Does Positivity Bias Explain Patterns of Performance on Wason's 2-4-6 Task?

Maggie Gale (m.gale@derby.ac.uk)
University of Derby
Western Road, Mickleover, Derby DE3 9GX, U.K.

Linden J. Ball (l.ball@lancaster.ac.uk)
Department of Psychology, Lancaster University,
Lancaster, LA1 4YF, U.K.

Abstract

In the standard form of Wason's (1960) 2-4-6 task, participants must discover a rule that governs the production of sequences of three numbers. Studies typically show success rates of approximately 20%, which Wason attributed to a cognitive deficit that he labeled 'confirmation bias'. In Tweney et al.'s (1980) formally equivalent Dual Goal (DG) form of the task, however, success rates are at least double to those seen on the standard task. If this facilitated performance could be accounted for, then this would go some way toward explaining the normally low performance on the standard problem. The present experiment examined two competing accounts of the DG superiority effect: Evans' (1989) positivity bias explanation, and Wharton, Cheng and Wickens' (1993) goal complementarity theory. The experiment independently manipulated the number of goals that participants had to explore (a single goal vs. two complementary goals) and the linguistic labels used to provide feedback (DAX and MED vs. 'fits the rule' and 'does not fit the rule'). Results supported the goal complementarity account in that facilitation was evident in both DG conditions irrespective of the polarity of the feedback provided. We also discuss a novel finding: that it is the production of at least a single 'negative' triple that is most closely associated with task success.

Introduction

Poletiek (2001) summarises hypothesis testing as comparing internal thoughts with external facts in order to interact with the world. For example learning a language can be characterised as hypotheses testing as the learner utters sounds and observes the listener's reactions. Hypothesis testing, therefore, can be viewed as a fundamental mode of mental functioning, and for this reason is of considerable interest to psychologists and cognitive scientists alike.

One important experimental paradigm that has been employed extensively in order to study hypothesis-testing behaviour is the 2-4-6 task, introduced by Peter Wason in 1960. The 2-4-6 task is a deceptively simple rule discovery task, which Wason originally devised to investigate whether people conformed to the contemporary scientific norm of hypothesis testing, namely falsification (Popper, 1959). In the standard version of the 2-4-6 task, participants seek to discover a rule which generates sequences of three numbers (referred to as *triples*). They are initially given an example, conforming triple (2-4-6), and are then required to produce further triples which the experimenter classifies as either conforming to, or not conforming to, the rule. The to-be-discovered rule is 'any ascending sequence'. Participants produce triples until they are confident that they know the rule, at which point they announce it. Despite the seeming simplicity of the task, participants perform poorly, with typically only around 20% correctly announcing the rule on the first attempt, (e.g., Tukey, 1986; Wason, 1960; Wharton, Cheng & Wickens, 1993). Many of these incorrect announcements are a more restricted version of the rule, for example, 'numbers increasing by two'. It has been suggested (e.g., Wetherick, 1962) that the initial 2-4-6 exemplar lures participants into formulating such overly-restricted hypotheses. Participants then produce triples motivated by these hypotheses (e.g., 8-10-12), which always receive positive feedback, since they form a subset of the target rule. Faced with repeated confirmations of their hypothesis, participants seemingly become increasingly confident of its correctness until they announce it as the rule. It is clear that unless participants change their testing strategy they will never discover that although their hypothesis is sufficient, it is not necessary.

