Interactive Graphical Representations Support the Formulation of Sound Hypotheses in a Rule Discovery Task

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

Abstract

In the 2-4-6 rule discovery task, reasoners seek to discover a rule that governs the arrangement of three numbers (or triple). The to-be-discovered rule is ‘any increasing sequence’. Upon being given the triple 2-4-6 as an initial example, however, reasoners are lured to formulate overly specific hypotheses. Traditionally, this task is conducted primarily from an internal representation of the triples and candidate hypotheses. More recently, substantial representational effects have been demonstrated wherein an external representation of the dimensions of the problem space facilitated successful rule discovery. In the current study, an interactive external representation was created by concurrently plotting each triple produced by the participants. Compared to a control group, participants who performed the task with this interactive external representation formulated sounder hypotheses which led to the production of more informative triples and to more successful rule discovery.

Introduction

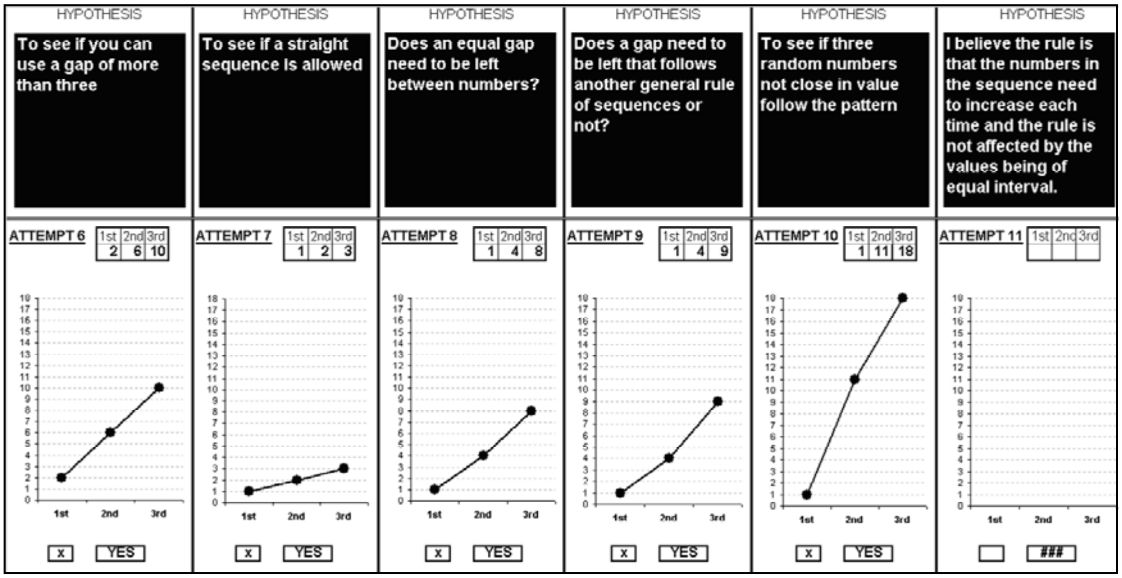
The importance of the physical presentation of a problem in fostering a productive representation distributed over the mind of the reasoner and the external appearance of the problem has been well documented (e.g., Simon, 1996; Zhang, 1997). Recent work on hypothesis-testing has explored the importance of the manner with which key elements of the reasoning task are physically presented. For example, Vallée-Tourangeau and Krüsi Penney (2005) examined hypothesis-testing behaviour in a simple inductive inference task where some of the computational cost underpinning the hypothesis-testing process could be delegated to the positioning of external artefacts manipulated by the participants. Vallée-Tourangeau and Krüsi Penney reported that such richer distributed representations fostered more productive and successful hypothesis-testing behaviour.

