

Caverni, J.P., Rossi, S., & P  ris, J.L. (in press). How to defocus in hypothesis testing: Manipulating the status of the initial triple in the 2-4-6 problem. In V. Girotto & P.N. Johnson-Laird (Eds.) *The shape of reason*. Hove: Psychology Press.

How to defocus in hypothesis testing:

Manipulating the status of the initial triple in the 2-4-6 problem

Jean-Paul Caverni^{*1}, Sandrine Rossi^{**}, & Jean-Luc P  ris^{*}

^{*}Aix-Marseilles I University, Marseilles, France

^{**}University of Caen, France

¹ Corresponding author: J.-P. Caverni, Universit   d'Aix-Marseille I, Laboratoire de Psychologie Cognitive (UMR-CNRS 6146), Case 66, 3 Place Victor Hugo, F-13331 Marseille Cedex 3 , France (Mail: caverni@up.univ-mrs.fr) .

We are indebted to, and would thank, Phil Johnson-Laird and Vittorio Girotto for helpful comments and suggestions.

One of the most widely used tasks for studying hypothesis testing in inductive reasoning is Wason (1960) 2-4-6 problem². In this task, the experimenter tells the participants that he has a rule in mind concerning triples of numbers. The triple 2-4-6 is presented as satisfying the rule, and the participants' task is to discover this rule by proposing triples. For each proposed triple, the experimenter says whether or not it follows the rule. The participants are asked to write down each triple, the hypothesis under test, and the experimenter's response. Whenever the participants think they have discovered the rule, they state it, and the experimenter says whether it is correct or incorrect. If incorrect, the participants continue to propose triples until they either discover the rule or give up.

Two phenomena have been observed. In general (1) participants propose triples that are positive examples of the hypothesis they claim they are testing (e.g. the "increasing linear series" hypothesis is tested with triples such as "10-12-14", "10-20-30", etc.), and (2) participants start by hypothesising the most specific rules (e.g. "an increasing series of numbers which differ by 2"), whereas the rule to be discovered is more general ("an increasing series of numbers").

Since 1960 until now, this task has been extensively investigated (for recent reviews, see Gorman, 1995; Poletiek, 2000), and the aforementioned results have been replicated even in studies where the task instructions were changed ("Be disconfirming!" as in studies by Mynatt, Doherty, & Tweney, 1977; 1978) or where the task was transformed (by specifying that the provided feedback may sometimes be erroneous, as in Gorman, 1986). The only ways to induce disconfirming behaviour seem to be to tell participants explicitly to test negative examples (e.g., Gorman & Gorman, 1984), or to tell them that the initial triple is a counterexample rather than an example of the tested rule (Rossi, Caverni & Giroto, 2001), or to ask them to find two rules, one that is satisfied by "dax" triples and one that is satisfied by "med" triples (Tweney, Doherty, Warner, & Pliske, 1980).

The prevailing interpretation for quite some time was that participants are prone to confirmation bias: they would rather verify (or confirm) their hypotheses than falsify (or disconfirm) them. For Evans (1983), however, this bias is not caused by a participants' deliberate choice, but by a cognitive difficulty in processing negative information: "It is not that participants do not wish to falsify, it is simply that they cannot think of the way to do it" (p. 143). In this case, then, we are not dealing with a confirmation bias, but with a positivity

² While we were writing this paper, Peter Wason died. We want to pay homage to his exceptional contribution to cognitive psychology.

bias. This idea conforms to the Wason's original work (e.g. 1959, 1965), which yielded evidence that individuals have difficulties in processing negation.

There seems, however, to be another way to describe the phenomenon. What information does the participant use to interpret the problem statement? One important piece of information here is that the particular triple "2-4-6" is presented to exemplify the rule to be discovered. Suppose a participant is taking an intelligence test and has to complete the series 2-4-6. Obviously, the simplest correct answer is 8-10-12. But in Wason's task, the initial triple is typical of an obvious rule (numbers increasing in intervals of 2) that is more specific than the one to be discovered (increasing numbers). It is neither socially conventional nor didactically efficient to choose examples having conspicuous properties that do not follow the law they are supposed to illustrate (see Armstrong, Gleitman, & Gleitman, 1983). As Girotto and Politzer (1990) remarked, the fact that the 2-4-6 triple evokes a more specific rule than the experimenter's one can be considered as a violation of the Gricean maxim of quantity (Grice, 1989). According to this view, we hypothesise that the example used in the original task misleads the participants by implicitly introducing irrelevant assumptions or constraints. This phenomenon is similar to the focussing process proposed by Legrenzi, Girotto, and Johnson-Laird (1993): the initial premise presented to the participants leads them to focus on certain mental models rather than others, so they restrict their thoughts to what is explicitly represented in their mental models.