In his analysis of participants' performance on the task, Wason showed that solvers and non-solvers could be differentiated in terms of both the number of triples they produced (with solvers producing reliably more triples), and the type of triples generated (with solvers producing a higher proportion of triples which received negative feedback). Wason viewed the non-solvers' strategy of testing positive instances of their hypothesised rule as a cognitive failing, which he labeled 'confirmation bias'. However, Klayman and Ha (1987), in an elegant conceptual analysis of the underlying structure of the 2-4-6 task and its variants, demonstrated that it is the *relationship* of the hypothesised rule to the target rule in the original task which causes participants' failure to discover the target rule. Klayman and Ha ague that what Wason regarded as a bias to seek confirmatory evidence could instead be conceptualised as a 'positive test strategy', which in certain circumstances is an effective method for yielding disconfirmations of a current hypothesis. Their essential point (cf. Wetherick's, 1962, argument noted earlier) is that in the standard task, the target rule ('any ascending sequence') has been deliberately designed to be more general than the hypothesis invited by the given triple, such that the application of a positive test strategy can never lead to the discovery of the target rule. For other target rule/hypothesis relationships, however, such as where the experimenter's rule (e.g., 'even numbers ascending by two and less than 10') is more restricted than the participant's initial hypothesis (e.g., 'numbers increasing by two'), then the implementation of positive testing would lead rapidly to falsification of the overly general initial hypothesis, and to accurate rule discovery.

Dual Goal Instructions

Tweney, Doherty, Worner, Pliske, Mynatt, Gross, and Arrkelin (1980) introduced a modified form of the task, in which participants were instructed to discover two rules, one called DAX, the other MED. The DAX rule governs triples of the traditional ascending type, all other triples are MEDs. Although formally equivalent to Wason's original task, this simple Dual Goal (DG) manipulation was seen to have a dramatic effect on success rates, with 60% of participants making a correct first announcement of the rule. This facilitated performance has been shown to be a robust finding that has been replicated many times (e.g., Farris & Revlin, 1989a, 1989b; Tukey, 1986; Wharton, Cheng & Wickens, 1993). Tweney et al., were at a loss to explain the facilitatory effect of the DG manipulation, although they felt that the explanation was somehow related to the way participants conceptualise the task, and how triples produced are related to their conceptualisation.

It has been noted that the DG manipulation has the effect of increasing the *number* of triples which are generated before rule announcement, and also the *variety* of triples (Gorman, Stafford & Gorman, 1987; Tukey, 1986; Tweney et al., 1980). Vallée-Tourangeau, Austin and Rankin (1995), in their replication and extension of DG instructional effects, formulated two measures of triple heterogeneity, namely *posvars* and *negtypes.* Posvars are triples which receive positive feedback but which do not increase by a constant. Thus, if the numbers that make up a triple are *a, b,* and *c*, a posvar is a triple in which *(b-a)*   (*c-b*). Negtypes are the eight possible types of triples which receive negative feedback, (e.g., descending triples, identical-number triples, etc.). Vallée-Tourangeau et al. found that using these indices of triple heterogeneity, the DG manipulation led to increased production of both posvars and negtypes compared to Single Goal (SG) instructions. They interpreted this as being indicative of participants considering a wider range of hypotheses, although they did not directly test this claim. Whilst these observations of triple heterogeneity are interesting, they are largely descriptive, and do little to explain the facilitatory effect of DG instructions.

Evans (1989) proposed that poor performance on the standard task could be attributed to the operation of a general 'positivity bias', which is a form of selective processing that causes people to attend to positive rather than negative information. According to this proposal, facilitated performance using DG instructions is caused by the labeling of triples that 'do not fit the target rule' as MED, thereby avoiding a negative label, and hence counteracting participants' tendencies not to attend to this information. Evans (1989) argues that in the standard version of the 2-4-6 task, if a participant forms the hypothesis 'numbers ascending by equal intervals is right', they have logically also formed the hypothesis 'numbers not ascending by equal intervals is wrong'. They are, however, not aware of this alternative hypothesis, and therefore do not test it. In the DG manipulation, however, because participants are attempting to discover two rules, they test both DAX (correct) triples, and MED (incorrect) triples. Participants are more successful with the DG instructions since by carrying out positive tests of their MED hypotheses they are effectively carrying out negative tests of their DAX hypotheses, thus eliminating the overly restrictive hypotheses typically announced by the SG non-solvers. In summary, then, Evans argues that DG instructions facilitate performance by changing participants' representation of the task by creating a positive label for the previously negative 'does not fit' feedback.

