Rational Hypothesis-Testing Strategies in a Rule Discovery Task

Frédéric Vallée-Tourangeau (psyqfv@herts.ac.uk) and **Martin New**
Department of Psychology, University of Hertfordshire
Hatfield, Hertfordshire, UNITED KINGDOM, AL10 9AB

Abstract

In Wason's (1960) inductive learning task, subjects must discover a rule that governs the production of sequences of three numbers, such as '2-4-6', by generating new triples that receive feedback. Data obtained with Wason's original procedure suggest that people test few hypotheses before announcing their guess and mostly proceed on the basis of a positive-test strategy. These features are commonly regarded as lamentable aspects of reasoning agents who fail to appreciate normative models of hypothesis-testing. Such interpretations, however, are relative to the inferential context in which the behavior is observed. In the present study, Wason's original procedure was modified such that in one condition desirable consequences were associated with the production of positive exemplars and undesirable consequences with negative exemplars. In a second condition, the consequences were reversed. Subjects in the latter condition produced more exemplars, of a greater variety, and were more likely to discover the rule than subjects in the first condition. It seems then that in this second condition the hypothesis-testing strategy emerging from the subjects' appreciation of the cost and benefit of generating certain kinds of triples *coincided* with the normative strategy. However, since subjects in both conditions aimed to achieve different goals their hypothesis-testing strategies can, in that respect, be characterized as rational.

Wason (1960) originated a simple rule discovery task to assess how people test hypotheses and whether the process of doing so could be said to approximate the then popular prescriptive philosophy of science, namely Popperian falsificationism. In this task, subjects seek to discover a rule that governs the creation of number triples by producing new triples which the experimenter classifies as conforming or not conforming to that rule; the to-be-discovered rule is 'any increasing sequences'. In the original Wason procedure (cf. Klayman & Ha, 1989) subjects produce new triples until they feel confident they know the rule and then announce it to the experimenter. Before subjects generate their first triple they are told that the triple '2-4-6' conforms to the rule. On the basis of this initial example, subjects are naturally lured to believe that the rule involves even numbers increasing by a constant (Kareev, Halberstadt, & Shafir, 1993; Wetherick, 1962) and new triples motivated by this hypothesis (e.g., '10-12-14') will receive positive feedback. The likely initial hypothesis thus falls within the scope of the target rule (since all sequences of even numbers increasing by a constant are increasing sequences; Klayman & Ha, 1987). Should subjects seek to test this initial hypothesis by producing sequences of even numbers increasing by a constant they will unfailingly encounter positive feedback from the experimenter, bolstering their confidence in the hypothesis. In fact this initial hypothesis is sufficient to produce triples that receive positive feedback, but not necessary since a sequence such as '1-5-19' will as well. The nature of the task instructions, the initial triple offered as an 'example' to the subjects, the initial hypothesis it strongly implies, and the to-be-discovered rule together configure a certain inferential context in which human reasoning is observed.

In that inferential context Wason found that 80% of his subjects offered an incorrect guess for their first announcement. Wason found that solvers and nonsolvers (on the first announcement) could be demarcated in terms of how hard they worked, with solvers producing a reliably greater number of triples before making their first announcement than nonsolvers (see left panel of Figure 1), and solvers produced a greater variety of triples than nonsolvers as evidenced by the reliably greater proportion of triples that received negative feedback (see left panel of Figure 2).

Wason's findings are now textbook wisdom. For example, Sutherland (1992) writes: "Why is it difficult to find this simple rule? The main reason is that people try to prove that their current hypothesis is correct -they test it by picking only examples that will confirm it and do not look for ones that would disconfirm it." (p. 136) Schustack (1988) draws pessimistic implications of this so-called confirmation bias: "(...) to the extent that [Wason's data] characterize behavior outside the experimental setting (and there is much evidence that it does), all of us probably hold many erroneous beliefs for which we can adduce much evidence, convincing ourselves and others of generalizations that are at least overly narrow." (p. 110).