What would happen if participants could be led to think that 2-4-6 is not necessarily a well-chosen example of the rule to be discovered? If their behaviour stems from their overconstraining the problem because of their knowledge of the properties of "good" examples, then it should be modified by the knowledge that the example provided is not necessarily a "good" one.

How could participants be led to think that 2-4-6 is not necessarily a well-chosen example of the rule? The experimental instructions for the task usually stipulate that the experimenter "has a rule in mind" and that the triple 2-4-6 conforms to the rule. They do not, however, specify the ways in which the experimenter selected the rule or the initial triple. Hence, it is natural for the participants to assume that the experimenter has chosen the rule, and chosen the triple to be a good example of the rule (with the usual properties described above). But, if the rule and the triple are said to have been "randomly drawn" from sets of rules and triples, then there is no implied guarantee about the quality of the initial triple.

In the real world, when somebody is trying to discover a natural rule, two possibilities exist. Either the rule is generated by a natural process (such as a natural law), or it has been

chosen by somebody (as in a parlour game). Given a rule to be discovered, whatever its provenance, then another problem concerns the origin of the initial instance of the rule. Once again, there are two possibilities: either the initial instance is picked at random from all the possible examples, or a human being has selected it with a particular intention in mind.

In the standard 2-4-6 problem, the instructions imply to the participants that both the rule and the example have been selected by the experimenter (i.e. with a goal in mind). In our view, the different ideas that participants have about the selection of the rule and its initial instance should induce different strategies for formulating and testing their hypotheses. When the participants think that both the rule and the initial instance have been selected at random, they should have no reason to prefer one possible rule to another, i.e. they should have no reason to focus on one rule rather than another. Hence, they should try to reject hypotheses with the goal to keep only the correct one at the end. It follows that the participants should use a strategy of disconfirming their hypotheses more often in this condition than in the standard one.

One potential problem concerns the respective roles of the rule and the initial triple in eliciting a strategy of disconfirmation. To answer this question we used three experimental conditions: the instructions stated that 1. both the rule and the triple were randomly selected, 2. only the rule was randomly selected, and 3. only the initial triple was randomly selected. The control condition used the instructions from the original classic version of the task.

For the sake of simplicity, we told the participants that the rule was selected before the triple. A similar effect should occur if the opposite were true: if the triple is selected first, it is the selection of the rule that should have an effect. On the one hand, if the triple is chosen by a presumably co-operative experimenter, then it should be a "good" example, whatever the rule or its origin. On the other hand, if the triple is picked at random, there is no reason to assume that it is a "good" example of the previously selected rule, even if this rule has been carefully chosen by the experimenter. Thus, a demonstration of an effect of the way the rule and the triple were selected would show that the classic confirmative strategy is caused, at least for a part, by the participants' expectations induced by the human origin of the rule.

Before to set out our experiment, we have to explain the way we traced the participants' strategies. The mere fact that participants try a triple that conforms to the hypothesis that they are testing does not show that they are trying to confirm the hypothesis. There are, at least, two lines of reasoning supporting this view.

First, as Klayman and Ha (1987, 1989) and Klayman (1995) pointed out, if the rule is more specific than the hypothesis being tested, the hypothesis can be rejected only by testing a triple that is an instance of the hypothesis but not of the rule. For example, if the rule is "three

consecutive even numbers" and the participant tests the hypothesis "three increasing even numbers" using the triple 4-8-10, then the feedback "no" from the experimenter disconfirms the participant's hypothesis. It is only because of the particular relation between the participant's hypothesis, which is very specific, and the rule, which is very general, that this strategy turns out to be ineffective in Wason's original task.

Second, a reason calling for a reconsideration of the categorisation of participants' trials is the nature of the feedback, *not that they get* but that they expect (Wetherick, 1962). Along this way, four trial types can occur, according to (1) whether the participant attempts to confirm or to disconfirm the actual hypothesis and (2) whether the triple is consistent or inconsistent with the actual hypothesis (Caverni, Rossi & Pérès, 2000). When the participant expects a positive feedback of a triple consistent with the hypothesis, or when he expects a negative feedback of a triple inconsistent with the hypothesis, then the participant attempts to confirm the actual hypothesis. At the opposite, when the participant expects a negative feedback of a triple consistent with the hypothesis, or when he expects a positive feedback of a triple inconsistent with the hypothesis, then the participant attempts to disconfirm the actual hypothesis. The four trial types are summarised in Table 1. Indeed, participants do not necessarily always expect a specific feedback. They have even been known to propose triples without having any particular hypothesis in mind to test (Tukey, 1986). As long as such cases are rare, however, they do not vitiate an analysis of performance.

