Contrast class cues and performance facilitation in a hypothesis-testing task: evidence for an iterative counterfactual model

Maggie Gale · Linden J. Ball

Published online: 9 November 2011
© Psychonomic Society, Inc. 2011

Abstract Hypothesis-testing performance on Wason’s (Quarterly Journal of Experimental Psychology 12:129–140, 1960) 2–4–6 task is typically poor, with only around 20% of participants announcing the to-be-discovered “ascending numbers” rule on their first attempt. Enhanced solution rates can, however, readily be observed with dual-goal (DG) task variants requiring the discovery of two complementary rules, one labeled “DAX” (the standard “ascending numbers” rule) and the other labeled “MED” (“any other number triples”). Two DG experiments are reported in which we manipulated the *usefulness* of a presented MED exemplar, where *usefulness* denotes cues that can establish a helpful “contrast class” that can stand in opposition to the presented 2–4–6 DAX exemplar. The usefulness of MED exemplars had a striking facilitatory effect on DAX rule discovery, which supports the importance of contrast-class information in hypothesis testing. A third experiment ruled out the possibility that the useful MED triple seeded the correct rule from the outset and obviated any need for hypothesis testing. We propose that an extension of Oaksford and Chater’s (European Journal of Cognitive Psychology 6:149–169, 1994) iterative counterfactual model can neatly capture the mechanisms by which DG facilitation arises.

Keywords Reasoning · Inductive reasoning · Heuristics

Many researchers who study high-level cognition ascribe a central role to hypothesis testing within much of human reasoning (e.g., Evans, 2007; Poletiek, 2001). The standard paradigm employed to study hypothesis-testing behavior is Wason’s (1960) rule discovery task. In its original form, participants are asked to discover a rule, known only to the experimenter, that governs the production of number triples. The to-be-discovered rule is “any ascending sequence.” Participants are then given 2–4–6 as a seed triple and are asked to generate further triples, each of which is described by the experimenter as either conforming or not conforming to the target rule. When the participants are confident that they have discovered the experimenter’s rule, they announce it. Despite the apparent simplicity of the task, only about 20% of reasoners are successful on their first rule announcement (Wason, 1960), with around 25% of people still being unable to discover the rule even after repeated announcement attempts (e.g., Farris & Revlin, 1989b; Wason, 1960).

This rule discovery paradigm was originally designed by Wason (1960) to reflect *scientific reasoning*, with participants being required both to generate hypotheses about the unknown rule and to devise tests to evaluate these hypotheses. It can also be argued that the task models hypothesis testing in *everyday reasoning*, since Popper’s (1959) view of scientific reasoning is that it is a special case of everyday reasoning. As such, the low success rate on the task when assessed in terms of initial rule announcements clearly seems paradoxical, in light of the supposed importance of hypothesis testing within the repertoire of human cognitive abilities. It is this paradox that is of interest in the present article, which reports an experimental manipulation that aimed to facilitate rule discovery on the 2–4–6 task so as to provide theoretical insights into the mechanisms underpinning people’s successful and unsuccessful

M. Gale
University of Derby,
Derby, UK

L. J. Ball (✉)
Department of Psychology, Lancaster University,
Lancaster LA1 4YF, UK
e-mail: L.Ball@lancaster.ac.uk

hypothesis-testing performance. First, however, we examine theoretical accounts of the commonly observed failure on the task, as well as approaches that have been taken to facilitate performance and theories that have been proposed to explain facilitation effects.

Wason (1960) noted that many incorrect announcements on the 2–4–6 task were overly restricted versions of the target rule (e.g., “numbers increasing by twos”), suggesting that people produced rules that were sufficient but not necessary to describe the triples that had received positive feedback. To understand people’s performance failure on the 2–4–6 task, Wason coded the triples that people generated as being either “enumerative” (i.e., fitting the current hypothesis and therefore reflecting attempts to support it) or “eliminative” (i.e., not fitting the current hypothesis and therefore reflecting attempts to falsify it). He concluded that many participants adopted an enumerative or confirmatory approach, and proposed that the resulting failure to attempt the falsification of a current favored hypothesis was the critical factor that undermined task success.

