reasoning would in this case require a confirmation strategy, since a disconfirmation strategy could

The point is that in this context “disconfirmation” can mean two different things: testing cases that contradict one's own hypotheses, or testing cases that do not follow the target rule. Klayman and Ha (1987) showed that in the 2-4-6 task these two things tend to coincide, because in it the hypothesized rule is in all likelihood a special case of the target rule; only under these conditions is Wason actually right that the only way to falsify a hypothesis is to propose examples that contrast with it and be told that they do follow the target rule. But for the purposes of falsification, what matters is not contradicting one's own hypothesis as such, but demonstrating that it is wrong and thus eliminating it.

Another case in which a hypothesis could be falsified would be by proposing examples compatible with it in the hope of receiving the answer “no, it does not follow the target rule.” In the 2-4-6 task such an answer is objectively impossible, but if one is interested in the subject's strategy, it is important and appropriate to also consider which examples are falsifying from the subject's own point of view. As Wetherick had already observed in 1962, one cannot actually know whether a subject consciously eliminates a hypothesis or not without asking, for every triad tested, whether the subject expects a positive or negative answer (that is, whether or not he or she expects it to follow the target rule). Gorman and Gorman (1984) likewise stressed the importance of distinguishing between examples predicted to be non-conforming to the target rule and examples that actually turn out to be non-conforming. “Disconfirmation” in a subject can therefore be revealed not only by considering whether he or she proposes examples that contrast with his or her hypotheses (Wason's criterion) but also by considering what the subject expects when proposing an example.⁵

Based on these considerations, in order to overcome the problem it would be appropriate to calculate the disconfirmation index by considering the cases in which a subject truly intends to see whether the hypothesis formulated is false or, if none has been formulated (we have already discussed the importance of this case), whether the attributes identified in “2-4-6” are irrelevant. These cases reduce to the following two: a triad congruent with the hypothesis (or with the identified attribute) but with an expected answer of “no”; and a triad not congruent with the hypothesis (the case considered by Wason), provided the expected answer is “yes.” As shown earlier, the first of these two cases cannot lead to any falsification within the 2-4-6 task; nevertheless, it can be regarded as falsifying from a subjective, intentional point of view.

Artificiality of the task

Here we will set out certain aspects that make the standard 2-4-6 task rather artificial — not only in terms of adequately emulating the strategies adopted in situations requiring induction, but also simply in terms of justifying the conclusion that difficulties in discovering the rule are due to the adoption of a wrong strategy.

a) As noted in the already-cited work of Gorman and Gorman (1984), allowing subjects to state several hypotheses and giving feedback on each one may discourage subjects from reasoning for themselves about the consequences of a hypothesis, and in particular about the possibility of eliminating it purely on the basis of evidence rather than because they were simply told to. In any case, this kind of feedback is not natural: a

never falsify the hypothesis (examples contrasting with the hypothesis would be numbers not increasing by 2, which would always also contrast with the “consecutive even numbers” rule, thus invariably producing a “no” answer equivalent to a verification of the hypothesis).

⁵For example, proposing “8-10-12” to test the hypothesis “numbers increasing by 2” would be a disconfirmation strategy for a subject who expected the answer “no, it does not follow the target rule” — and indeed such an answer would eliminate the hypothesis. Conversely, proposing “1-10-20,” again to test “numbers increasing by 2,” while expecting the answer “no,” would actually be a confirmation strategy for one's own hypothesis, even though the example chosen contrasts with it.

scientist, for example, is never told whether his or her hypotheses are right or wrong, but must find that answer by questioning nature.

b) The standard instructions do not provide subjects with clear, uniform criteria for deciding when a rule can be considered acceptable — and therefore statable. Stating the rule “when you are sure” may, for some subjects, mean when it is “plausible,” for others when it is “convincing,” and for still others (very few) when it is “proven.” The instructions refer to “certainty,” not to logical necessity, yet the results are then interpreted as demonstrating the existence of logical fallacies in subjects' reasoning.