These sentiments illustrate two important misconceptions about the implications of the data obtained via Wason's original inferential context. First, seeking to disconfirm one's hypothesis is not sufficient to lead to discovery. For example, assume that a reasoner's current hypothesis is 'evens increasing by a constant (where the constant = 2)'. Following a strategy that aims to disconfirm the hypothesis, the reasoner could test '6-4-2' or '1-7-23', but the latter is clearly more informative. Thus, an appreciation of the logic of disconfirmation alone does not guarantee successful induction (Tweney, Doherty, Worner, Pliske, Mynatt, Gross, & Arkkelin, 1980). No hypothesis testing methodology can guarantee true inductive inferences, although a broad or creative exploration of the space of possible experimental manipulations (Klahr & Dunbar, 1988), or in this context, of the space of possible triples, helps.

Second, Schustack's characterization suggests that the positive test strategy (Klayman & Ha, 1987) exhibited by subjects in the Wason task has nefarious consequences. However, as Klayman and Ha (1987) demonstrated elegantly, an evaluative characterization of a reasoning 'strategy' is relative to an inferential context. In the original Wason

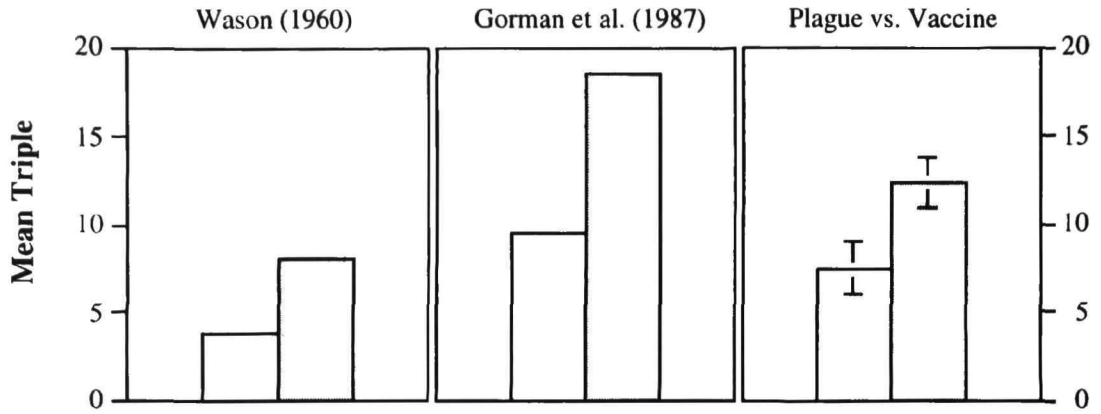

Figure 1. Mean number of triples produced by nonsolvers (white bar) and solvers (shaded bar) in Wason (1960) left panel, by subjects working with single-goal instructions (white bar) or dual-goal instructions (shaded bar) in Gorman et al. (1987) middle panel, and by subjects in the present study with the *Plague* scenario (white bar) and the *Vaccine* scenario (shaded bar) right panel.

procedure an unvarying positive test strategy might not have fruitful consequences. In other contexts, however, such a strategy might be the only one to adopt (e.g., if the to-be-discovered rule describes a subset of items characterised by the subjects' current hypothesis -see Klayman & Ha, 1987, Figure 3).

Textbook wisdom thus treats the original Wason inferential context as the canonical context from which to evaluate the consequences of the ways in which most subjects go about testing their hypotheses. There is no inherent virtue to the original Wason procedure that should grant it such a privileged status. Other inferential contexts offer equally legitimate perspectives from which to examine the consequences of hypothesis-testing strategies.