The research reported here examines hypothesis-testing behaviour using the 2-4-6 rule discovery task in conditions where the hypothesis-testing process is supported by a rich external representation of the problem. In the traditional 2-4-6 task (Wason, 1960), participants must discover a rule that governs the generation of sequences of three numbers (triples). The rule to be discovered is ‘any increasing sequence’. However, participants are informed at the start that the triple 2-4-6 satisfies the rule. Participants are instructed to generate new number sequences to test their hypotheses until they feel highly confident they have discovered the rule. The triple 2-4-6 encourages participants to formulate overly narrow hypotheses such as ‘even numbers increasing in twos’, ‘numbers increasing by a constant’, “third number = first number + second number”, that are too algebraically specific. Because the scope delineated by these hypotheses is much narrower than, and at times nested in, the one delineated by the correct hypothesis (viz., ‘any increasing sequence’), a simple positive-test strategy will unfailingly yield positive feedback (Klayman & Ha, 1987) from which participants draw growing confidence that they are on the right track. Yet, the vast majority of participants fail to announce the correct ‘any increasing sequence’ rule on their first attempt: 79% failed to do so in Wason (1960), a finding much replicated since (e.g., Tweney, Doherty, et al., 1980; Vallée-Tourangeau & Krüsi Penney, 2005).

The hypothesis-testing behaviour of participants engaging in the traditional 2-4-6 task is characteristically indolent and prosaic. That is, participants do not work very hard before announcing their best guess, testing on average five number sequences. Clearly, the need to work harder at discovering the rule is hampered by the abundance of positive feedback which participants invariably receive for these initial triples. Second, participants are not creative in the kinds of number sequences they produce and test, exploring a very narrow region of the space of possible triples. Again, in light of the abundance of positive feedback, participants experience little pressure to create more adventurous or unusual number sequences. There is no hard and fast method that guarantees success at this task (cf. Gorman & Gorman, 1984). It is clear, however, that those who do succeed exhibit, in relative terms, considerably more diligence and creativity than those who don’t, producing a significantly greater number of triples of a much broader variety before announcing their first guess.

Distributed Representation

This simple rule discovery task packs an important inferential challenge, namely that of mapping the scope and generalisability of hypotheses. In this respect, the 2-4-6 task is representative of real-world hypothesis testing. However, in its original formulation and many of its replications, the task is of limited ecological validity in that much of the hypothesis-testing behaviour is canvassed in the head, that is, it proceeds primarily from an internal representation of triples and possible hypotheses. To be sure, participants write number sequences on an answer sheet against which


HYPOTHESIS: To see if you can use a gap of more than three

HYPOTHESIS: To see if a straight sequence is allowed

HYPOTHESIS: Does an equal gap need to be left between numbers?

HYPOTHESIS: Does a gap need to be left that follows another general rule of sequences or not?

HYPOTHESIS: To see if three random numbers not close in value follow the pattern

HYPOTHESIS: I believe the rule is that the numbers in the sequence need to increase each time and the rule is not affected by the values being of equal interval.

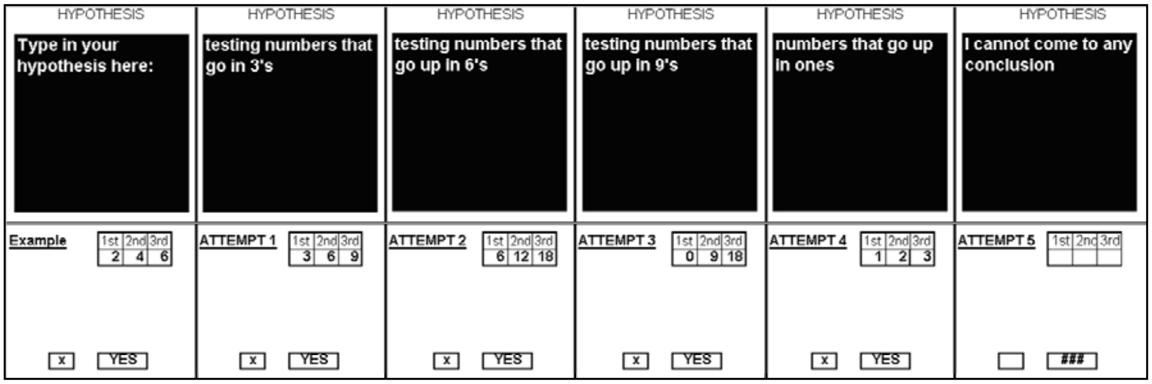
Figure 1: Actual protocol from Participant 2 in the Experimental Condition. The participant announced the correct rule after generating ten triples.