c) More generally, the subject is not encouraged to construct a “mental model” of the task (Johnson-Laird, 1983). The importance of this aspect was shown by Green (1989), who modified the task instructions by asking a group of subjects, before administering the usual test, to think of and write down possible rules, specifying the relationships among the various hypothesized rules, which triads they would use to test them, and what they would do if all the triads turned out to conform to the rule. This group's performance was facilitated because, presumably, they succeeded in constructing a mental model of the task — that is, in representing alternative rules and the relationships among them.⁶

d) No account is taken of subjects' expectations regarding how to conceptualize the task. These conceptual expectations interfere with task performance in various ways: for example, they can affect the analysis of the initial example given, when the subject is searching for relations among the three numbers. These expectations can be regarded as resulting from the activation of schemas or frames in Minsky's sense (1975), with the important characteristic of containing default information for those aspects that have not been explicitly specified.

An example of this kind of expectation concerns the level of generality of the rule. Rosch's research (1975) has shown that, in the absence of contrary information, elements to be categorized are by default assumed to belong to a “basic” category rather than a super-ordinate or sub-ordinate one. In Wason's task this translates into a greater plausibility for rules such as “even numbers” rather than “any numbers in ascending order” or “even numbers increasing, below 10, with a difference of 2 between the digits.”⁷ More generally, subjects' expectations interact with the way the task is presented. For instance, as observed by Legrenzi and Rumati (1985), presenting three even numbers increasing by two “says more than necessary” for a general rule and “less than necessary” for a very specific rule.⁸

These observations thus indicate that, in using Wason's task, due importance must be given to analyzing the schemas subjects use to understand the task, but also that the importance of how the presentation instructions are formulated — and especially of the “preliminary examples” usually offered to explain the task — must not be neglected. These conceptual expectations can be influenced by appropriately modifying such instructions.

⁶Unfortunately, this article does not describe the procedure followed in detail. In particular, it would be important to know whether a preliminary example was presented.

⁷Another expectation, which we have not addressed in this study, may concern whether or not the order of the elements is relevant as a relation among them. In this case, order does not by default appear to be relevant, since the triad is usually taken as a set of numbers linked by some relation but statable in any order.

⁸If one assumes that the same rules that apply to communicative exchanges hold in this context, this information can mislead subjects because it violates the so-called “Gricean maxim”: the recipient of a message assumes that the sender does not say anything more than necessary. Moreover, subjects may be misled by the fact that they are dealing with a psychology experiment, and that this type of numerical series often characterizes certain intelligence tests, where the subject is expected to continue the series having grasped not an arbitrarily chosen rule but the most “natural” one.

Empirical Study

Study methodology

In the empirical part of the study, while keeping the original 2-4-6 task unchanged so as to allow comparisons with previous research, modifications were introduced both in the instructions and in the analysis of the results. We will discuss the latter when presenting the results themselves; here we will examine the modifications to the instructions, which were aimed at overcoming the methodological difficulties outlined above and will be taken up again here. The conditions described below concerned the treatment of the subjects making up the experimental group, whereas Wason's original procedure was followed for the others.

a) In order to encourage, as far as possible, an explicit and complete encoding of hypotheses and the development of a mental model of the task, subjects were first asked to write down all the rules that came to mind (as already done by Tukey, 1986, and Green, 1989). To ensure that they did so as clearly as possible, it was specified that the statement of these rules would need to be comprehensible to other subjects to be examined subsequently.

b) To overcome the problem of the lack of correspondence between stated hypotheses and chosen examples, an effort was made to ensure that subjects, in choosing the examples to test, were aware of whether or not those examples were consistent with the hypothesis or the identified attribute; they were therefore asked to give a written justification for the choice of each example. In addition, they were asked to indicate what feedback they expected from each triad. Recording these expectations is important for redefining disconfirmation in subjective terms.

c) To ensure that the decision to state the rule was based exclusively on logical necessity, the instructions allowed subjects only a single statement of the rule, to be made only when they believed it had been proven (as in Gorman & Gorman, 1984); this device was also intended to eliminate the unnatural feedback on any previous attempts to state hypotheses.