Early criticism of Wason’s (1960) “confirmation bias” account came from Wetherick (1962), who claimed that it was simply the experimenter’s essentially devious provision of the 2–4–6 exemplar triple that lured participants into the production of the overly restricted hypotheses that are commonly announced. More recent support for Wetherick’s proposal has come from research by Van der Henst, Rossi, and Schroyens (2002), who invoked “relevance theory” (e.g., Sperber, Cara & Girotto, 1995) to argue that participants *expect* that the information provided by the presented seed triple will be of value for solving the task. In examining this pragmatic relevance account, Van der Henst et al. introduced an experimental manipulation that led participants to believe that the regularities in the 2–4–6 exemplar were actually generated by chance, and found that this undermined the usual tendency for people to produce overly restrictive initial rule announcements (but see Vallée-Tourangeau & Krüsi Penney, 2005, for a methodological critique of this study). In a similar vein, Cherubini, Castelvechio and Cherubini (2005) conducted experiments that varied the types and number of *perceivable relationships* in the example triple, which demonstrated that participants retained as much relational information as possible in their initial hypotheses, which again attests to people ascribing particular value to experimenter-based exemplars. Although these findings provide insights into how people formulate initial, overly narrow hypotheses, they do little to elucidate why participants subsequently fail in testing hypotheses adequately or how people revise hypotheses in the face of new information.

Efforts to improve performance on the 2–4–6 task by instructional manipulations that have emphasized falsification have generally proved unsuccessful (Gorman & Gorman,

1984; Gorman, Gorman, Latta & Cunningham, 1984; Gorman, Stafford & Gorman, 1987; Kareev, Halberstadt & Shafir, 1993; Mynatt, Doherty & Tweney, 1977; Mynatt, Doherty & Tweney, 1978; Tweney, Doherty, Worner, Pliske, Mynatt, Gross & Arkkelin, 1980). With the benefit of hindsight, this finding may not be so surprising, in view of Poletiek’s (1996) proposal that falsification in hypothesis testing is psychologically implausible because of the paradoxical relationship between testing behavior and the correctness of one’s current hypothesis. Poletiek (1996) argues that a person’s current hypothesis in the 2–4–6 task can be viewed as their “best guess” at the rule. However, their selecting a test triple with the intention of falsifying this hypothesis must mean that this hypothesis is, in fact, not their best guess, since they anticipate that it will be false. Thus, if a current hypothesis is one’s best guess, one cannot attempt to falsify it, whereas if one can attempt to falsify it, then it cannot be one’s best guess. Indeed, Poletiek (1996) showed empirically that participants generally expect their best guess to be confirmed by the tests that they make, and they therefore rarely encounter falsifying evidence; when they do observe such evidence, however, they readily seem to understand its implications.

Despite the failure of falsification instructions to improve task performance in the 2–4–6 paradigm, one facilitatory manipulation that has emerged in the literature is to employ “dual-goal” (henceforth, DG) instructions, as introduced by Tweney et al. (1980). In this DG version of the task, participants are required to discover *two* rules that the experimenter has in mind, one called DAX and the other called MED. Participants are given 2–4–6 as an example of the DAX rule and are then asked to generate triples that are classified as either conforming to the unknown DAX rule (which is “any ascending sequence”) or to the unknown MED rule (which is “all other sequences”). Using these amended instructions, success rates on the task for initial rule announcements typically rise to over 60% (e.g., Gale & Ball, 2006; Gorman et al., 1987; Tukey, 1986; Vallée-Tourangeau, Austin & Rankin, 1995; Wharton, Cheng & Wickens, 1993).

Theoretical attempts to account for the DG facilitation effect have generally proved unsatisfactory. Evans (1989) proposed that the facilitation results from the *positive labeling* (as “MED”) of “does not fit” feedback, since such feedback might normally be seen as irrelevant because of preconscious biases to ignore negated information. However, this explanation was revealed to be inadequate by Gale and Ball (2006), who showed that relabeling negatively valenced “does not fit” feedback as MED in the standard, single-goal (SG) version of the 2–4–6 task failed to improve performance.

Another account—Wharton et al.’s (1993) *goal complementarity theory*—claims that two aspects of the

DG task are critical for success: (1) the complementary nature of the DAX and MED rules (i.e., triples can only be DAX or MED) and (2) people's tendency to employ a "positive test strategy," whereby triples are tested that match a current hypothesis (Klayman & Ha, 1987, 1989). In support of their theory, Wharton et al. reported a study in which DG instructions *failed* to elicit facilitated performance when the complementary nature of the rules was left implicit. However, a recent experiment by Gale and Ball (2009) directly manipulated the relationship between the experimenter's two rules such that triples could be DAX, MED or "neither DAX nor MED" (i.e., the DAX and MED rules were no longer logically complementary), yet the level of facilitated performance was equivalent to that observed with the usual DG instructions. Gale and Ball's (2009) findings therefore confirm that rule complementarity is not the causal factor underpinning the DG facilitation effect (see also Vallée-Tourangeau et al., 1995, for a study similar to that of Gale & Ball, 2009, which likewise produced evidence opposing the goal complementarity hypothesis).

