Goal-Driven Hypothesis Testing in a Rule Discovery Task

Frédéric Vallée-Tourangeau & Teresa Payton
Psychology Research Unit, Kingston University
Kingston-upon-Thames UNITED KINGDOM KT1 2EE
f.vallee-tourangeau@kingston.ac.uk

Abstract
The Wason 2-4-6 task was embedded in a practical reasoning scenario where number sequences had well defined utilities in the process of achieving a goal. Reasoners’ hypothesis-testing behaviour was clearly goal-driven and was significantly influenced by whether the utilities favoured positive or negative sequences. In the version of the scenario where generating positive sequences had greater benefits than generating negative ones, participants performed poorly at the task as measured by their ability to guess the correct rule and by the nature and number of triples tested before making an announcement. In contrast, the scenario that assigned a greater utility to the production of negative sequences fostered significantly more diligent and creative hypothesis testing behaviour, and participants were significantly more likely to discover the rule. These results suggest that the poor performance observed in Wason’s traditional 2-4-6 task reflects a hypothesis-testing process that by default assigns greater epistemic utility to the production of sequences that conform to the initial triple, and hence receive positive feedback. However, participants are not averse to producing negative sequences, and understand their implication, if their epistemic utility is made relevant in the process of achieving goals.

Introduction
The Wason (1960) 2-4-6 rule discovery task provides an interesting window onto hypothesis-testing behaviour. In this simple task, participants are instructed to discover the rule that governs how sequences of three numbers are put together. To discover the rule, participants produce new number sequences, or triples, which are classified by the experimenter as either conforming or not to the rule. The to-be-discovered rule is ‘ascending sequence’. Before participants produce their first triple, they are provided with a crucial but misleading piece of information: The triple ‘2-4-6’ conforms to the rule. In its transparent orderliness, this initial example naturally lures participants to formulate hypotheses such as ‘evens increasing in twos’ that are too specific relative to the target rule. That is, sequences that conform to these initial hypotheses also conform to the broader ‘ascending sequence’ rule. Hence a positive test strategy (Klayman & Ha, 1987) leads participants to produce number sequences that receive positive feedback, strengthening the plausibility of these initial hypotheses. And this is indeed what tends to happen: Participants generate few sequences that receive positive feedback and rarely test the scope of these initial hypotheses by formulating number sequences (e.g., ‘1-9-43’) that would produce important disconfirming information. As a result, few participants discover the correct rule, even though instructions stress that they should announce a rule only when they feel highly confident. Nearly 80% of Wason’s (1960) participants failed to announce the ‘ascending sequence’ rule on their first attempt, a finding that has been consistently replicated (e.g., Vallée-Tourangeau & Krüsi Penney, 2005; Gale & Ball, 2006).

The hypothesis-testing behaviour of unsuccessful reasoners in this task can be characterised in terms of two important features. The first is the relative lack of diligence exerted before announcing a best guess, participants testing on average five or fewer sequences. The second is the nature of the number sequences tested: Participants tend to produce sequences that explore but a very narrow range of possibilities, failing to generate a more informative sample of number sequences. In this simple inferential reasoning task, just as in science (e.g., Feyerabend, 1987), no hypothesis-testing strategy guarantees success. However, the participants who announce the correct rule tend to exhibit more diligence and more creativity, that is they generate a greater number of triples that exhibit a greater degree of heterogeneity, producing a sample of observations that provide a better platform from which to discover the ‘ascending sequence’ rule (Vallée-Tourangeau, Austin, & Rankin, 1995).