These manipulations were not expected merely to make it easier for subjects in the experimental group, compared with the control group, to adopt a falsification strategy (a variable that was controlled for in all the studies that modified Wason's original design); of more specific interest was whether:

- a) they would follow the strategy of successive hypothesis examination (scanning) to a greater extent;
- b) they would achieve greater success in solving the task in relation to the strategy adopted;
- c) they would, in the end, formulate more articulated hypotheses, and whether this would help their performance.

Finally, to assess the influence of expectations induced by the preliminary example regarding the rule's level of generality, that example used rules that were more or less general depending on the subgroup to which subjects belonged. This assessment was carried out on an exploratory basis, given the small number of subjects who could be assigned to these subgroups.

It should be added that the task was presented as the discovery of the law governing the response given by a computer whenever the appropriate numbers appeared in the three cells displayed on the monitor. In this way, the rule is more readily regarded as a "law of nature," less tied to the experimenter's arbitrary choice.

Procedure

The subjects are 50 students who take part in the study voluntarily. 25 subjects, chosen at random, make up the control group, for whom the task follows Wason's original procedure; the description below therefore concerns the other 25 subjects, who make up the experimental group.

The task is presented as the discovery of the rule built into the computer program which, depending on the numbers present in the three cells displayed on the monitor, responds by writing "YES" or "NO" when the subject presses a key. The monitor used by the subject is located in a room adjoining the one containing the central unit, and the subject does not have a keyboard available, only a large button; the numbers to be entered into the cells are communicated to the experimenter, who makes the change. On the subject's monitor, all the triads supplied and their respective feedback remain displayed progressively; no other message appears. Subjects are given the opportunity to write down any notes they wish. The instructions explain the nature of the task using a preliminary example involving animal names (dog, wolf, hyena). For 15 randomly chosen subjects, the rule is general (any three animals); for the other 10 subjects the rule is more specific: for 5 it is basic (three quadruped animals) and for 5 it is specific (three carnivorous quadruped animals). Otherwise, the instructions are the same for all subjects in the group, requiring them first to think of all the rules they could hypothesize and to write each one down as clearly as possible (it is specified that they must be comprehensible, since they will be used by other subjects). Likewise, for every triad proposed, subjects must write down the reason for the choice and what outcome is expected from the test. Subjects are told that the rule must be stated only on the basis of solid reasons that demonstrate its validity, and that it will not be possible to state further rules afterward.

Results and discussion

a) Effects on the scanning-focusing strategy

First, we examined whether the modifications introduced had any effect on subjects' strategy. Since the instructions aimed to elicit more coherent strategies from subjects in the experimental group, it was expected that they would decisively adopt the strategy of successive hypothesis examination (scanning).

Since previous research had not taken the scanning-focusing dimension into account,⁹ a dependent measure capable of revealing the difference between the two strategies had to be devised. The idea underlying the operational measure chosen for this assessment starts from the observation that a scanning strategy, being centered on a hypothesis, tends to produce a certain number of examples addressing the same attribute until the hypothesis is abandoned, whereas a focusing strategy leads to continuous variation of the attributes motivating the choice of triads. A measure was therefore devised, called the perseverance index, of the quantity of examples presented to test the same attribute, excluding the first occasion on which that attribute is proposed. This index was obtained using the following formula:

$$\text{Index} = \frac{\sum a_i - n}{\sum a_i} \times \sum a_i^2$$

(where a_i are the frequencies of the individual attributes and n is the number of attributes) — that is, the difference between the total number of attributes and the number of distinct attributes, expressed as a proportion of the total number of attributes. The formula is completed by multiplying this ratio by the sum

⁹Only Tukey (1986) classified the strategies used by his subjects as centered on the hypothesis (scanning) or on the attribute (focusing). To do this, however, he did not rely directly on the protocols of the inductive task but on a self-assessment of strategy made retrospectively by the subjects themselves, with reference to a list provided by the experimenter.

of the squares of the individual frequencies for each attribute, which serves as a correction to give greater weight to the number of times the same attribute is proposed again (otherwise, for an equal number of attributes, it would not be possible to distinguish between a case in which the attributes have equal frequency and a case in which one has a very high frequency and another a very low one).