< Insert Table 1 about here >

Experiment

The purpose of the experiment was to show that when participants are led to consider that the initial 2-4-6 triple and/or the to be discovered rule have been picked up at random, then they use more disconfirmative trial than they classically do.

Method

Participants

Eighty-three volunteers who were psychology undergraduates at the University of Burgundy in Dijon, France participated in the experiment.

Design

Participants were given the 2-4-6 Wason's task (1960). They had to discover a rule by proposing triples. The experimenter presented an initial triple "2-4-6", which he stated was an instance satisfying the rule that the participants had to discover. Each participant was randomly assigned to one of the four experimental groups, with 20 participants in each group. Two of the groups were told that the rule was "in the experimenter's mind", and two of the groups were told that the rule was "drawn at random by the experimenter from the set of possible rules concerning triples of numbers". One of each pair of these groups was told that 2-4-6 triple was "satisfied the rule" and other group in each pair was told that the triple was "drawn at random from the set of all the triples satisfying the rule". Thus, the design manipulated two variables: the status of the rule and the status of the triple, to yield four groups. Participants wrote down each triple they declared aloud, their hypothesis, the reply they were expecting from the experimenter, and the experimenter's actual reply, all on the same sheet of paper. The experimenter answered "yes" if the proposed triple conformed to the rule, and "no" if it did not. During the experiment, the experimenter did not know what the participants wrote down, but each tested triple. The participants continued the task until they either discovered the correct rule or gave up. Three independent judges scored the participants' protocols. The scoring was done according to the principles described above (see Table 1).

Results

Table 2 presents the percentages of the four sorts of trial, depending on whether the triple was an instance of the participants' hypothesis and expectation about feedback, in each of the four experimental groups. We rejected a total of 3 participants' data because their triples could not be classified in a uniform way by the three judges.

< Insert Table 2 about here >

Number of hypotheses

The mean number of hypotheses tested by each participant was not reliably affected by the randomness of the rule, or the randomness of the initial triple, or by an interaction between the two (all three F ratios <1).

Expected feedbacks

The percentage of trials on which the participants expected negative feedback was 30% (10% were triples that were counterexamples to the hypothesis to which the participants expected negative feedback, and 20% were triples that were examples of the hypothesis to which the participants expected negative feedback). The percentage was reliably higher when the rule was said to be selected at random than when its provenance was unstated (36% vs. 24%, $F(1,76) = 4.86$, $MSe = 624.35$, $p < .03$) (Table 3). The provenance of the initial triple, i.e., whether or not it was selected at random ($F(1,76) = 2.35$) and interaction between the two factors ($F(1,76) = 0.38$) did not reliably modify the percentage of triples on which the participants expected negative feedbacks.

< Insert Table 3 about here >

It should be to note that, when the participants tried to disconfirm, they did it almost exclusively expecting negative feedback with triples that were positive instances of their hypotheses (Table 2). Only 1.5% of the participants' trials were attempts to disconfirm a hypothesis using triples that were counterexamples to the hypothesis under test. In other words, when participants do try to falsify their hypotheses, they are 13 times ($20 / 1.5$) more likely to do so by expecting negative feedback to an instance of their hypothesis than by testing a triple that violates their hypothesis.

Randomness of rule and triple

The origin of the initial triple has a reliable effect on the percentages of attempts to confirm hypotheses (71% when the triple chosen at "random" as opposed to 88% when its origin is unspecified, $F(1,76) = 12.26$, $MSe = 4.49$, $p < .001$) (Table 4). Neither the origin of the rule nor the interaction between these two factors had a reliable effect on the confirmation rate (both $F < 1$).

Discussion

When the experimenter attributed a random origin to the initial triple, the participants were more likely to try to disconfirm their hypotheses. That means that when participants can find good reasons to consider alternative hypotheses (i.e. not to focus on the more salient

properties of the initial triple), they do so. The origin of the rule that they have to discover, however, does not modify this rate. The results fit nicely with our "goodness-of-the-example" hypothesis. If the participants have reasons to believe that the initial triple contains only relevant properties, then manipulating factors that induce this belief may improve performance. On the other hand, since we told participants that the rule was picked out before the triple, the way the rule was chosen has no relation to the goodness of the triple, and so the participants' performance should not be affected by the status of the rule.