Epistemic Utility
The hypothesis testing performance in Wason’s 2-4-6 task suggests that reasoners ascribe much greater importance to generating ascending sequences than to generating non-ascending ones. In other words, the epistemic utility (cf. Evans, Manktelow, & Over, 1993) of ascending sequences is perceived to be much greater, and not just any ascending sequences, but rather those that reflect the algebraic specificity of the initial example. There is no hard and fast method that guarantees success in the 2-4-6 task, but a narrow exploration of the space of possible triples is unlikely to provide information to spark discovery. If participants considered that non-ascending sequences had intrinsic value, it might encourage them to produce more such triples. The broader the exploration of the triple space, hence the more heterogeneous the sample of triples tested, the more likely participants are to hit upon key informative contrasts, such as the one between ascending sequences that receive positive feedback and
the rule that characterises all ascending sequences, and the second, often labelled Med, is the rule that applies to all non-ascending sequences. Participants test number sequences in the usual manner, but instead of obtaining positive or negative feedback, the triples are labelled Dax or Med, respectively. Dual-rule instructions have a substantial impact on diligence, creativity and success at discovering the ascending sequence rule. These instructions encourage the production of a greater number of triples as well as a greater number of ‘negative’ or Med triples which enhances participants’ ability to home in on the correct rule for Daxes. The success of the dual-rule manipulation cannot be simply reduced to a labelling phenomenon (Gale & Ball, 2006). As Tweney et al. suggested, the effect reflects a deeper change in the participants’ “entire conceptualization of the problem” (p. 121).

Dual-rule instructions work so well because reasoners’ efforts to discover the Med rule naturally lead them to construct a broader and more informative sample of triples from which to base their inferences. In effect, we propose that dual-rule instructions enhance the epistemic utility of non-ascending or Med triples, and this change in perceived utilities is a key component of the new ‘conceptualization’ that is fostered with dual-rule instructions. Dual-rule instructions elicit interest in negative or Med sequences and reasoners value their production. As a result, they produce a larger more heterogeneous set of number sequences from which they formulate their best guess.

The study presented here is motivated by the idea of creating a 2-4-6 task isomorph wherein the utility of ascending and non-ascending triples is manipulated explicitly (see Vallée-Tourangeau & New, 1998). In effect, we sought to couch a 2-4-6 rule discovery task in a scenario that emphasizes either the utility of producing positive sequences or the utility of producing negative sequences. Since the production of a broader more informative set of triples fosters the discovery of the ‘ascending numbers’ rule we conjectured that a task scenario that encouraged participants to generate negative sequences would enhance the likelihood of discovering the rule.

The scenario elaborated for this experiment was one where number sequences had explicitly defined benefits. In the scenario, number sequences were codes that gained access to information (a ‘database’). Reasoners’ goal was to access the database in order to extract information deemed important in the context of the task scenario. In one version, database access was possible with sequences that met the ‘ascending numbers’ rule, while in another access was possible if the sequences did not meet the rule. In both versions of the task, the reasoner’s aim was to discover the ‘ascending numbers’ rule, and thus at this level of analysis the task was identical to Wason’s 2-4-6 task. However, the task scenario enabled the manipulation of the utility of different kinds of number sequences. We predicted that reasoners for whom the utility of generating sequences that received negative feedback was greater than the utility of generating sequences that received positive feedback would (i) work longer at the task, (ii) generate a more heterogeneous and more informative set of number sequences, and hence (iii) be more likely to discover the ascending numbers rule.

Method

Task Scenario

Wason’s (1960) 2-4-6 task was presented in terms of a scenario inviting participants to imagine working as undercover police officers. In that role they were investigating a company that imported toys but were acting on a tip that the company directors were also using certain incoming toy shipments to bring in drugs. Participants were informed that the directors kept a ‘restricted entry’ database on the company’s computer that split into two files, one that contained the company’s legitimate transactions (which was of little interest to the undercover operation) and the other secret information that pertained to the directors’ illegal drug trafficking operations (which was of great interest to the undercover operation). The task scenario continued: To access the database of interest, a code was requested. Participants were instructed that this code was made up of a sequence of three numbers and was based on a rule; this rule was crucial in determining to which of the two files (the legitimate business file or the drug trafficking file) a user was directed upon entering three numbers.