Dual-Goal Instructions (Dax-Med)

Tweney et al. (1980) modified the original Wason procedure by instructing subjects to discover two rules, a Dax rule that produced triples of the kind '2-4-6' and a Med rule. Dax triples are those that conform to the to-be-discovered rule in Wason's original procedure and Med triples to those that don't. This *dual-goal* (DG) manipulation (a term coined by Wharton, Cheng, & Wickens, 1993) transforms the usually unsuccessful, 'lazy', 'uncreative' triple generators, into successful, 'hard-working', 'creative' ones. That is, the rate of rule discovery is doubled and sometimes tripled with DG instructions; DG instructions encourage the production of a greater number of triples (see middle panel[1] of Figure 1) and a greater proportion of 'negative' or Med triples (see middle panel of Figure 2 -see Gorman, Stafford, & Gorman, 1987; Tukey, 1986; Tweney et al., 1980). In their replication of the DG manipulation, Vallée-Tourangeau, Austin and Rankin (1995) formulated two new indices of creative exploration of the triple space, namely *posvars* and *negtypes*. *Posvars* are triples that receive positive feedback and for which the increment between numbers is not constant. Thus if a, b, and c are the three number that make up a triple, a *posvar* is a positive triple for which $(b - a) \neq (c \quad b)$. *Negtypes* refer to the 8 possible types of triples that receive negative feedback[2]. Vallée-Tourangeau et al. found that the DG inferential context fostered a greater number of *posvars* compared with the traditional single-goal (SG) procedure, as well as a greater number of *negtypes* (left panel of Figure 3). Thus DG instructions foster a creative exploration of the triple space.

The originator of the DG inferential context were puzzled by the potency of the manipulation: "(...) the key to an explanation lies, we feel, in an understanding of the relation between the subjects' entire conceptualization of the problem at hand and the way empirical evidence is related to the components of that conceptualization." (Tweney et al., p. 121). Wharton, Cheng, and Wickens (1993) suggested that the DG effect hinged on subjects conceptualizing the Dax-Med rules as being complementary. However, explicit violations of the rules' complementarity in the task instructions do not mitigate the DG effect. For example in one of the conditions of Experiment 2 of Vallée-Tourangeau et al. subjects were given DG instructions but told that triples could be Dax, Med or neither. Changes in the conceptualization of the kinds of triples did not alter the beneficial effect of the DG instructions.

DG instructions encourage the production of 'negative' or Med triples. This should not be thought of as attempts to disconfirm Dax hypotheses, but rather as attempts at discovering the nature of the Med rule (Evans, 1989). In fact, a dual positive-test strategy seems to characterize hypothesis-testing in the DG inferential context. A by-product of this broader exploration of the triple space is a larger more informative sample of triples on which to base inferences and as a consequence reasoners are more likely to discover the (Dax) rule. The challenge has been to explain why in the DG manipulation subjects seek to explore an entirely new region of the space of triple, one populated by non-increasing or Med sequences. Such an explanation may

[1] The to-be-discovered rule in Gorman, Stafford, and Gorman (1987) was 'three different numbers'. The relationship between the to-be-discovered rule and the one implied by the initial triple '2-4-6' is the same as in the original Wason procedure however.

[2] There are eight possible patterns that produce a negative triple: 1. $a > b > c$; 2. $a = b = c$; 3. $a > b < c$; 4. $a < b > c$; 5. $a = b < c$; 6. $a = b > c$; 7. $a > b = c$; and 8. $a < b = c$.