The present results can be considered with those reported by Gigerenzer, Hell, & Blank, (1988), which showed that random sampling of a description is crucial to the participants' internal representation of a problem as it lead them to use or to neglect bases rates in probability revision. Our results must also be considered with those reported by Van der Henst, Rossi, & Schroyens (2002), who designed the 2-4-6 problem in such a way that there was no presumption of relevance accompanying the triple 2-4-6 and showed that participants performed better when the salient characteristics of the 2-4-6-triple resulted from a random procedure (a jackpot) than when a presumption of relevance accompanied such a triple (as in communication). However, our results differ from the previous ones from at least one point of view: when both Gigerenzer et al. and Van der Henst et al. argued that random sampling is effective when performed and observed by participants, we showed that random sampling may be effective too when only declared to participants: so a "pragmatic" way is not necessary.

Our findings about what can determine confirmative vs. disconfirmative strategies must be also compared with those obtained by Paolo Legrenzi himself. He showed that confirmatory behaviour arises from the constraints of the task and the particular social situation in which it is performed (cf. Butera & Buchs, in this book). Butera, Legrenzi & Oswald (1997a, 1997b) showed that, when hypothesis testing takes place during situations of social confrontation, a minority source induced defocussing, and a majority source induced focussing.

We end with some reflections on the shape of reasoning as it has been traced in our experiment. When the participants were asked to express what feedback they expected, they expected negative feedback on a fair proportion (30%) of their trials . It therefore seems wrong to assume that when participants do not have to express their expectations, they always expect positive feedback. Hence, the usual coding in terms of "confirmation" or "disconfirmation", which is based on this implicit assumption, may not reflect reality.

A further problem with the usual coding of the task is that the only pattern it calls "disconfirmation" is a negative example with an (assumed) positive expected feedback. This pattern is the least common in our data (1%), whereas the participants' preferred way to disconfirm was a triple satisfying their current hypothesis but with a negative expected feedback (20%).

An apparent paradox in our data is the fact that the confirmation rate is the same as the rate usually obtained using the classical coding (about 80%). This result seems rather puzzling: to take the participants' expected feedback into account reveals previously unnoticed disconfirmations, and so the confirmation rate should be lower in our experiment than in usual studies. The comparison, however, is not appropriate, because the procedure in which the participants state their expected feedback also defines differently some trials as confirmation instead of disconfirmation, i.e., in the case of a triple that is a counterexample to the current hypothesis but the participants expect negative feedback).

What do we know about the shape of reason? We know only what the tool we use to study reason reveals about it! If the tool is wrong, the results are misshapen. There is only one way to try to avoid such mistakes: researchers must search systematically for alternatives. Paolo has long recognized this principle, which may explain why Paolo he has influenced so many psychologists and made so many friends!

References

- Amstrong, S.C, Gleitman, L., & Gleitman, H. (1983). What some concepts might not be. Cognition, 13, 263-306.
- Butera, F., Legrenzi, P., & Oswald, M. (1997a). Is context a bias? Swiss Journal of Psychology, 56, 59-61.
- Butera, F., Legrenzi, P., & Oswald, M. (Eds.) (1997b). Contexts and biases. Bern: Hans Huber.
- Caverni, J.-P., Rossi, S., & P  ris, J.L. (2000). The alternatives taken into account in hypotheses testing: Two new paradigms for investigating strategies In J.A. Garcia-Madruga, N. Carriedo, & M.J. Gonzalez-Labra (Eds.), pp. 133-141. Mental Models in Reasoning. Madrid: UNED.

- Evans, J. St B. T. (1983). Selective processes in reasoning. In J. St B. T. Evans (Ed.), Thinking and reasoning: Psychological approaches. London: Routledge and Kegan Paul.
- Gigerenzer, G., Hell, W. & Blank, H. (1988). Presentation and Content: The use of Base Rates as a Continuous Variable, Journal of Experimental Psychology: Human Perception and Performance, 14, 513-525.
- Giroto, V., & Politzer, G. (1990). Conversational and world-knowledge constraints in deductive reasoning. In J.-P. Caverni, J.-M. Fabre & M. Gonzalez (Eds), Cognitive biases. Amsterdam: North Holland
- Gorman, M. E. (1986). How possibility of error affects falsification on a task that models scientific problem-solving. British Journal of Psychology, 77, 85-96.
- Gorman, M. E., & Gorman M. E. (1984). Comparison of disconfirmatory, confirmatory and control strategies on Wason's 2-4-6 task. Quarterly Journal of Experimental Psychology, 36A, 629-648.
- Gorman, M.E. (1995). Hypothesis testing. In S.E. Newstead & J. St B. T. Evans (Eds), Perspectives on Thinking and Reasoning. Hove (UK) & Hillsdale (USA): Lawrence Erlbaum.
- Grice, H.P. (1989). Studies in the way of words. Cambridge (Mass.): Harvard University Press.
- Johnson-Laird, P.N. (1983). Mental models. Cambridge (UK): Cambridge University Press.
- Klayman, J. (1995). Varieties of Confirmation Bias. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), Decision Making from the Perspective of Cognitive Psychology. New York: Academic Press.
- Klayman, J., & Ha, Y-W. (1987). Confirmation, disconfirmation and information in hypothesis testing. Psychological Review, 94, 211-228.
- Klayman, J., & Ha, Y-W. (1989). Hypothesis testing in rule discovery: Strategy, structure and content. Journal of Experimental Psychology: Learning, Memory and Cognition, 15, 596-604.
- Legrenzi, P., Giroto, V. & Johnson-Laird, P.H. (1993) Focussing in reasoning and decision-making, Cognition, 49, 37-66.
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. Quarterly Journal of Experimental Psychology, 24, 326-329.
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. Quarterly Journal of Experimental Psychology, 30, 85-96.