Wharton et al. (1993) proposed a subtly different mechanism by which DG instructions improve performance. They invoke Klayman and Ha's (1987) proposal that a central feature of hypothesis testing behaviour is a tendency for individuals to adopt a positive test strategy which leads to the generation of triples that match their hypothesised rule. As we noted earlier, in the standard 2-4-6 task positive testing will never enable participants to discover the overly restrictive nature of their hypothesis as they will never generate a triple which lies outside of their hypothesis yet is still within the experimenter's target rule. With DG instructions, however, even though the exemplar triple for DAX suggests the same restricted hypothesis, the requirement to discover the second (MED) rule should encourage participants to form a second hypothesis (e.g., 'numbers ascending by intervals other than two' are MED). On carrying out a positive test of the MED hypothesis, (e.g., 5-10-15) participants will (unexpectedly) receive DAX feedback, thus causing them to alter both their DAX and MED hypotheses. This sequence of events is repeated until satisfactory rules for DAX and MED are discovered. Thus, according to Wharton et al., it is the *complementary* nature of the two rules that leads to task success.

It is clear that Evans' (1989) positivity-bias account and Wharton et al.'s (1993) goal-complementarity account make very different predictions regarding performance on the 2-4-6 task. Evans' theory predicts that participants given positively-labeled DAX and MED feedback in relation to generated triples will perform better than those given a combination of 'fits the rule' (positively labeled) and 'does not fit the rule' (negatively labeled) feedback, and that this dissociation will be present irrespective of whether participants are asked to discover a single target rule or two complementary rules. In contrast, the goal-complementarity account proposed by Wharton et al. predicts that participants given the task of discovering two complementary rules will be more successful than those seeking a single rule, regardless of whether feedback is given as DAX/MED or 'fits'/'does not fit'. Previous studies of the facilitatory effect of DG instructions have always confounded these two variables, such that participants given DG instructions have always been given DAX/MED feedback, whilst those given SG instructions have always received 'fits'/'does not fit' or 'yes'/'no' feedback. The present experiment was designed to discriminate between the positivity-bias account and the goal complementarity theory, by manipulating these two factors independently. To this end, the participant's goal (i.e., discovery of one rule vs. discovery of two complementary rules) was systematically crossed with the linguistic label of the feedback (i.e., DAX/MED vs. 'fits the rule'/'does not fit the rule').

Method

Participants

Sixty undergraduates of varying backgrounds and ethnicity from the University of Derby took part in the experiment in exchange for course credits. They had not received any teaching on the psychology of reasoning before the experiment.

Design

A fully between participants design was employed that manipulated two factors: Goal (Single Goal vs. Dual Goal), and Linguistic Labeling (DAX/MED vs. Fits/Does Not Fit). Fifteen participants were randomly assigned to each of the four resulting conditions.

Procedure

Participants were tested in groups of up to four in a quiet laboratory. Standardised instructions were read to each group. Single Goal (SG) instructions referred to a unique rule: *'I have in mind a rule that specifies how to make up sequences of three numbers (triples), and your task is to discover this rule'.* In what we subsequently refer to as the SG—Fits condition, participants were asked to discover the target rule by generating triples which they would be told either 'fitted' or 'did not fit' the rule that the experimenter had in mind. On the other hand, in what we refer to as the SG—DAX condition, participants were told that triples that fitted the rule were called DAX triples and those that did not fit the rule were called MED triples. It was explained to participants that on generating a triple they would be informed as to whether it was a DAX or a MED triple.

The Dual Goal (DG) instructions emphasised that there were two rules to be discovered: '*...your task is discover this rule, and also a second rule for categorising the triples that do not fit my rule*'. In the standard DG task (i.e., DG—DAX), participants were additionally informed that triples that fitted the rule were called DAX triples and those that did not fit the rule were called MED triples. They were instructed to produce further triples, which the experimenter would describe as either DAX or MED. In the DG—Fits condition, participants were told to generate triples which the experimenter would classify in terms of whether they 'fitted' or 'did not fit' the rule.

Participants in all conditions were given 2-4-6 as the example triple. All participants were provided with an answer sheet and were asked to write 2-4-6 on the first row, and either 'fits' or DAX in the feedback column, as appropriate. They were instructed that they could produce as many triples as they wished, and that when they were sure of the rule(s) they should write it (or them) on the answer sheet. In line with Gorman (1992), participants were allowed only one guess at the rule(s).