the experimenter affixes feedback, and in that respect there is an external record of past triples. Yet real-world hypothesis testing usually proceeds on the basis of a tight coupling of artefacts (cognitive or otherwise) and ideas. For example, the apparatus, measuring instruments and methodologies scientists employ to ply their trade encourage and constrain the formulation and test of certain kinds of hypotheses. Test results are also represented in graphical formats, some more likely than others to act as further catalyst of productive hypothesis formulation and testing (cf. Cheng, 1996; Reinmann, 1999).

Vallée-Tourangeau and Krüsi Penney (2005) examined the impact on hypothesis-testing behaviour of a richer distributed representation of the problem space in a 2-4-6 task isomorph. In this version of the task, sequences could be made of numbers ranging from 1 to 6. However, the triples were created by manipulating three traditional six-sided dice. Participants rotated the face of the dice or interchanged their order to produce new triples. A group of control participants engaged in the 2-4-6 task without the dice, but also with numbers ranging from 1 to 6. Even with this considerably reduced space of triples ($6^3 = 216$ triples) only 21% of the control participants announced the correct rule, in line with the first-announcement performance observed in the original Wason (1960). In contrast, 66% of the participants with the dice isomorph announced the correct rule. These participants produced more triples, of a more varied kind, before announcing their guess than control participants. It appeared that providing an external, manipulable, representation of the triple space made the number permutations perceptually salient and easier for reasoners to implement. The external environment was thus configured in a way that naturally encouraged diligence and creativity.

Vallée-Tourangeau, Krüsi Penney, and Payton (2005) examined the impact of creating a visual representation of the generated number sequences on hypothesis-testing behaviour in the 2-4-6 task. They created a task isomorph where numbers could range from 1 to 8, mapping out a space of 512 possible triples. Using a stack of small paper grids (4.5cm by 6.5cm) where the x axis was labelled 1^{st}, 2^{nd}, 3^{rd} and the y axis ranged from 1 to 8, participants plotted each new number sequence before producing the next one, creating a graphical record of the triples they tested. Compared to control participants who did not have access to such a graphical record of their tested triples, graph participants were more likely to announce the correct rule on first announcement.

The productive impact of the graphs on reasoning in the 2-4-6 task may be due to their ability to 'limit abstractions' (Stenning & Oberlander, 1995). Graphs offer perceptually transparent representations of the simple linear relationships between consecutive numbers. They also offer a medium other than strings of numbers with which to formulate new hypotheses, that is a medium where hypotheses can be expressed using more qualitative concepts as opposed to quantitative, numerical ones. It is thus plausible to suggest that participants in the graph condition of Vallée-Tourangeau et al. (2005) were freer to contemplate a range of hypotheses that were not constrained by the algebraic parameters implied by the initial 2-4-6 triple. However, Vallée-Tourangeau et al. did not ask their participants to explicitly formulate and write down a hypothesis for each triple generated. Thus, the impact of the external representation on the nature of the hypotheses formulated by participants remains conjectural.



HYPOTHESIS	HYPOTHESIS	HYPOTHESIS	HYPOTHESIS	HYPOTHESIS	HYPOTHESIS
Type in your hypothesis here:	testing numbers that go in 3's	testing numbers that go up in 6's	testing numbers that go up in 9's	numbers that go up in ones	I cannot come to any conclusion
Example	**ATTEMPT 1**	**ATTEMPT 2**	**ATTEMPT 3**	**ATTEMPT 4**	**ATTEMPT 5**
1st 2nd 3rd 2 4 6	1st 2nd 3rd 3 6 9	1st 2nd 3rd 6 12 18	1st 2nd 3rd 0 9 18	1st 2nd 3rd 1 2 3	1st 2nd 3rd
x YES	x YES	x YES	x YES	x YES	###

Figure 2: Actual protocol from Participant 7 in the Control Condition. The participant tested four triples and gave up.