At this point, the utility of the different types of number sequences was manipulated. In the first version of the scenario participants read:

“If the code input meets the rule, you are directed to the legitimate toy business data file that you are not interested in.”
Thus in this scenario, the utility of generating sequences that did not conform to the rule implied by the initial triple was greater than the utility of generating sequences similar to the initial example. That is, these instructions were designed to encourage participants to generate number sequences that did not conform to the pattern suggested by the initial sequence, ‘2-4-6’, since such sequences provided access to desired information. In the second version of the scenario, the utility assignment was reversed such that sequences similar to the initial ‘2-4-6’ example, sequences that met the rule, enabled participants to access the desirable database, whereas sequences that did not conform to the rule gained them access to the undesired database. In this version of the scenario, participants should exert relatively little effort generating sequences that receive negative feedback.

In addition to manipulating the utility of different types of number sequences, a second independent variable was designed to provide an added incentive to discover the rule. One version of the scenario specified that participants had to discover the rule to gain access to the desired database, while in the other version access had already been gained, but out of curiosity, participants sought to discover the secret code. The instructions read:

“You have managed to rig up one of the director’s computer so that when he logged out of the secret data base on Friday night, the secret file was kept open. In the low incentive condition, access to the database was already secured:

(….)After having successfully copied all the information you require for your investigation, you are curious and want to discover the rule behind the entry code. You reboot the computer to return to the database entry screen. To try to discover the rule, you have just started inputting sequences of three numbers to see which files in the database they direct you to.”

In the high incentive condition, access to the database was not secured:

However, as you expected, you have been ejected from the file and returned to the database entry screen because a code was not entered after the first two minutes. To regain entry to the secret file and then to stay there long enough to obtain the critical information required for your investigation, you need to discover the rule behind the coding. To try and discover the rule, you have just started inputting triples to see which files in the database they direct you to and then if the input codes allow you to remain there.

There were thus four different versions of the scenario, reflecting the factorial combination of two independent variables, namely whether conforming or non-conforming sequences were valued, and whether the incentive to discover the rule was high or low.

**Experimental Design and Procedure**

Participants were assigned randomly to each of the four experimental conditions. Participants were run one at a time in a quiet room. An instruction sheet outlining the task scenario was placed in front of them, which they read at their own pace. Once participants’ questions were answered they were given a response sheet on which they wrote new number sequences. The answer sheet had four columns and 21 un-numbered rows. Each number that made up a new sequence tested by the participant was entered in each of the first three columns, while the fourth was reserved for the feedback, which the experimenter wrote on the answer sheet. In all four conditions the first row of the answer sheet showed the example ‘2-4-6’. The labels of the feedback column differed as a function of the epistemic condition. In the two conditions where the production of non-conforming sequences was valued, the labels read “YES: Satisfies the rule (access legitimate toy data)/NO: Does not satisfy the rule (access drug trafficking data)”. The labels were reversed in the two conditions where the production of conforming sequences was valued, namely “YES: Satisfies the rule (access drug trafficking data) /NO: Does not satisfy the rule (access legitimate toy data).”

Participants were instructed to tell the experimenter their guess when they felt “highly confident” they had discovered the rule. Participants had only one opportunity to guess the rule, and if they were incorrect, they were not invited to continue. Participants were not asked to formulate a hypothesis before every triple they tested.

**Measures**

The proportion of participants announcing the correct ascending numbers rule was calculated in all four conditions. In addition, the participants’ hypothesis-testing performance was measured in a number of ways. First their diligence was gauged in terms of the number of triples generated. Their creativity was measured by calculating the number of triples that received positive and negative feedback. Using a, b, and c to represent the first, second, and third number of a triple, the proportion of positive triples that did not exhibit a constant increment, that is where \((b - a) \neq (c - b)\), was also calculated (we termed those triples ‘variable positives’). Finally the different types of non-ascending sequences were calculated: In addition to a descending triple where \(a > b > c\), there are 7 different types of non-descending triple, namely (i) \(a > b < c\), (ii) \(a < b > c\), (iii) \(a = b = c\), (iv) \(a = b > c\), (v) \(a = b < c\), (vi) \(a > b = c\), and (vii) \(a < b = c\).