It was decided to treat perseverance-index values, calculated in this way, that fell below the mean as indicative of an overall structure tending toward focusing, and values above the mean as indicative of a tendency toward scanning. Since, according to expectations, the experimental group should have adopted the scanning strategy more, its perseverance index should have been higher.

Given this, if we now examine the results, we can see that they ran contrary to expectations: despite the modifications to the instructions, no significant differences were found between the perseverance indices of the two groups (see Table 1a). In both groups these indices remained fairly low, indicating a preference for an attribute-centered, or focusing, strategy. This result lends further weight to doubts about whether subjects always have some hypothesis to test.

b) Effects on disconfirmation and success

The main purpose of the empirical part of the present study was to consider the possibility that confirmation bias hinders success in the task. First, measures of confirmation or disconfirmation were calculated following Wason's standard criteria: a disconfirmation index was therefore calculated, consisting of the ratio — for each subject — between the number of examples incompatible with the current hypothesis or with the attribute currently under consideration and the total number of examples produced (cf. Gorman & Gorman, 1984, p. 646), regardless of the answer the subject expects. Disconfirmation calculated in this way was higher, though not significantly so, in the experimental group (Table 1d).

The disconfirmation index was also calculated separately for subjects who solved the task and for those who did not. With this treatment and analysis of the data, it turns out that subjects who solve the task disconfirm significantly more than the others, both in the experimental group (Table 2c) and in the control group (Table 2b). Moreover, if we reconsider the superiority found, in terms of disconfirmation, in the experimental group, this difference is significant among the subjects who do not solve the task (Table 1e). This is essentially in line with what emerges from the comparison in Table 1c: the manipulation appears to have produced more disconfirmation but not more success. Even stopping the analysis at this point, then, one might already conclude that the conclusions reached by Tweney et al. (1980), and those that emerged from our own earlier experiment (Greco, 1989), are well founded — namely, that the use of confirmatory or disconfirmatory strategies is irrelevant to rule discovery.

There are, however, other indications leading to the same conclusion. Since, as noted, in the present experiment subjects in the experimental group were asked, among other things, to indicate for each triad proposed for testing which answer they expected, it was possible to obtain, for the subjects in this group, a measure of confirmation or disconfirmation different from the one considered so far.

This revised index is not based on the number of examples compatible or incompatible with the current hypothesis but instead treats as “disconfirming” those examples that meet the requirements identified in the section Falsification from the subjective point of view:

- a) they contrast with the hypothesis under consideration at a given moment, and the subject's expected answer is “YES”;
- b) they do not contrast with that hypothesis, and the subject's expected answer is “NO”. The remaining examples are obviously treated as “confirming.”

If the comparison between the confirmation scores of subjects who solve the task and those who do not is made using this revised index, subjects who solve the task appear to disconfirm slightly more than the others, but the difference is no longer significant (Table 2d).

This fact can be explained by an analytical examination of the scores contributing to the disconfirmation indices calculated in the two ways. Table 3 shows how the triads tested in the control group are distributed according to type of test (compatible – incompatible), expectation (YES – NO), and success; the table also shows (under “Analysis”) how these are treated for the purposes of the confirmation-disconfirmation score by the traditional analysis and by the revised analysis proposed here. As can be seen, the most frequent triads, after those compatible with a YES expectation, are those incompatible with a NO expectation: using Wason's procedure, this type of example could not be detected and therefore could not be correctly regarded as a case of confirmation-seeking; indeed, being incompatible, these triads unrealistically inflated the disconfirmation score.