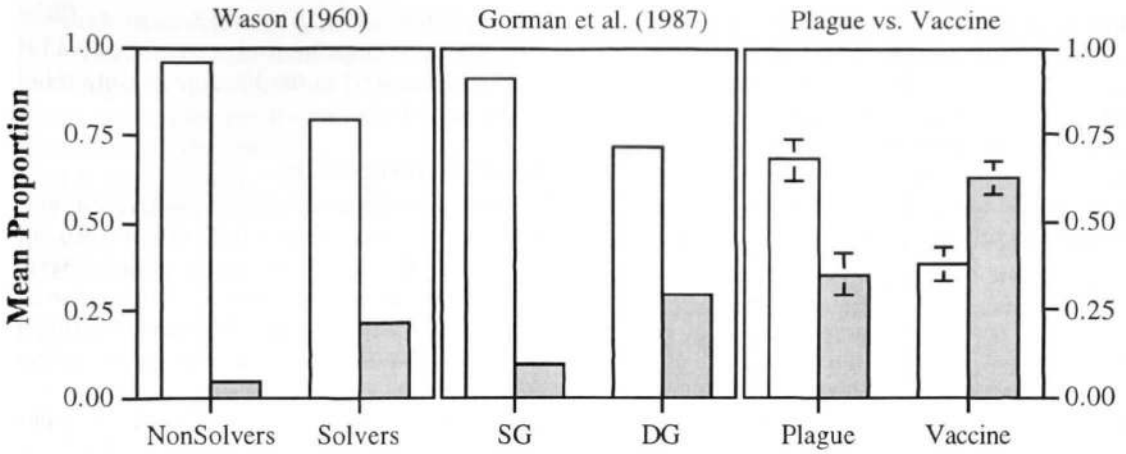

Figure 2. Mean proportion of positive (white bars) and negative triples (black bars) produced by subjects in Wason (1960) left panel, by subjects with SG and DG instructions in Gorman, Stafford & Gorman (1987) middle panel, and by the subjects in the two different scenarios used in this study, right panel.

be sketched in terms of the epistemic utility (cf. Evans & Over, 1996a) attached to both Dax and Med triples. In other words, with DG instructions, subjects become *interested* in both positive and negative triples (disguised as Daxes and Meds) and thus value their production.

Different regions of the triple space may vary in their perceived utility and this suggests that a decision-making perspective on this rule-induction task may yield new ways of characterizing hypothesis-testing behavior. Our aim in this study was to explore this idea by creating an inferential context where the underlying cost and benefits associated with the production of different kinds of number sequences were made more explicit. We created two inferential contexts: in one subjects were informed that the production of positive triples had more beneficial consequences than the production of negative triples, whereas in the other the utility assignment was reversed. Subjects were observed to favor the kinds of triples that had the highest benefits, and when those were the triples that received negative feedback, subjects were more likely to make the appropriate inductive inference.

Method

Subjects

Sixty-two undergraduates from the University of Hertfordshire received course credits for their participation.

Design & Procedure

A rule discovery task was used where subjects sought to infer the nature of a rule governing the production of sequences of three numbers. The rule was 'any increasing sequences'. Subjects were given the sequence '2-4-6' as an initial example that conformed to the rule. Subjects produced new sequences of numbers which were classified by the experimenter.

Two different inferential contexts were instantiated in two different scenarios. In the first scenario, the *Plague* scenario, desirable consequences were associated with the production of triples that received positive feedback but undesirable consequences were associated with the production of triples that received negative feedback. Subjects in this condition read the following instructions:

> "The plague sewer rats have invaded a major city. This breed of rats has developed immunity to all commercially available brands of rat poison. Health officials fear an outbreak of the plague if these rats are not exterminated.
>
> Chemist at the local university have isolated three chemical elements, call them P, Q, R, which when combined together kill the rats. Chemists however do not know the law that regulates which combinations of the three elements and in which quantity, but preliminary tests have shown that the following combination killed a captured sewer rat:
>
> | P | Q | R | OUTCOME |
> |---|---|---|---|
> | 2 | 4 | 6 | RAT KILLED |
>
> Your task is to test new combinations of the chemicals in any quantity you choose, in order to discover the general rule that determines which combinations of the chemical elements are lethal for the sewer rats. It is important to discover the rule to save as many people from contacting the plague as possible."

Thus in this scenario a 'positive' triple is one for which the combination of chemical elements kills the animal, and a 'negative' one is one which fails to do so. From the subjects' perspective positive and negative consequences mapped onto positive and negative triples respectively.

The second inferential context was instantiated in a scenario, the *Vaccine* scenario, which reversed the assignment of consequences to triple type such that desirable consequences were associated with negative triples and undesirable consequences with positive triples. Subjects in this condition read the following instructions:

> "The Teragglia parasite has infested most of the elephant calves in their natural habitat. The parasite causes fatal heart disease before the animals reach maturity.
>
> Researchers working on a treatment have identified three chemical elements, call them P, Q, R, which when combined together destroy the parasite, but

which in most tests kill the calf as well. Researchers believe that there exist combinations of chemicals whose interactions should be lethal only to the parasite and not the elephant. Researchers do not know the law that regulates which combinations of the three elements and in which quantity are lethal to both the parasite and the calf, but preliminary tests have shown that the following combination is lethal to both:

P	Q	R	OUTCOME
2	4	6	CALF KILLED

Your task is to test new combinations of the chemicals in any quantity you choose, in order to discover the general rule that determines which combinations of the chemical element are lethal for the calves. It is important to discover the rule to save as many of this endangered species as possible."[3]

As in the *Plague* scenario, a 'positive' triple was one for which the combination of the chemical elements killed the host organism and a 'negative' triples was one which did not. However, from the subjects' perspective positive consequences were associated with negative triples and negative consequences with positive triples.

In both conditions, subjects were instructed to produce new chemical combinations (number triples) until they felt confident they knew the rule that killed the rats/calves. They then wrote their answer at the bottom of the response sheet. At that point, the experimenter told them what the target rule was. Subjects were then debriefed.

Thirty one subjects were randomly assigned to each inferential context. Subjects were run individually in a quiet room.

Results

Success

Of the 31 subjects in the *Plague* condition, 8 discovered the rule, and 23 announced an incorrect rule. Of the 31 subjects in the *Vaccine* condition, 16 discovered the rule and 15 announced an incorrect rule. Thus twice as many subjects discovered the rule in the condition which attached desirable consequences to the production of negative triples. The difference was reliable, $\chi^2(1) = 4.35, p < .05$.

Triples

Subjects in the *Plague* condition produced an average of 7.5 triples (SE = 0.90) before announcing their guess while those in the *Vaccine* condition produced an average of 12.3 triples (SE = 1.48; see also the right panel of Figure 1); this difference was reliable, $F(1, 61) = 7.47, p < .009$.

The mean proportion of triples that received positive feedback were 0.67 (SE = 0.05) and 0.38 (SE = 0.05) in the *Plague* and *Vaccine* conditions respectively (the means are also plotted in the right panel of Figure 2). The difference was reliable, $F(1, 61) = 17.4, p < .001$. The mean proportions of negative triples were thus 0.33 in the *Plague* condition and 0.62 in the *Vaccine* condition (same standard errors, same F ratio).

Triple Heterogeneity

Posvars. The mean number of positive triples where $(b - a) \neq (c - b)$, or *posvars*, were 1.52 (SE = 0.40) and 1.94 (SE = 0.60) in the *Plague* and *Vaccine* conditions respectively (see also the right panel of Figure 3). While subjects in the latter condition seemed to have produced slightly more varied positive triples, the difference was not reliable, $F < 1$.

Negtypes. The mean number of different types of negative triples were 2.16 (SE = 0.33) and 3.68 (SE = 0.44) in the *Plague* and *Vaccine* conditions respectively (see right panel of Figure 3). Subjects in the *Vaccine* condition produced a reliably greater number of different types of negative triples, $F(1, 61) = 7.66, p < .008$.

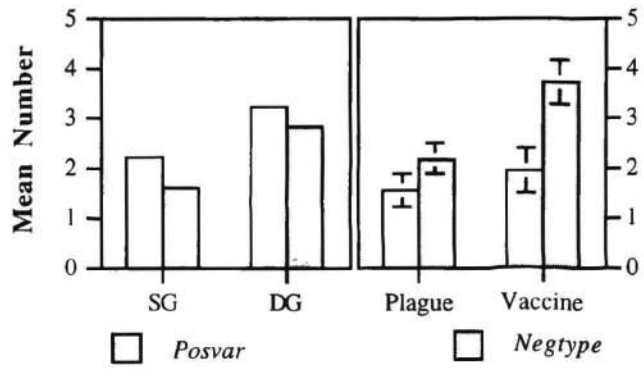

Figure 3. Mean number of variable positive triples (*posvars*) and types of negative triples (*negtypes*) in the SG and DG groups of Vallée-Tourangeau et al. (1995), left panel, and in the two groups of the present study, right panel.