Results

Success Table 1 shows the frequency of correct and incorrect announcements by participants in each of the four experimental conditions.

Table 1: Frequency of correct announcements by condition.

Condition	N	Solvers	Non-Solvers
SG—DAX	15	3	12
SG—Fits	15	3	12
DG—DAX	15	12	3
DG—Fits	15	11	4

Table 2: Mean number of triples (and type of triples) produced by condition.

Condition	Total Triples	Posvars	Feedback	Negtypes
SG—DAX	7.6 (6.25)	0.33 (0.62)	0.73 (1.28)	0.33 (0.62)
SG—Fits	5.87 (2.75)	0.53 (1.36)	0.93 (1.58)	0.80 (1.42)
DG—DAX	10.27 (6.30)	1.13 (1.06)	2.67 (1.95)	1.20 (0.86)
DG—Fits	8.33 (3.22)	1.07 (1.10)	1.4 (1.24)	0.93 (0.80)

Note: SD in parenthesis.

Labeling of feedback (DAX/MED vs. Fits/Does Not Fit) seems to have had little effect on the likelihood of success; the proportions of solvers were similar between both the two SG groups and also the two DG groups. However more than three times the number of participants in the DG conditions than in the SG conditions announced the correct rule. A contingency table chi-square analysis was performed on the frequencies of correct and incorrect announcements (pooled over the labeling of feedback), and revealed a highly significant effect of the SG versus DG manipulation, $\chi^2 (1) = 19.288$, $p < .001$.

Feedback Analyses were also performed to ascertain whether the manipulated factors had any effect on the type or number of triples produced (see Table 2). Again, the labeling of the feedback appeared to make little difference to the number or type of triples produced by participants. With regard to the SG versus DG instructions, however, there were significant main effects on three of the measures: number of triples produced, F(1, 56) = 4.09, p < .05; number of triples receiving negative feedback, F(1, 56) = 9.1, p < .01; and number of variable positive triples, F(1, 56) = 5.86, p < .05. The difference in the number of negtypes produced across SG and DG conditions also approached significance, F(1, 56) = 3.96, p = .052. There were no significant interactions for any of the measures.

Presence of Triple Types Although the analyses of triple type are interesting they do not give insight into the absolute importance of the production of the triple types. For this reason, it was decided to carry out further analyses in which the production of either a posvar or a negative triple was crossed with success on the task. In this way it would be possible to test whether the production of such triples is necessary for task success.

Table 3: Frequency of correct announcements by production of at least one posvar.

	Solvers	Non-Solvers	Total
Posvar produced	25	10	35
No Posvar produced	6	19	25
Total	31	29	60

A contingency table was, therefore, produced in which the production of *at least one* posvar was crossed with success (see Table 3). The table clearly demonstrates that the production of a single posvar is associated with success on the task, with four times the number of participants who produced a posvar making a correct announcement compared to those who did not produce one. A chi-square analysis confirmed the reliability of this observation, $\chi^2(1) = 13.137$, p < .001.

Table 4 shows a contingency table in which the production of at least one negative triple is crossed with success. Here the association is even more marked than in the case of the production of at least a single posvar, with there being only one instance of a participant who had not produced a negative triple correctly announcing the rule. In contrast, of the 34 participants who did produce a negative triple, 28 solved the task. A chi-square analysis revealed that these differences were highly significant, $\chi^2(1) = 36.363$, p < .001).

Table 4: Frequency of correct and incorrect announcements by production of a negative triple.