Yet another account of DG facilitation is the *triple-heterogeneity theory* proposed by Vallée-Tourangeau et al. (1995). This theory builds on Wason's (1960) observation that solvers can be differentiated from nonsolvers in terms of the types of triples they produce (i.e., solvers have a higher eliminative/enumerative index). Vallée-Tourangeau et al. argued that DG instructions, with their emphasis on the need to discover two hidden rules, induce reasoners to carry out a wider and more creative search of the problem space, which is evident in the sheer *range* of triples that are produced. It is not clear, however, why DG instructions should necessarily lead to a greatly enhanced breadth of triple testing. It is also not the case that such breadth is consistently observed with DG instructions (Gale & Ball, 2009).

Previous research by Gale and Ball (e.g., 2003; 2006; 2009) has suggested an alternative mechanism by which DG instructions might facilitate task performance. This mechanism is based on the "contrast class cues" that are afforded by MED feedback, with these cues enabling participants to clarify the scope of the DAX rule as relating to "ascending" sequences. The "contrast class" notion is viewed by Oaksford (2002) as fundamental to understanding hypothesis-testing behavior in general and the processing of negations in particular. He proposed that a contrast class is a "psychological" rather than a "logical" concept, in that it does not refer to the strict complement of a set, but instead is "made up of the most likely or relevant members of the complement set" (Oaksford, 2002, p. 140). Oaksford provided the example of a reasoner who learns that someone, say John, is not drinking coffee. The immediate "hypothesis" that comes to mind is that John is drinking some other hot beverage—in all likelihood, "tea"—rather

than one of a multitude of other drinks that would fall into the *logical* complement set (e.g., whisky, water, cola, or cough mixture). In this way, contrast class construction can be viewed as being driven by perceived "relevance." Furthermore, Oaksford and Stenning (1992) proposed that the process of computing the most relevant contrast class in hypothesis testing will exploit all available sources of information, including phonetic, syntactic, semantic, and pragmatic cues.

Gale and Ball's (e.g., 2006; 2009) contrast class account of hypothesis testing in the 2–4–6 paradigm stems directly from these notions, with the idea being that the seemingly oppositional nature of DAX and MED provides participants with cues (via triple generation) that soon facilitate insight into the ascending nature of the DAX rule. In particular, receiving MED feedback for a descending triple such as 6–4–2 in the DG task is argued to alert reasoners to the realization that MED and DAX triples may fall at opposite poles of the dimension "descending versus ascending." This account gains support from the observations that DG solvers are far more likely to produce a "contrasting" descending triple than are DG nonsolvers (Gale & Ball, 2003, 2006).

The contrast class account can also explain the impoverished level of rule discovery in the SG 2–4–6 paradigm. In this paradigm if reasoners were to produce triples such as 3–2–1 then these would receive "does not fit" feedback, which would provide useful contrast class cues as to the scope of the "fits" rule. However, according to Gale and Ball (2006), a critical difference between the SG and DG paradigms is that people generate far fewer descending triples in the former than in the latter case, and therefore have less opportunity to experience the resulting contrast class cues. In this way, the contrast class account can explain the DG facilitation effect and can also explain why those who spontaneously generate descending triples in the SG task are more likely to discover the rule than are those who do not generate such triples (Gale & Ball, 2006).

We note, however, that the contrast class account of DG facilitation has, to date, been based solely on the post hoc analysis of people's hypothesis-testing strategies. The primary aim of the present study was, therefore, to test the theory directly by manipulating the initial task scenario given to participants in a way that would cue either "useful" or "nonuseful" contrast class information. To achieve the desired contrast class manipulation we (1) employed a standard DG paradigm, in which the experimenter's DAX rule was the usual "any ascending sequence" and the experimenter's MED rule was the usual "all other sequences," and (2) presented 2–4–6 as the DAX exemplar but additionally informed participants that an example of the MED rule was either the triple 6–4–2 or the triple 4–4–4.

We predicted that the 6–4–2 MED exemplar would act as a useful contrast class cue that would promote identification of the DAX rule. This prediction was based on the fact that 6–4–2 and 2–4–6 are oppositional on the salient and crucial dimension of “ascending” versus “descending.” We predicted that the 4–4–4 MED triple would act as a nonuseful contrast class cue that would fail to promote successful DAX discovery. This is because 4–4–4 and 2–4–6 are oppositional on the salient but irrelevant dimension of “three identical numbers” versus “three different numbers.”

Experiment 1

Method

Participants A set of 43 first-year psychology undergraduates at the University of Derby participated as volunteers. None had received training in reasoning or logic.