The Present Study

The study reported here sought to determine the exact nature of the hypotheses entertained when participants engage in the 2-4-6 task with or without the benefit of an enriched distributed representation Participants engaged in a 2-4-6 task isomorph with numbers ranging from 0 to 18 specifying a space of over 6,000 possible triples. Before the formulation of each triple, participants were required to explicitly formulate a hypothesis. In addition, a number of important modifications to the procedure employed by Vallée-Tourangeau et al. (2005) were implemented, namely (i) the interactive nature of the graphical representation, (ii) the opportunity to continue after announcing an incorrect rule, and (iii) the computer-driven feedback. These improvements are discussed in turn.

First, the external representation was interactive in that it synchronously reflected the selection of numbers in the creation of a test triple. Thus, triples were plotted one number at a time as participants selected individual numbers. In Vallée-Tourangeau et al. the representation was created *after* having generated the triple. In contrast, then, participants in the present study could alter their number selection more dynamically on the basis of perceptual feedback.

Second, in the task employed in Vallée-Tourangeau et al., (2005), participants were permitted only one announcement. If they announced an incorrect rule, they were not given the opportunity to continue. Clearly announcing an incorrect rule may significantly alter the nature of the hypotheses entertained. Hence, in the present study, participants who announced an incorrect rule were invited to continue with the task, thus permitting an assessment of hypothesis-testing behaviour before and after a first announcement.

Third, in nearly all past versions of the 2-4-6 task, the experimenter provides feedback for each generated triple, a process that conspires to create a pupil-teacher communicative context that might enhance perceived accountability for triples produced (cf. Lerner & Tetlock, 1999). There is evidence to suggest that automating feedback using a computer-controlled version of the 2-4-6 task makes it substantially harder (e.g., Vanderhenst, Rossi, & Schroyens, 2002, Exp. 2). In the present study, the feedback was generated by the computer and hence eliminated participant-experimenter interactions for each triple generated.

The goal of the study was to determine whether an enriched, interactive external representation encouraged the formulation of a qualitatively different set of hypotheses, weakening the algebraic specificity constraints implied in the initial 2-4-6 triple.

Method

Task

Participants engaged in a task isomorphic with Wason's (1960) 2-4-6 rule discovery task where numbers could range only from 0 to 18. Their task was to discover the rule that governed the production of 'correct' sequences of three numbers. They did so by producing new sequences that would be categorised as satisfying, or not, the to-be-discovered rule. As in the original Wason task, participants were informed at the outset that the triple 2-4-6 was a number sequence that satisfied the rule.

Experimental Design & Procedure

Participants were assigned to an Experimental or a Control group on a random basis. Participants engaged in the task using a specially configured Excel worksheet split into a top half where participants typed in their hypotheses (see Fig. 1) and a bottom half where they entered new three-number sequences and then clicked on the feedback box to receive feedback. These halves were segmented with vertical dividing lines into separate columns each corresponding to a new hypothesis-testing attempt.

In the Experimental group, as participants entered a new triple in the bottom half of the worksheet, the number sequence was automatically and synchronously plotted on a

2-axis grid, where the x-axis coded the number position in the triple (first, second, or third) and the y-axis ranged from 0 to 18 (see Fig. 1). The task instructions in the Experimental group were as follows:

> *The present task consists in discovering why certain numbers go together in a sequence. To start you off, I can tell you that 2-4-6 is a sequence that satisfies the rule I have in mind. In order to discover my rule, you should produce new number sequences from 0 to 18 by typing your sequence into the relevant boxes. Before generating each number sequence please enter the reason for your choice in the blue box (type it in).*
>
> *These numbers will then be plotted on the individual graphs; to discover whether your sequence meets the rule input "x" into the bottom left hand box, the right hand box will then tell you whether your sequence meets the rule. You can produce as many or as few sequences as you wish, but proceed to tell me your best guess only when you feel highly confident that you have discovered the rule that I have in mind.*