**Participants**

One hundred and twenty two Kingston University undergraduates (69 males, 53 females) volunteered as
participants and were randomly allocated to one of the four experimental conditions, 30 were assigned to the negative valued/low incentive condition, 31 to the positive valued/low incentive condition, 31 to the positive valued/high incentive condition, and 30 to the positive valued/high incentive condition.

**Results**

The number and percentage of participants correctly announcing the ascending numbers rule as well as the mean number of triples and the mean number of the different kinds of triples are reported in Table 1.

**Correct Announcement**

In the two conditions where the utility of generating non-conforming sequences was emphasized, 20 participants announced the correct rule, 12 (or 40%) in the low incentive condition, and 8 (or 26%) in the high incentive condition, the difference between incentive levels was not significant, $\chi^2(1, N = 61) = 1.39, p > .05$. In the two conditions where the utility of generating conforming sequences was emphasized, 8 participants announced the correct rule, 5 (or 16%) in the low incentive condition, and 3 (or 10%) in the high incentive condition. Again the difference across incentive levels was not significant, $\chi^2(1, N = 61) = 0.50, p > .05$. Collapsing across incentive conditions, there were significantly more participants who announced the ascending numbers rule (20 out 61 or 33%) in conditions where non-conforming sequences were valued than in conditions where positive sequences were valued (8 out of 61 or 13%), $\chi^2(1, N = 122) = 6.67, p < .01$.

**Triples**

A series of 2 (utility emphasis) by 2 (incentive level) between-subjects analysis of variance (ANOVA) were conducted on the number and kind of triples.

**Overall Number of Triples.** Participants in the two conditions that emphasised the utility of producing positive triples appeared to produce more number sequences before announcing their best guess ($M = 8.18$) than participants in the conditions that emphasised the utility of producing positive triples ($M = 5.90$). In turn the incentive manipulation within each utility condition did not appear to have influenced diligence to the same extent. The 2x2 ANOVA confirmed these observations. The main effect of utility was significant, $F(1, 118) = 8.73, MSE = 18.45, \eta^2 = .07, p < .005$, but neither the main effect of incentive, $F(1, 118) = 2.39, MSE = 18.45, \eta^2 = .02, p > .05$ nor the interaction, $F(1, 118) = 1.07, MSE = 18.45, \eta^2 = .01, p > .05$, were significant.

**Number of Positive Triples.** Participants in all four groups produced a similar number of triples that received ‘yes’ feedback, means ranging from 4.35 to 5.17. In the 2x2 ANOVA, none of the main effects were significant, all $F$s < 1.

**Number of Negative Triples.** Participants appeared to have tested more triples that received ‘no’ feedback in the two conditions that emphasised the utility of testing negative sequences ($M = 4.83$ and $M = 2.84$, in the low and high incentive conditions, respectively) than in the two conditions that encouraged testing positive triples ($M = 1.74$ and $M = 0.53$ in the low and high incentive conditions, respectively). In the 2x2 ANOVA the main effect of utility was significant, $F(1, 118) = 22.71, MSE = 9.78, \eta^2 = .16, p < .001$ as was the main effect of incentive, $F(1, 118) = 8.00, MSE = 9.78, \eta^2 = .07, p < .01$; however the interaction was not significant, $F < 1$.

**Proportion of Variable Positive Triples.** Participants in the two conditions that emphasised the utility of negative sequences may have tested the same number of positive triples as the participants in the conditions that emphasised the utility of positive sequences, however they were more likely to produce positive triples that increased in variable or non-constant increments. The mean proportion of variable positive triples was 31.3% in

### Table 1: Rule discovery performance using scenarios wherein negative or positive triples are valued and wherein the incentive to discover the rule is high or low in terms of number of participants who announced the correct rule (and in percent), mean number of triples before announcements, number of triples that received positive and negative feedback, percentage of variable positives and number of negative types (SEM = standard error of the mean).