On the basis of these observations, we believe it is possible to cast doubt on empirical data that measure confirmatory or disconfirmatory strategies without accounting for subjects' expectations, and to revise the conclusions drawn regarding the relationship between such strategies and success. In particular, these results once again lend support to the conclusions of the studies cited above (Tweney et al., 1980; Greco, 1989), which cast doubt on the connection between confirmation or disconfirmation and rule discovery.

c) Effects on the articulation of hypotheses

An attempt was also made to see whether the different treatment had any effect on the explicitness and articulation of any hypotheses formed. This variable is not easy to assess empirically, but a sufficiently indicative measure can be the number of jointly mentioned attributes cited by subjects when proposing each triad to be tested. For instance, if a subject expresses his or her hypothesis, in testing “8-10-12,” by mentioning the attributes “even” and “increasing by 2,” this formulation can be considered more explicit than one that mentions only “even,” since in the latter case the subject, while in fact testing both attributes, would show awareness of only one of them.

Inducing subjects to preliminarily formulate a certain number of hypotheses verbally, in writing, and thus in an articulated way, had the effect of significantly increasing, compared with the control group, the number of different attributes mentioned when proposing the triads to be tested (Table 1b). It can therefore be said that the diversity of experimental conditions favored more articulated hypotheses. Moreover, if this effect is related to the results concerning success, it emerges that the number of different attributes mentioned (and hence the degree of articulation of hypotheses) is significantly higher for subjects in the experimental group who discover the rule than for those who do not (see Table 2a). This result highlights the importance, for success, of formulating a model of the task; nevertheless, the fact remains that the manipulations did not improve the performance of the experimental group as a whole compared with the control group.

d) Effect of the instructions

One final assessment (as noted, exploratory in nature) concerned the impact of manipulating the level of generality of the preliminary example. This change did not produce a significant improvement in performance (see Table 4a), but subjects in the subgroups whose preliminary example was basic or specific examined a significantly higher number of attributes (Tables 4b and 4c) and more frequently proposed, as their final rule, a rule that included the correct one but was more specific (e.g., “increasing and even numbers,” etc.; see Table 4e). It seems to us that, although one cannot say that subjects whose example was general were facilitated, one can certainly say that the others were hindered. A finding of particular interest (Table 4d) is the very high proportion of negative examples (i.e., examples that do not follow the rule to be discovered)

encountered by subjects in the subgroup whose example was specific — a proportion that nonetheless did not help them, precisely because they were misled by the expectation of a very specific rule.

Conclusion

In the present study we have considered some problems that lead us to doubt the claim that experiments based on the 2-4-6 task demonstrate that confirmation bias hinders the process of inductive discovery. The reasons for this doubt lie essentially in certain difficulties in the usual methodological approach to this task. In particular, it does not seem tenable to speak of confirmation or disconfirmation solely with reference to hypotheses about the rule, when it is evident that subjects simply tend not to reason by examining successive hypotheses (that is, to adopt a scanning strategy), even when — as in the present empirical study — they were encouraged and helped to do so. It seems preferable instead to assess confirmation or disconfirmation with reference to the attributes or operators examined progressively in the tested examples (without necessarily regarding them as “hypotheses” about the rule), while taking into account which answers subjects expect from their tests.

The results have certainly not ruled out the existence of confirmation bias; indeed, measuring this variable in the way proposed above (that is, treating as confirming even those triads incompatible with the examined attributes when the expected answer is “no”) makes it clear that the tendency toward confirmation is greater than what emerged with the traditional analysis. However, there clearly appears to be no relationship between the use of confirmation or disconfirmation and success or failure in the task.