Design An independent-measures design was employed in which we manipulated the usefulness of the contrast class cue that was afforded by the example triple presented for the MED rule. One group of participants received a useful contrast class cue (6–4–2), while the other group received a nonuseful contrast class cue (4–4–4). Participants were randomly assigned to these two groups, with the DG paradigm being used in both cases.

Materials and procedure The participants were tested individually in a quiet laboratory. Standard DG instructions were read to the participants, as follows: “I have in mind two rules that specify how to make up sequences of three numbers (triples), and your task is to discover these rules. Triples that fit one of my rules are called DAX triples, and those that fit my other rule are called MED triples.” All participants were given 2–4–6 as an example DAX triple. Those in the useful contrast class cue condition were given 6–4–2 as a MED exemplar, while those in the nonuseful contrast class cue condition were given 4–4–4 as the MED exemplar.

Participants were asked to produce exactly 10 triples and received feedback for each triple in the form of “DAX” or “MED,” in accord with the experimenter’s rules, where DAX was “any ascending sequence” and MED was “all other sequences.” After 10 triples had been generated, participants were asked to write down their best guess at the two rules. In line with the procedure of Gorman (1992), participants were allowed only one guess at these rules.

The decision to impose a 10-triple limit derived from two previous studies (i.e., Klayman & Ha, 1989; Vallée-Tourangeau et al., 1995) that had used a similar constraint as a method to standardize any influence of triple quantity

on rule discovery, while still affording participants an opportunity to examine a variety of triple types should they wish to. Vallée-Tourangeau et al. suggested that the production of 10 triples is an especially good cutoff, since many participants are eager to announce their hypotheses just before this point is reached. This means that this criterion is likely to be minimally disruptive to the participants’ natural triple-testing behavior while still permitting tight controls over the quantities of triples that are tested across conditions.

Results

Solution success across conditions The analysis of solution success data focused solely on correct announcements for the DAX rule. This scoring method has become standard practice in studies with DG task variants from the time it was established by Tweney et al. (1980). Table 1 shows the frequency of correct announcements for the DAX rule across the useful and nonuseful contrast class cue conditions for Experiment 1. It is clear that the usefulness of the given contrast class cue had a dramatic effect on rule discovery, with success rates being over three times greater for participants who have been given the useful contrast class cue relative to those who had been given the nonuseful contrast class cue. A chi-square analysis indicated that this effect was significant, $\chi^2(1, N = 43) = 12.44, p < .001, \phi .54$.

We note that the success rates presented in Table 1 for the useful versus the nonuseful contrast class cue conditions of Experiment 1 are, in fact, very similar to those typically reported in the literature for the DG version of the task versus the SG version of the task, respectively (e.g., Gale & Ball, 2006; Wharton et al., 1993). This observation supports the notion that the DG instructions per se may not lead to task success in the DG paradigm, but rather the fact that DG instructions facilitate participants’ production

Table 1 Frequencies of DAX solvers versus nonsolvers across conditions in Experiments 1 and 2

Condition	Success				
	<i>n</i>	Solvers		Nonsolvers	
Experiment 1					
Useful contrast class cue (6–4–2)	23	17	(74%)	6	(26%)
Nonuseful contrast class cue (4–4–4)	20	4	(20%)	16	(80%)
Experiment 2					
Useful contrast class cue (6–4–2)	20	15	(75%)	5	(25%)
Multiple-dimensions contrast class cue (9–8–1)	22	8	(36%)	14	(64%)
Nonuseful contrast class cue (4–4–4)	22	5	(23%)	17	(77%)

of a salient contrast class—which, in turn, promotes successful discovery of the target DAX rule.

Production of descending triples The purpose of this study was to test the idea that the key to success on the DG task relates to the availability or discovery of useful contrast class information that facilitates the identification of the potential scope of the DAX rule. In our previous research (Gale & Ball, 2006, 2009), we noted that successful solvers are those who uncover at least one descending triple during their hypothesis testing, and that DG instructions promote the production of at least one descending triple. With these findings in mind, we examined the present data set for any effect of the presence versus absence of the participants' production of a descending triple on their task success. Table 2 shows data collapsed across the useful-versus-nonuseful contrast class cue manipulation. Although over 75% of participants produced at least one descending triple (with a majority of these subsequently discovering the DAX rule), there was no single instance of a participant solving the task in the absence of having produced a descending triple. A Fisher's exact test indicated that this effect was reliable, $p = .001$.