In the Control group, the triples were not plotted, and hence participants proceeded to discover the rule in the absence of that external representation (see Fig. 2). The task instructions were the same as those for the Control group with the omission of the phrase referring to numbers being plotted. In both groups, participants who announced an incorrect rule were encouraged to continue and test new triples and make further announcements.

Measures

Hypothesis-testing behaviour was measured in three ways. First, the proportion of Experimental and Control participants who announced the correct hypothesis at the time of their first announcement and at any later announcements. Second, the number and type of triples tested. Specifically, how many received positive and negative feedback, and of those positive triples, how many displayed variable increments (e.g., '1-5-8') in contrast to those that displayed constant increments (e.g., '3-6-9'). The generation of such variable positives has been shown to correlate significantly with the ability to discover the correct rule (Vallée-Tourangeau & Krüsi Penney, 2005). Third, the number and kind of hypotheses produced by participants. Hypotheses were coded in terms of algebraic specificity. Hypotheses low in algebraic specificity were those that did not stipulate explicitly a specific algebraic rule governing the composition of a number sequence. Thus, hypotheses such as "all even numbers" "random order", "increasing sequence" were classified in the low algebraic specificity category. In turn, hypotheses such as "gap is 2", "numbers add to 12", "second number + first number = third number", "three times table", were classified in the high algebraic specificity category.

Participants

Fifty-six university undergraduates received course credits for their participation. Participants were assigned to either the Experimental (N = 28) or the Control (N = 28) group on a random basis. As a result, the Experimental group was composed of 28 females and no males, with an overall mean age of 21, while the Control group was composed of 27 females, and 1 male with an overall mean age of 20.

Results

In the analyses reported below a .05 rejection criterion was employed throughout unless otherwise indicated.

Correct Announcement

The frequencies, and proportions, of participants announcing the correct rule in both Experimental and Control conditions are reported in Table 1. On first announcement, 11 or 39.3% of the Experimental participants announced the correct rule while 4 or 14.3% of the Control participants did so. Of the 14 participants in the Experimental group that continued with the task beyond the first announcement, 10 or 71.4% discovered the rule. Of the 19 participants in the Control group who chose to pursue the task beyond the first announcement, 9 or 47.4% discovered the rule. Over all announcements, 21 or 75% of the Experimental participants discovered the correct rule, while 13 or 46.4% of the Control participants did so. Chi square analyses revealed that (i) a significantly greater number of participants in the Experimental group announced the correct rule on their first attempt than Control participants, $\chi^2(1) = 4.46$, N = 56; (ii) over all announcements a significantly greater number of participants in the Experimental group discovered the rule, $\chi^2(1) = 4.70$, N = 56; (iii) but that there was no significant difference in the rate of success beyond the first announcement $\chi^2(1) = 1.91$, N = 33.

Table 1: Number (*and percentage*) of participants who announced the correct rule on first announcement and over subsequent announcements.

	Announcement			
	First		Subsequent	
	Correct	Incorrect	Correct	Incorrect
Experimental	11 *39.3%*	17 *60.7%*	10 *71.4%*	4 *28.6%*
Control	4 *14.3%*	24 *85.7%*	9 *47.4%*	10 *52.6%*

Triples

The number and kind of triples generated by the participants are reported in Table 2. Leading up to the first announcement, Experimental participants tested slightly more triples of a slightly greater variety than Control participants, but none of the group differences were significant. Over subsequent announcements, Experimental participants tested fewer triples (means of 3.5 vs. 5.2, $t(31) = 1.78$, $p < .09$) and produced a significantly higher