	Solvers	Non-Solvers	Total
Negative triple present	28	6	34
Negative triple absent	1	25	26
Total	29	29	60

Discussion

The results of the present experiment clearly support the goal complementarity account (Wharton et al., 1993) of the facilitatory effect of DG instructions on the 2-4-6 task. Evans' (1989) positivity-bias account, on the other hand, fails to find support in the evidence presented. The results show that DG superiority cannot be attributed to the re-labeling of negatively valenced 'does not fit' feedback as positive 'MED' feedback, as participants in the DG conditions performed significantly better than participants in the SG conditions, regardless of the nature of their feedback. This leads to the conclusion that the typically poor success rates on the standard form of the task cannot be accounted for by participants selectively attending to positive information and thus ignoring a potentially informative set of triples. In relation to this point, the analyses of triple type and triple production show that participants in the DG condition produced a greater number and variety of triples. It could, therefore, be argued that it is not that SG instructions lead to selective processing of negative information, but rather that SG instructions do not promote the exploration of negative information in the first place (cf. Wharton et al., 1993).

The final set of analyses also revealed a hitherto unremarked phenomenon. It has long been noted that people who solve the 2-4-6 task tend to produce more triples as well as a greater proportion of negative triples (Wason, 1960). It has also been demonstrated more recently that solvers generate a greater variety of triples (e.g., Vallée-Tourangeau et al., 1995). What has not previously been shown, however, is that it is the production of at least a *single* negative triple that is so closely associated with success on the task. Indeed it remains possible that other indices of success such as the total number of triples produced or overall triple variety may well be mediating factors through which the critical negative triple is produced as a result of task manipulations. This is an area which would seem to require closer investigation.

The basic observation that negative-triple production is so closely related to task success, does, at first sight, appear rather paradoxical The point is, that given the typically overly-restrictive hypotheses which participants form, it seems intuitively obvious that it should be the production of the discriminatory *posvars* (rather than negative triples) that would be most strongly associated with task success. Although our results do indicate that posvar generation is significantly linked to correct initial rule announcements on the task, it remains striking that the production of negative triples is even more predictive of task success. Why might this be the case?

One possibility is that the production of a *descending* triple (and its associated MED or 'does not fit' feedback) somehow makes the general dimension of *ascending numbers* appear to be relevant to the target DAX or 'fits' rule. The concept 'descending' may have this effect by facilitating the establishment of a salient contrast class that promotes an insight into the potential scope of the target rule. Closer investigation of the precise role of negative triples in facilitating task success - perhaps through the invocation of clear contrast sets within the space of possible triples - would, therefore, appear to be essential. To achieve this a finer-grained system of codifying the triples that participants produce may be required.

In summary, the results of this study clearly support a goal complementarity account of facilitated performance using DG instructions on the 2-4-6 task. Participants in the DG conditions were more successful at the task than those in the SG conditions. The lack of effect with regard to the labeling of feedback would appear to undermine a standard positivity-bias account. Further work, however, is vital to understand the role that negative triples play in determining task success.

References

Evans, J. St. B. T. (1989). *Bias in human reasoning*: *Causes and consequences.* Hove: Lawrence Erlbaum Associates, Inc.

Farris, H., & Revlin, R. (1989a). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory and Cognition, 17,* 221-232.

Farris, H., & Revlin, R. (1989b). The discovery process: A counterfactual strategy. *Social Studies of Science, 19,* 497-513.

Gorman, M. E. (1992). Experimental simulations of falsification. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking: Vol. 1*. Hemel Hempstead: Harvester Wheatsheaf.

Gorman, M. E., Stafford, A., & Gorman, M. E. (1987). Disconfirmation and dual hypotheses on a more difficult version of Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology* 39A, 1-28.

Klayman J., & Ha, Y-W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Poletiek, F. H., (2001). *Hypothesis-testing behaviour*. Hove: Psychology Press.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

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., Gross, K. A., & Arrkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology, 32,* 109-123.

Vallée-Tourangeau, F., Austin, N. G., & Rankin, S. (1995). Inducing a rule in Wason's 2-4-6 task: A test of the information-quantity and goal-complementarity hypotheses. *Quarterly Journal of Experimental Psychology, 48A,* 895-914.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology, 14,* 129-140.

Wharton, C. M., Cheng, P. W., & Wickens, T. D. (1993). Hypothesis-testing strategies: Why two goals are better than one. *Quarterly Journal of Experimental Psychology, 46A,* 743-758.