The data were further examined to determine whether the type of contrast class cue that had been provided influenced whether or not at least one descending triple was produced (Table 3). This analysis (collapsing across solvers and nonsolvers) showed that 100% of participants receiving a useful contrast class cue produced at least one descending triple, whereas only 50% of those receiving a nonuseful contrast class cue produced such a triple. This effect was shown to be significant using a Fisher's exact test, $p = .001$. It therefore appears that the contrast class cue manipulation was successful in terms of its capacity to induce the generation of descending triples.

Triple types produced by solvers and nonsolvers Notwithstanding the support arising from the previous analyses for the contrast class theory of DG facilitation, it also seemed appropriate to examine the triple-generation profiles of DAX solvers and nonsolvers from the perspective of triple-heterogeneity theory. Triples were therefore codified using the system introduced by Vallée-Tourangeau et al. (1995),

Table 2 Frequencies of DAX solvers versus nonsolvers by presence versus absence of at least one descending triple in Experiment 1

Descending triple	Success				
	<i>n</i>	Solvers		Nonsolvers	
Absent	10	0	(0%)	10	(100%)
Present	33	21	(64%)	12	(36%)

Table 3 Frequencies of participants producing at least one descending triple across conditions in Experiment 1

Condition	Descending Triple				
	<i>n</i>	Absent		Present	
Useful contrast class cue (6–4–2)	23	0	(0%)	23	(100%)
Nonuseful contrast class cue (4–4–4)	20	10	(50%)	10	(50%)

in which positive triples (i.e., ascending sequences) are categorized as either “constant positives” (rising by equal increments) or “variable positives” (rising by nonequal increments), and negative triples (i.e., nonascending sequences) are subdivided into eight discrete “negtype” categories according to the following formulae (where the first number of a triple is a , the second b , and the third c): (i) $a > b > c$, (ii) $a = b > c$, (iii) $a > b = c$, (iv) $a = b = c$, (v) $a = b < c$, (vi) $a < b = c$, (vii) $a > b < c$, and (viii) $a < b < c$. Using the latter, negtype scheme, it is possible to ascribe a “negtype score” to each participant (ranging from 0 to 8) to reflect their coverage of these negtypes. The negtype score is distinct from the total number of generated triples receiving negative feedback. As one final index of triple heterogeneity, we also computed the overall number of different types of triples generated by each participant, whether positive or negative. This index, which ranged from 1 to 10, is referred to as a participant's “triptype score” and is a more accurate measure of the full breadth of triples that a participant has generated (see Gale & Ball, 2009).

Table 4 summarizes participants' triple-generation profiles across solvers and nonsolvers for the indices described. The only significant difference on any measure of triple variety related to the number of descending triples produced, with solvers reliably producing more descending triples than did nonsolvers, $t(41) = 5.77$, $p < .001$, $d = 1.25$. The lack of any association between triple variety and task success provides evidence against the triple-heterogeneity theory of DG facilitation. We also note that Wharton et al.'s (1993) goal complementarity theory predicts that the production of variable positive triples is a key determinant of DG success. Again, the data presented here do not support this notion (see Gale & Ball, 2009, for similar evidence against the goal complementarity theory).

Discussion

The aim of Experiment 1 was to test Gale and Ball's (2006) contrast class account of the DG facilitation effect in Wason's (1960) rule discovery paradigm. This account posits that the use of DG instructions per se is not what gives rise to task success, but rather the fact that standard DG instructions promote the production of a salient

Table 4 Participants' triple-generation profiles for DAX solvers versus nonsolvers collapsed across conditions in Experiments 1 and 2

Heterogeneity Measure	Experiment 1				Experiment 2			
	Solvers		Nonsolvers		Solvers		Nonsolvers	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Constant positives	4.71	0.25	3.86	0.45	4.11	0.31	3.64	0.25
Variable positives	0.71	0.23	0.64	0.25	1.00	0.23	0.97	0.21
Negtype score	1.62	0.21	2.41	0.50	2.07	0.28	2.69	0.23
Negative feedback	4.62	0.21	5.18	0.37	4.79	0.24	5.44	0.28
Descending triples	3.81	0.29	1.36	0.37	2.93	0.26	1.94	0.29
Triptype score	3.00	0.29	3.55	0.41	3.57	0.31	4.19	0.23

SE standard error. Triple-generation profiles were generated in terms of mean number of constant positive triples, mean number of variable positive triples, mean negtype score (ranging from 0 to 8), mean number of triples receiving negative feedback, and triptype score (ranging from 1 to 10)

descending triple for the MED rule, which then enables the successful scoping of the DAX rule as relating to *ascending* triples. In Experiment 1, we predicted that DG facilitation would be supported by providing participants with a MED exemplar such as 6–4–2 that would act to cue the relevance of the “ascending” versus “descending” dimension. We also made the novel prediction that DG facilitation would be suppressed by giving participants a MED exemplar such as 4–4–4 that would *inhibit* the production of a critical descending triple. This is because 4–4–4 and 2–4–6 contrast on a salient but nonrelevant dimension of “three identical numbers” versus “three different numbers,” which does nothing to cue the importance of the ascending-versus-descending dimension.