Discussion

In this modified Wason rule discovery task, subjects were more likely to discover the rule "any increasing numbers" in a context in which benefits were associated with the production of triples that would have received negative feedback in the original Wason procedure than in a context in which they were associated with the production of positive triples. Subjects in the *Vaccine* inferential context seemed to have been more successful than subjects in the *Plague* context because they worked harder, as indexed by the reliably greater number of triples they produced, and generated a more creative set of triples, as indexed by the reliably greater proportion and greater variety of triples that received 'negative' feedback. As a result, the *Vaccine* inferential context encouraged the production of a more informative set of triples over which subjects were naturally more likely to infer the 'correct' rule. These data also bolster the explanation of the DG instructions effect in terms of a similar albeit more implicit consideration of the relative importance or epistemic utility of generating triples of a certain kind.

[3] An analogous real-word set of circumstances arose with the development of the anthrax vaccine which, in its early stages of development, was often lethal to the recipient organism. These difficulties were largely resolved by the research of Max Sterne in the 1930's which lead to an effective yet safe vaccine (the Sterne anthrax spore vaccine).

This experimental manipulation also offers a clear illustration of the fact that terms such as 'successful' or 'correct' or indeed 'negative' and 'positive' used to quantify kinds of inferences and triples are thoroughly relativistic. From the perspective of the subjects in the *Plague* inferential context a 'merely sufficient' hypothesis such as 'evens increasing by 2s' that unfailingly produces positive triples (in this context, triples that kill plague-carrying rats) is a very *successful* inductive inference. Indeed to experiment further would be costly in delaying the extermination of rats! In turn, from the perspective of the subjects in the *Vaccine* scenario such a sufficient hypothesis simply won't do. And the reason is not because these subjects abhor such "satisficing" reasoning or seek to abide by loftier canons of hypothesis testing. Rather these subjects are motivated to produce 'negative' triples (and hence save elephant calves). The richer more creative set of triples produced by the subjects in the *Vaccine* scenario is a by-product of these goal-directed efforts. Describing the subjects in the *Vaccine* condition as being more successful than their counterparts in the *Plague* condition makes sense only from the perspective of Wason's original inferential context. Outside this frame of reference, and outside the cognitive psychologist's laboratory, reasoners test hypotheses to achieve goals.

A useful distinction is made by Evans and Over (1996a, b) between *rationality*$_1$ which is characteristic of "reasoning in such a way as to achieve one's goals [in contradistinction to] *rationality*$_2$ [which conforms to a] relevant normative system such as formal logic or probability theory." (1996a, p. 357). Normative considerations of hypothesis-testing practices orthogonal to reasoners' goals are unlikely to do justice to what may otherwise be "rational" reasoning in the sense of Evans and Over's *rationality*$_1$. The experimental conditions designed in this study set out different goals. Subjects adopted these goals and as a result employed hypothesis-testing strategies that yielded triples that differed in quantity and quality. The output of these strategies clearly reflected the subjects' efforts to achieve their respective goals. Hence, subjects in both conditions exhibited equally adaptive reasoning behavior.

Acknowledgments

We thank Neville Austin, Ken Manktelow, and Ryan Tweney for thoughtful comments on a previous version of this paper.

References

Evans, J. St. B. T. (1989). *Bias in human reasoning.* London: Lawrence Erlbaum Associates, Publishers.

Evans, J. St. B. T. , & Over, D. E. (1996a). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological review, 103*, 356-363.

Evans, J. St. B. T. , & Over, D. E. (1996b). *Rationality and Reasoning*. Hove, UK: The Psychology Press.

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.

Kareev, Y., Halberstadt, N., & Shafir, D. (1993). Improving performance and increasing the use of non-positive testing in a rule-discovery task. *Quarterly Journal of Experimental Psychology, 46A*, 729-742.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

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.

Schustack, M. W. (1988). Thinking about causality. In R. J. Sternberg & E. E. Smith (Eds.), *The psychology of human thought* (pp. 92-115). New York: Cambridge University Press.

Sutherland, S. (1992). *Irrationality*. London: Penguin

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., & Arkkelin, 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.

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.

Wetherick, N. E. (1962). Elimination and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology, 14*, 246-249.