<table>
<thead>
<tr>
<th></th>
<th>Correct Announcements</th>
<th>Triples</th>
<th>Positive Triples</th>
<th>Negative Triples</th>
<th>% Variable Positives</th>
<th>Negative Types</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Incentive: Low</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>n</td>
<td>%</td>
<td>M</td>
<td>SEM</td>
<td>M</td>
<td>SEM</td>
</tr>
<tr>
<td>30</td>
<td>12</td>
<td>40%</td>
<td>9.20</td>
<td>0.87</td>
<td>4.37</td>
<td>0.65</td>
</tr>
<tr>
<td>31</td>
<td>8</td>
<td>26%</td>
<td>7.19</td>
<td>0.74</td>
<td>4.35</td>
<td>0.44</td>
</tr>
<tr>
<td>61</td>
<td>20</td>
<td>33%</td>
<td>8.18</td>
<td>0.58</td>
<td>4.36</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Incentive: High</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>n</td>
<td>%</td>
<td>M</td>
<td>SEM</td>
<td>M</td>
<td>SEM</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>16%</td>
<td>6.10</td>
<td>0.76</td>
<td>4.35</td>
<td>0.44</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>10%</td>
<td>5.70</td>
<td>0.73</td>
<td>5.17</td>
<td>0.69</td>
</tr>
<tr>
<td>61</td>
<td>8</td>
<td>13%</td>
<td>5.90</td>
<td>0.52</td>
<td>4.75</td>
<td>0.41</td>
</tr>
</tbody>
</table>
the conditions where negative sequences were valued but just 11.6% in the conditions where positive sequences were valued. A 2x2 ANOVA confirmed these observations: The main effect of utility was significant, $F(1, 118) = 13.61, MSE = .088, \eta^2 = .10, p < .001$ but the main effect of incentive was not, $F < 1$, nor was the interaction, $F < 1$.

**Types of Negative Triples.** Participants in the conditions that encouraged the production of sequences that received negative feedback not only produced more triples that received ‘no’ feedback, they also appeared to produce more different types of negative triples ($M = 2.23$ and $M = 1.48$ in the low and high incentive conditions, respectively) than participants in the conditions that encouraged the production of positive triples ($M = 0.77$ and $M = 0.50$ in the low and high incentive conditions, respectively). Looking at the four means, the low incentive conditions seemed to have encouraged a greater number of different types of negatives than the high incentive conditions. In a 2x2 ANOVA, the main effect of utility was significant, $F(1, 118) = 22.75, MSE = 1.92, \eta^2 = .17, p < .001$ as was the main effect of incentive, $F(1, 118) = 4.17, MSE = 1.92, \eta^2 = .03, p < .05$ but the interaction was not significant, $F < 1$.

**Discussion**

In the original Wason task participants aim to discover the rule that govern sequences of three numbers by producing new sequences that receive ‘yes’ or ‘no’ feedback. Participants are given the sequence ‘2-4-6’ as an initial example. These key features also defined the Wason task isomorph developed for this experiment but in addition participants were encouraged to either generate sequences that conformed to the rule or that did not. In conditions where the utility of generating negative sequences was greater than the utility of generating positive sequences, participants were 2.5 times more likely to discover the ascending number rule. Their success was associated with two important factors that have been identified as predictors of discovering the ascending rule in the Wason 2-4-6 task (e.g., Vallée-Tourangeau et al., 1995; Vallée-Tourangeau & Krüsi Penney, 2005), namely a greater degree of diligence and creativity before announcing their best guess. Thus in the experimental conditions that emphasised the importance of producing sequences that receive negative feedback, participants generated a greater number of triples before announcing their hypothesis than participants in the conditions that stressed the importance of producing sequences that receive positive feedback. In addition, when the utility of negative sequences was stressed, participants produced a more heterogeneous sample of triples. This heterogeneity was reflected in a number of ways. First, a greater proportion of positive sequences that increased in variable increments characterised the sample of positive sequences; variable increasing sequences offer very important information to reasoners, acting to defuse the attractive but misleading feature of the initial triple ‘2-4-6’. Second, participants produced more negative sequences and a greater number of different types of negative sequences when their utility was greater than positive sequences.