TABLE 2: Hypothesis-testing profile of participants in the Experimental and Control groups as measured up to their first announcement and then over all other annoucements in terms of mean triples produced, mean types of triples (positive or negative), and mean percentage of positive triples that increased in variable increments (pos var). s. e. = standard error

	First Announcement								All Other Announcements							
			Triple Type								Triple Type					
	Triples		Positive		pos var (%)		Negative		Triples		Positive		pos var (%)		Negative	
	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
Experimental	6.7	0.9	5.3	0.7	23.1%	5.2%	1.4	0.4	3.5	0.4	2.6	0.3	50.5%	6.8%	0.9	0.2
Control	5.8	0.6	4.8	0.4	12.2%	4.4%	1.0	0.3	5.2	0.6	3.8	0.5	22.2%	4.4%	1.4	0.2

proportion of variable positive triples, 50.5% vs. 22.2%, *t*(31) = 2.73, than Control participants.

Hypotheses

Over all announcements participants generated on average 6.35 hypotheses in the Experimental group and 6.64 hypotheses in the Control group, a statistically non-significant difference. However, examining the relative proportion of the types of hypotheses, significant differences were observed (see Fig. 3). Leading up to the first announcement, the vast majority (84%) of the hypotheses offered by the participants in the Control group were classified as highly specific compared with 64% in the Experimental group. However, after announcing an incorrect rule, the nature of the hypotheses formulated changed significantly in both groups: Fewer specific hypotheses were formulated overall, although the proportion of algebraically specific hypotheses was lower in the Experimental group (34%) than in the Control group (56%). In a 2 (Groups: Experimental vs. Control) by 2 (Announcements: First vs. Subsequent) mixed analysis of variance, the main effect of Groups was significant, *F*(1, 28) = 14.7, as was the main effect of Announcement *F*(1, 28) = 12.7, but not the interaction (*F* < 1). The results of this analysis confirm that Control participants generated a greater proportion of hypotheses that expressed specific algebraic links between numbers than Experimental participants, and that in both groups, more such hypotheses were formulated leading up to the first announcement than over subsequent announcements.

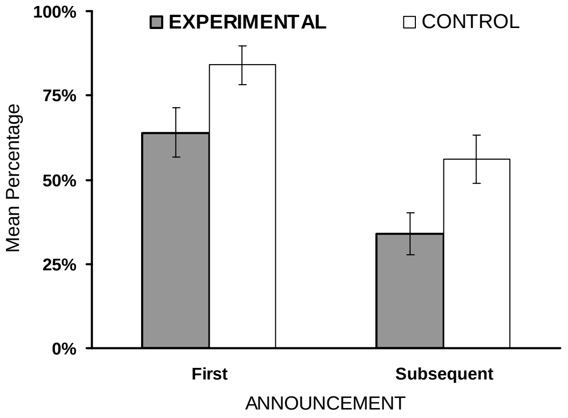


Figure 3: Mean percentage of hypotheses categorised as high in algebraic specificity up to the first announcement and over all subsequent announcements.

Discussion

In the present study participants engaged in a task isomorph of Wason's (1960) 2-4-6 task. The goal of the task was to discover the rule 'any increasing sequence' by using numbers ranging from 0 to 18. The resulting space of possible triples while still large, was considerably smaller than in the one specified in Wason's original task. Nonetheless, the task remained very hard. Of the 28 participants in the Control group who engaged in the representationally impoverished version of the task, only four announced the correct rule on their first attempt. Of the 19 participants who continued with the task, only 9 announced the correct rule over their subsequent attempts. Thus the majority of Control participants abandoned the task without discovering the correct hypothesis.

In the Experimental condition, each triple was automatically and synchronously plotted as participants entered the sequence of numbers. The graphical representation of the linear relationship among adjacent numbers helped participants discover the rule: 11 out of 28 announced the correct rule on their first attempt and of the remaining 14 participants who persevered after making an incorrect first announcement, 10 announced the correct rule over subsequent attempts. Thus 75% of the Experimental participants discovered the rule.