- Poletiek, F. H. (2000). Hypothesis-testing behaviour. Hove, UK: Psychology Press.
- Rossi, S., Caverni J.-P., Girotto, V. (2001). Hypothesis testing in a rule discovery problem : when a focused procedure is effective. The Quarterly Journal of Experimental Psychology, 54A (1), 263-267.
- 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., Warner, W.J., & Pliske, D.B. (1980). Strategies of rule discovery in an inference task. Quarterly Journal of Experimental Psychology, 32, 109-124.
- Van der Henst, J.B., Rossi, S, & Schroyens, W. (2002). When participants are not misled they are not so bas after all: A pragmatic analysis of a rule discovery task. In W.D. Gray & C Schunn (eds), Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Assicates.
- Wason, P.C. (1959). The processing of positive and negative information. Quarterly Journal of Experimental Psychology, 11, 92-107.
- Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12, 129-137.
- Wason, P.C. (1965). The contexts of plausible denial. Journal of Verbal Learning and Verbal Behavior, 4, 7-11.
- Wetherick, N.E. (1962). Eliminative and enumerative behaviour in a conceptual task. Quarterly Journal of Experimental Psychology, 14, 246-249.

Table 1

The four trial types to confirm or disconfirm

TEST	<i>EXPECTED FEEDBACK</i>	STRATEGY
Positive hypothesis test (+HT)	Positive (+FB)	Confirmation
Negative hypothesis test (-HT)	Negative (-FB)	Confirmation
Positive hypothesis test (+HT)	Negative (-FB)	Disconfirmation
Negative hypothesis test (-HT)	Positive (+FB)	Disconfirmation

Table 2

The mean percentages of the trial types produced by participants in each of the four experimental groups.

		<i>EXPERIMENTAL GROUPS</i>				
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	
		<i>RULE</i>				
		<i>UNS</i>		<i>RAND</i>		
		<i>TRIPLE</i>				
STRATEGY	TRIAL TYPE	UNS	RAND	UNS	RAND	MEANS
<i>Confirmation</i>	+HT/+FB	82 .23	67 .24	66 .27	60 .26	69
	-HT/-FB	5 .12	8 .14	22 .22	6 .12	10
<i>Disconfirmation</i>	+HT/-FB	13 .17	22 .18	11 .23	33 .27	20
	-HT/+FB	0 0	3 .06	1 .02	1 .03	1

Group 1: The way both the rule and the initial triple were selected was unspecified (UNS)

Group 2: The way the rule was selected is unspecified; the initial triple was presented as selected at random (RAND)

Group 3: The way the rule was selected was said "at random"; the way the initial triple was selected was unspecified

Group 4: The way both the rule and the initial triple were selected was said "at random"

+HT: the proposed triples (T) were positive (+) examples of the proposed hypotheses (H)

-HT: the proposed triples (T) were negative (-) examples of the proposed hypotheses (H)

+FB: the expected feedback (FB) were positive (+) feedback

-FB: the expected feedback (FB) were negative (-) feedback

Table 3

Means percentages of trials on which the participants expected negative feedback (-FB), both when they generated a negative instance of their hypothesis (-HT) and when they generated a positive instance of their hypothesis, in the four sorts of group.

TRIAL TYPE	<i>RULE</i>				<i>MEANS</i>
	UNSC		RAND		
	UNSC	RAND	UNSC	RAND	
-HT/-FB	5	8	22	6	10
+HT/-FB	13	22	11	33	20
<i>MEANS</i>	24		36		