The scenarios employed were also designed to enhance the effort to discover the rule by implying that access to the desired information was either already obtained or not, and hence in the latter case discovering the rule would provide an added incentive to successfully complete the task. As the data showed, this instructional manipulation had little impact on the participants’ hypothesis-testing performance. If anything there was a trend to suggest that the manipulation had the opposite effect. Thus, when the scenario cast the discovery of the rule as a puzzle to be solved out of ‘curiosity’ and the outcome of which did not shoulder the success of the fictional operation, participants appeared relatively more creative.

**Contextual Determinants of Sound Hypothesis Testing**

Discussions of the 2-4-6 task in the literature (e.g., Politiek, 2001) are based on the original Wason procedure and the hypothesis-testing profile drawn from this task is grim indeed: People invest little energy and creativity before announcing their best guess, and a large majority fail to discover the rule on their first attempt. It has been argued that the nature of the task is too unusual to provide an accurate window onto people’s hypothesis-testing abilities. For one, the ‘2-4-6’ initial example is, of course, misleading (Wetherick, 1962), and were the initial example any different the simplicity of the to-be-discovered rule becomes transparent (cf. Rossi, Caverni, & Girotto, 2001). Others have argued that the conversational pragmatics that underpin the exchange of information between experimenter and participant further encourage participants to believe that the initial example is particularly relevant to solving the task (Van der Henst, Rossi, Schroyens, 2002). Yet we would argue that the misleading nature of the initial example does not vitiate the importance of the task as a tool to explore hypothesis testing. After all, nature is not a kind tutor to the naive observer. Hypotheses are formulated to best explain facts, but sometimes they reflect an inappropriate conceptualisation of these facts.

The experiment reported here further illustrates that participants’ difficulty at discovering the rule in the original procedure is not simply the result of poor hypothesis testing skills, but rather is a function of the context of reasoning. In this experiment, the context was not altered by offering a richer representation of number sequences (cf. Vallée-Tourangeau & Payton, 2008) but rather by manipulating the utilities ascribed to the triples. In effect, we transformed the 2-4-6 task into a ‘practical
reasoning’ problem (Evans, Over, & Manktelow, 1993) wherein participants sought to achieve a certain goal. In one version of the task, this goal was best achieved by exploring negative sequences, and as a result, participants were more likely to discover the rule because they were better able to generate an informative sample of triples. As Evans et al. (1993) propose, “(…) reasoning in the real-world supports decision-making and is aimed at the achievement of goals” (p. 165). The results of this experiment demonstrate that it is possible to embed the ‘2-4-6’ rule discovery task in a context of reasoning that aligns the goal and utilities of reasoners with a sounder hypothesis-testing strategy, that is one where a larger set of triples are tested, and where triples exhibit important features, such as positive sequences that increase in variable increments, that better position reasoners to evaluate the scope of their hypotheses.

In light of the hypothesis-testing limitations documented with the original 2-4-6 task procedure, Wason (1960, p. 139) and others have enjoined researchers to identify the psychometric dispositions that favour creativity and discovery (e.g., Vartanian, Martindale, & Kwiatkowski, 2003). In contrast, our research efforts have focused on the characteristics not of the reasoners (and their shortcomings) but rather of the reasoning context (e.g., Vallée-Tourangeau & Krüsi Penney, 2005). In some of these contexts, people exhibit considerably more diligence and creativity than is observed in Wason’s traditional 2-4-6 task. These results call into question the representative nature of the hypothesis-testing behaviour observed with the original Wason task procedure.

Acknowledgements

This research was supported by a grant from the Economic and Social Research Council (UK) awarded to FVT.

References