The graphical representations helped Experimental participants produce a greater proportion of varied positive triples, especially after a failed first announcement. Positive triples that increase in variable increments provide crucial evidence against a rule that specifies an invariant algebraic relationship among adjacent number, the very kind of rule that is naturally suggested by the initial triple 2-4-6.

The nature of the hypotheses formulated by Experimental participants also supports the contention that the graphical representation attenuated the tendency to think of algebraically specific candidate hypotheses. Over all announcements 70% of the hypotheses formulated by the Control participants identified a specific algebraic feature, whereas 55% of the ones formulated by the Experimental participants did so. The key to discovering the correct rule in the 2-4-6 task is to abandon the plausible algebraically-specific hypotheses suggested by the initial triple. It is no surprise that the Experimental participants did so much better in this 2-4-6 task isomorph.

Proceeding only on the basis of a numerical representation, reasoners in the Control group could but formulate hypotheses that were highly specific algebraically. Control participants only had numbers on which to anchor the creation of hypotheses. Numbers naturally invited formulaic recipes that would transform the first number of the triple into the second, and the second into the third. Given that the correct rule eschews such algebraic specificity, the representational medium of the task in the Control group was not conducive to discovering the correct rule.

In contrast, examining the nature of the hypotheses formulated in the Experimental condition suggests that the interactive graphical representation released participants from such a narrow numerical focus. The plotted lines encouraged the formulation of hypotheses that sought to capture the observed visual trend in simple non-algebraic terms, thereby more naturally converging on the correct ‘any increasing sequence’ rule. To be sure, reasoners more expert at curve fitting and calculus might have formulated, for example, hypotheses that mathematically described negatively or positively accelerating curves. Had the experimental sample included such participants, it remains uncertain whether the interactive graphical representations would have fostered such a high degree of successful rule discovery. Be that as it may, participants in either group never once formulated a hypothesis that could be considered an attempt at mathematically fitting a curve.

The results reported here lend further support to the contention that successful rule-discovery behaviour in the 2-4-6 task is to a significant extent determined by contextual and representational factors *external* to the reasoner. The hypothesis-testing profile drawn from the typically poor performance in Wason’s 2-4-6 task is textbook lore (e.g., Poletiek, 2001). Yet the results reported here paint a considerably more positive picture of performance in the 2-4-6 task. When the hypothesis-testing process is distributed over an internal representation of the task as well as a rich external representation of the triples, participants formulate more pertinent hypotheses and create a more informative set of number sequences that position them favourably to discover the correct rule. The fact that real-world hypothesis testing rarely proceeds solely on the basis of the reasoner’s internal representation of the problem and test results, calls into question the representativeness of hypothesis-testing behaviour as observed in the traditional version of the Wason 2-4-6 task.

Acknowledgment

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

References

Cheng, P. C.-H. (1996). Scientific discovery with law encoding diagrams. *Creativity Research Journal*, *9*, 145-162.

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.

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

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*, 255-275.

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

Reinmann, R. (1999). The role of external representations in distributed problem solving. *Learning and Instruction*, *9*, 411-418.

Simon, H. A. (1996). *The sciences of the artificial* (3rd Edition). Cambridge, MA: MIT Press.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: logic and implementation. *Cognitive Science*, *19*, 97-140.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, *32*, 109-123.

Vallée-Tourangeau, F., & Krüsi Penney, A. (2005). The impact of external representation in a rule discovery task. *European Journal of Cognitive Psychology*, *17*, 820-834.

Vallée-Tourangeau, F., Krüsi Penney, A., & Payton, T. (2005). Representational effects in a rule discovery task. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 2236-2241). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Vanderhenst, J.-B., Rossi, S., Schroyens, W. (2002). When participants are not misled they are not so bad after all: A pragmatic analysis of a rule discovery task. In W. D. Gray and C. D. Schunn (Eds.), *Proceedings of the Twenty Fourth Annual Conference of the Cognitive Science Society* (pp. 902-908). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, *21*, 179-217.