The results supported these predictions, with the success rates of participants who were given nonuseful contrast class information being equivalent to the rates observed in the SG version of the task. As far as we are aware, this is only the second report of task performance *not* being facilitated using DG instructions, with the other occasion having arisen in Wharton et al.'s (1993) study. Our analyses of the triples produced by participants also provided further support for the contrast class account of DG facilitation. No single case was evident of a participant solving the task in the absence of the generation of a descending triple. Moreover, all participants in the useful contrast class cue condition produced at least one descending triple, whereas only half of the participants in the nonuseful contrast class cue condition did so, thus offering good support for the notion that the production of a descending triple is important for task success.

Our analyses of participants' triple-production profiles additionally allowed us to explore other theories of DG facilitation, such as the triple-heterogeneity theory (e.g., Vallée-Tourangeau et al., 1995), which argues that DG instructions engender a more creative exploration of the

triple space, and the goal complementarity theory (Wharton et al. 1993), which proposes that the DG paradigm elevates the production of variable positive triples that provide the reasoner with insight into the need for a more generalized DAX rule. Contrary to the predictions of triple-heterogeneity theory, however, we found that solvers differed from nonsolvers in only one aspect of their triple production: the number of descending triples produced. The fact that solvers generated both a similar number and a similar range of triple types relative to nonsolvers suggests that DG facilitation cannot readily be attributed to increases in the creative exploration of the problem space. Likewise, the failure to observe any difference between solvers and nonsolvers in terms of their generation of variable positive triples does not appear to support Wharton et al.'s (1993) goal complementarity theory, which emphasizes the importance of such triples in producing the DG facilitation effect.

Experiment 2

Experiment 1 demonstrated the importance of contrast class information for solving Wason's (1960) rule discovery task. As previously mentioned, we view the 6–4–2 MED exemplar that was provided in the useful contrast class cue condition as being oppositional to the DAX exemplar (2–4–6) on a single, salient dimension, that of “ascending” versus “descending.” We reiterate that the 6–4–2 exemplar is not the logical opposite of the most commonly reported DAX hypothesis (i.e., “numbers rising by 2”) but instead represents its psychological complement set (i.e., “numbers descending by 2”). In this respect, we subscribe to Oaksford's (2002) view, discussed previously, which is that a contrast class comprises the most *likely* or *relevant* members of the complement set. We also follow Cherubini et al. (2005) in ascribing to hypothesis testers clear

capacities both to identify “perceivable relationships” between the numbers included in the example triples and to use this relational information to formulate hypotheses for both the DAX and MED rules in the DG paradigm.

The universe of possible descending triples is, of course, infinite and contains triples with a wide variety of properties (e.g., only odd numbers, unequal intervals, prime numbers, etc.). Our choice of 6–4–2 as a descending MED triple in Experiment 1 was made judiciously, so as to maximize the usefulness of the psychological contrast class information that it provided. What might happen, though, if a MED exemplar were provided that varied from the DAX exemplar on multiple dimensions? Presumably, such an exemplar would act as another form of *nonuseful* contrast class cue (much like the 4–4–4 MED exemplar in Exp. 1) and would similarly fail to facilitate DAX rule discovery within a DG paradigm.

To explore this possibility, we decided to run a second experiment that manipulated the number of dimensions on which the MED exemplar varied from the DAX exemplar. We therefore presented 9–8–1 as an illustrative MED triple, since this contrasts with 2–4–6 on several salient but task-irrelevant dimensions (e.g., mixed odd and even numbers vs. only even numbers, unequal vs. equal intervals, and middle number not the arithmetic mean of the outer numbers vs. middle number is the arithmetic mean of the outer numbers), while also contrasting with 2–4–6 on a relevant dimension (i.e., descending vs. ascending numbers). We additionally ran the same two conditions as in Experiment 1 (i.e., using a nonuseful 4–4–4 MED exemplar for one group and a useful 6–4–2 MED exemplar for another group), in an attempt to replicate the findings that had been derived for these DG conditions.

Method

A set of 64 first-year students at the University of Derby took part in this experiment as volunteers. The procedure employed was identical to that of Experiment 1, with the exception that participants were randomly assigned to one of three groups: one group was told that 6–4–2 was a MED triple, the second group was told that 4–4–4 was a MED triple, and the third was informed that 9–8–1 was a MED triple.