The possibility that seems most worth considering, instead, is that success is linked to the use of strategies broader than confirmation or disconfirmation. Indeed, the fact that the perseverance index of subjects who solved the task is significantly lower than that of subjects who did not (see Table 2e) may indicate that subjects who identify the rule prefer an attribute-centered, or focusing, strategy. This would not be surprising, moreover, since Bruner et al. (1956) had already observed that this strategy requires fewer cognitive resources, in terms of memory and inferential capacity, than the other one. What may be surprising is why research adopting Wason's paradigm has taken the use of a scanning strategy for granted.

Overall, however, the variables most relevant to inductive discovery, if there are any, do not seem to be identifiable using the 2-4-6 paradigm. It might be said that what actually facilitates or hinders discovery of the rule continues to elude us — assuming there is anything beyond sheer “luck,” as Tweney et al. (1980) concluded at the end of their Experiment II.

In our view, however, a conclusion of this kind is inevitable only if one continues to neglect an overall analysis of the cognitive process brought into play in tasks that call for inductive inquiry. This is a process (which we might call “use of evidence” or, if one prefers, “investigative style”) that can be assumed to develop somewhat independently of inductive reasoning proper, since its adoption is also tied to implicit assumptions, individual styles, and so on, which are not directly connected to what is specifically being investigated. Despite the substantial lack of analysis of this process, researchers did not hesitate to claim that the results of experiments with the 2-4-6 task demonstrated that confirmation bias hindered rule discovery. It seems to us, instead, that an investigation into the real causes of difficulty or facilitation, in this context as in similar ones, should take into account the procedures subjects put into practice when tackling the task, and the expectations that may mislead them.

In more general terms, the problem is essentially to study how expectations, mental models, ways of managing cognitive resources, and procedures interact to constitute those standardized systems of evidence

use that we call “strategies.” What is especially important is to bear in mind that general strategies (scanning/focusing or confirmation/falsification) are not options external to the task that subjects choose to guide its execution, but form an integral part of the representation of the task itself. Within this broader framework, some of the issues discussed in the present paper are better understood — such as the difficulty of assessing the intention to confirm or disconfirm with reference to “hypotheses” rather than “expectations.” Hypotheses are more elaborate, non-immediate constructions that emerge from the comparison between expectations and data. Strategies, on the other hand, might be better conceptualized as the particular type of schema known as scripts (Schank & Abelson, 1977), and the difficulty may depend on the fact that, depending on the type of script used, only certain evidence is expected, while, on the other hand, each script provides for only a certain use of that evidence. We believe that future exploration of inductive discovery tasks along these lines could lead to interesting results.

Tables

Table 1. Comparison between the experimental group and the control group

	S	C	t	df	p
a. Perseverance index	4.79	6.75	−0.54	48	0.70
b. Number of different attributes	5.4	3.7	2.47	48	0.003*
c. Success (percentage)	44.0	40.0	0.28	48	0.39
d. Disconfirmation	24.0	19.0	0.68	48	0.25
e. Disconfirmation among subjects who do not solve the task	18.7	5.6	1.48	25	0.075*

S = experimental group; C = control group. * Significant values at $p < 0.10$

Table 2. Comparison between subjects who solve and subjects who do not solve the task

	m	R	NR	t	df	p
a. Number of different attributes (experimental group)	5.4	6.7	4.5	1.87	23	0.037*
b. Disconfirmation index (control group)	19.0	31.4	5.6	2.60	23	0.008*
c. Disconfirmation index (experimental group)	24.0	32.0	18.7	1.39	23	0.089*
d. Revised disconfirmation index (experimental group)	18.4	21.6	16.3	0.77	23	0.22
e. Perseverance index	4.8	2.2	6.5	−1.37	23	0.091*

R = subjects who solve; NR = subjects who do not solve; m = mean. * Significant values at $p < 0.10$

Table 3. Percentage composition of triads tested in the control group, and related expectations

Strategy		R	NR	m	Analysis	
test	expectation				traditional	revised
C	YES	54.8	71.2	64.6	conf.	conf.
C	NO	13.1	10.2	11.3	conf.	dis.
I	YES	8.5	6.1	7.1	dis.	dis.
I	NO	23.6	12.5	17.0	dis.	conf.
Total		100	100	100		

C = compatible; I = incompatible. R = subjects who solve; NR = subjects who do not solve; m = overall mean. conf. = confirming; dis. = disconfirming (falsifying)

Table 4. Comparison between groups given instructions with different levels of generality

	g	b	s	F	df	p
a. Success (proportion)	0.33	0.60	0.40	0.52	2	n.s.
b. Mean total number of attributes mentioned	5.87	8.20	9.20	3.78	2	< 0.03*
c. Mean number of different attributes mentioned	4.13	6.80	7.80	4.55	2	< 0.02*
d. Proportion of negative examples encountered	14.5	19.0	40.2	4.68	2	< 0.02*

	g	bs	t	df	p
e. Final failure with a hypothesis more specific than the correct one	0.07	0.30	-1.57	23	0.06*

g = general rule; *b* = basic rule; *s* = specific rule; *bs* = basic or specific rule. * Significant values at $p < 0.10$

References

- Bruner, J.S., Goodnow, J.J., Austin, G.A., & Brown, R.W. (1956). A study of thinking. New York: Wiley & Sons (Italian translation: Il pensiero, Roma: Armando, 1969).
- Gorman, M.E., & Gorman, M.E. (1984). A comparison of disconfirmatory, confirmatory and control strategies on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, 36A, 629-648.
- Greco, A. (1989). Confirmation bias e feedback nel compito 2-4-6. *Ricerche di Psicologia*, 2, 115-130.
- Green, D.W. (1989). Confirmation bias, problem-solving and cognitive models. (In press).
- Johnson-Laird, P.N. (1983). Mental models. Cambridge: Cambridge University Press (Italian translation: Modelli mentali, Bologna: Il Mulino, 1988).
- Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 2, 211-228.
- Legrenzi, P., & Rumiati, R. (1985). Regole in Kripke: sviluppi sperimentali in psicologia. *Nuova civiltà delle macchine*, 3-4, 55-59.
- Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.
- Moneta, G.B. (1987). Are psychological experiments on hypothesis-testing strategies a good simulation of scientific problem-solving? *Epistemologia*, X, 29-54.
- Polanyi, M. (1966). The tacit dimension. London: Routledge & Kegan Paul (Italian translation: La conoscenza inespresa, Roma: Armando, 1979).
- Popper, K.R. (1959). The logic of scientific discovery. New York: Basic Books (Italian translation: La logica della scoperta scientifica, Torino: Einaudi, 1970).
- Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General*, 104, 129-233.
- Schank, R.C., & Abelson, R.P. (1977). Scripts, plans, goals and understanding. Hillsdale, NJ: Erlbaum Associates.
- Tukey, D.D. (1986). A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, 38A, 5-33.

- Tweney, R.D., Doherty, M.E., Worner, W.J., Pliske, D.B., & Mynatt, C.R. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, 32, 109-123.
- Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129-140.
- Wason, P.C., & Johnson-Laird, P.N. (1968). On the failure to eliminate hypotheses [...] A second look. In P.C. Wason & P.N. Johnson-Laird (Eds.), *Thinking and reasoning*. Middlesex: Penguin.
- Wason, P.C., & Johnson-Laird, P.N. (1972). *Psychology of reasoning*. London: Batsford (Italian translation: *Psicologia del ragionamento*, Firenze: Martello Giunti, 1977).
- Wetherick, N.E. (1962). Eliminative and enumerative behavior in a conceptual task. *Quarterly Journal of Experimental Psychology*, 14, 246-249.

[Received 12 June 1989]

[Accepted 12 October 1990]

*Translated from the Italian original: Greco, A. (1991). "Tendenza alla conferma e uso dell'evidenza nel compito 2-4-6." *Giornale Italiano di Psicologia*, a. XVIII, n. 1, febbraio 1991, pp. 79-96.*