Results and discussion

Solution success across conditions A comparison of success rates across conditions (Table 1) shows that DAX rule discovery in the useful contrast class cue condition (6–4–2) was more than double what was achieved in the multiple-dimensions contrast class cue condition (9–8–1), and over three times what arose in the nonuseful contrast

class cue condition (4–4–4). A chi-square test indicated that these differences were reliable, $\chi^2(2, N = 64) = 12.38, p = .002, \phi = .44$.

Partitioning the data (see Siegel & Castellan, 1988) showed no significant difference in success rates between participants given the nonfacilitatory 9–8–1 and 4–4–4 exemplars, $\chi^2(1, N = 44) = 0.36, p = .08, \phi = .14$, suggesting that the multiple-dimensions contrast class cue did indeed hinder performance in this DG version of the task. A significant difference was found, however, when the performance of participants given the nonfacilitatory cues (collapsed across conditions) was compared to the performance of those given the useful 6–4–2 exemplar, $\chi^2(1, N = 64) = 11.54, p < .001, \phi = .42$.

Overall, these findings replicated and extended the observations that arose in Experiment 1 and attest to the role of contrast class cues in sensitizing reasoners to either appropriate or inappropriate rule formulations that can impact markedly on eventual task success.

Triple types produced by solvers and nonsolvers To explore further the strategies pursued by solvers and nonsolvers, we analyzed participants’ triple-generation profiles in terms of the types of triples produced (see Table 4). As with Experiment 1, solvers and nonsolvers differed only in the number of descending triples that were generated, $t(62) = 2.41, p = .019, d = 0.56$.

Experiment 3

Experiment 2 replicated the facilitatory influence on DAX rule discovery of a 6–4–2 MED triple (75% solvers), as observed in Experiment 1 (74% solvers). This result validated our prediction that high levels of DAX solutions can be engendered in the DG task by providing an example MED triple (6–4–2) that cues useful contrast class information versus the seed DAX triple (2–4–6). In essence, 2–4–6 and 6–4–2 are oppositional to each other on a predominant critical dimension: ascending versus descending. Experiment 2 also showed, as predicted, that a 9–8–1 MED exemplar is not helpful for DAX rule discovery, presumably because it contrasts with the 2–4–6 triple on a several task-irrelevant dimensions (e.g., unequal intervals vs. equal intervals) as well as on the task-relevant descending-versus-ascending dimension.

Although the high success rates observed in the useful contrast class conditions of Experiments 1 and 2 corroborate our predictions, an alternative account of the power of the contrast class manipulation is that the 6–4–2 MED triple simply removed the inductive challenge from the outset by immediately seeding the correct answer for the DAX rule,

with subsequent triple testing merely conforming to the seeded hypothesis. This account might also explain why the only difference observed in triple-testing behavior between solvers and nonsolvers in Experiments 1 and 2 was the increased generation of descending triples in the solvers, perhaps reflecting a tendency for them to generate repeated confirming instances of the assumed MED hypothesis. This account is weakened, however, by our previous findings from studies using the “standard” DG paradigm, which likewise showed that solvers produce more descending triples than do nonsolvers (Gale & Ball, 2006, 2009). Thus, descending-triple production is not a localized aspect of behavior associated with the explicit provision of a contrast class cue, but is instead a general feature of successful hypothesis-testing behavior in the DG paradigm.

A good way to arbitrate decisively on the issue of whether the 6–4–2 MED triple simply removes the inductive challenge by directly priming the correct DAX solution would be to ask participants to record their best guess at DAX before they formulate their own first test triple. If participants immediately stated the correct DAX rule, this would support a direct-priming explanation of our observed effect. We report such an experiment below, in which we asked participants to attempt the useful contrast class cue version of the DG task, in which 6–4–2 was used as an example of a MED triple and 2–4–6 as an example of a DAX triple. Participants provided their initial best guess for DAX before formulating their first test triple. We also asked for participants’ best guesses at the DAX rule after each subsequently generated triple, so as to trace their hypotheses across the whole experiment. Finally, at the end of the procedure we asked all apparent solvers to answer whether or not they believed 3–22–25 was a DAX triple. This procedural requirement was included to ensure that participants had genuinely formulated the correct DAX rule and did not believe it to be something narrower (e.g., “ascending with equal intervals”), while failing to mention this.

Method

A set of 15 first-year undergraduates at Lancaster University participated for course credit. The experiment involved a single condition in which participants were given 6–4–2 as an example of a MED triple and 2–4–6 as an example of a DAX triple. The procedure was equivalent to that of Experiment 1, with the exception that participants were asked to record their best guess at the DAX rule following presentation of the exemplar triples as well as after the delivery of DAX/MED feedback for each generated triple. A participant’s understanding of the DAX rule was checked subsequent to experiment completion by presenting the triple 3–22–25 and asking whether or not it fitted the DAX rule.

Results and discussion

Solution success In this experiment, 67% of participants (10 out of 15) successfully announced the DAX rule. A chi-square test for independence showed no difference in the proportions of solvers for this experiment as compared to the combined data from the equivalent conditions in the two previous experiments, $\chi^2(1, N = 58) = 0.35, p = .56, \phi = .07$. All 10 solvers readily and spontaneously identified the 3–22–25 triple presented after the experiment as being a DAX. This finding indicates that solvers were well aware of the true nature of DAX rule and understood that it neither placed constraints on the intervals between numbers nor on the numbers being even or odd.

DAX hypotheses generated A key variable of interest was the nature of participants’ first DAX hypothesis at the start of the experiment, after they had been given the exemplar triples. The data showed that 87% of participants (i.e., 13 out of 15) produced a restricted hypothesis as their initial best guess at the DAX rule. Of these, 11 stated “numbers increasing by twos,” and 2 stated “numbers increasing by equal intervals.” The high number of participants producing overly restricted initial DAX hypotheses (which included 9 of the 10 solvers) runs counter to the idea that the 6–4–2 MED triple simply removed the inductive challenge and directly primed the correct DAX solution. This view is also challenged by the observation that solvers only reached an awareness of the correct DAX rule after a mean of 4.90 self-generated triples ($SD = 1.05$), which is similar to the value that has been reported with the standard DG paradigm (e.g., Wharton et al. 1993). We therefore see no evidence to support a simple explanation of the facilitatory effects of the 6–4–2 MED triple as arising from the direct seeding of the correct DAX rule from the outset.

Descending triples produced We finally assessed the types of descending triples produced by both solvers and nonsolvers. This was again done to examine the alternative account of our findings, whereby direct and immediate seeding of the correct DAX rule might result in tests of MED via stereotyped, conforming triples such as those descending by twos. Our descriptive analysis of the percentages of descending triples that were of various types in fact revealed that only 20% had the form “descending by twos,” with two-thirds of these involving even numbers and one-third involving odd numbers. We further observed that while a majority of descending triples (66%) involved numbers that decreased by equal intervals, 34% nevertheless involved numbers decreasing by unequal intervals. Overall, it was clear that participants were testing a range of descending triple types and were engaged in active testing that promoted a broader understanding of the MED and DAX rules.

General discussion

The aim of the reported experiments was to test the contrast class account of the DG facilitation effect (see, e.g., Gale & Ball, 2006, 2009), which proposes that hypothesis testers are acutely sensitive to the dimensions on which oppositional hypotheses such as DAX and MED lie. In particular, Gale and Ball (2006, 2009) argued that the experimenter's delivery of MED feedback in response to the participant's generation of a descending triple in the DG paradigm functions to alert the participant to the possibility that MED and DAX triples may fall at opposite ends of the descending-versus-ascending dimension. This realization may then trigger the participant's insightful discovery of the DAX rule as being "any ascending sequence."

In the present series of experiments, we tested this contrast class account of DG facilitation by cueing either useful or nonuseful contrast class information in terms of the specific MED exemplars given to participants at the start of the rule discovery procedure. Our findings support the contrast class theory both in terms of the predicted differences in task success across useful versus nonuseful contrast class conditions and in terms of participants' triple-generation profiles. In both Experiments 1 and 2, the provision of a MED example (6–4–2) that cues useful contrast class information versus the seed DAX triple (2–4–6) was observed to have a highly facilitatory effect on DAX rule discovery, as predicted. On the other hand, the nonuseful MED exemplar in Experiment 1 (i.e., 4–4–4) failed to promote discovery of the DAX rule, in spite of the use of a standard DG paradigm. In Experiment 2, we employed an alternative nonuseful MED exemplar (9–8–1) in addition to the 4–4–4 exemplar used in Experiment 1. The 9–8–1 triple contrasted with the 2–4–6 DAX exemplar on a variety of dimensions, unlike 4–4–4, which only contrasted with 2–4–6 on a single salient dimension. The fact that our failure to observe DG facilitation generalized across both the 4–4–4 and the 9–8–1 conditions serves to confirm the robustness of the inhibitory effects on rule discovery of nonuseful contrast class cues.

In terms of participants' triple-generation profiles, Experiments 1 and 2 revealed a reliable difference between solvers and nonsolvers on only a single measure: the number of descending triples produced, with solvers being significantly more inclined to generate descending triples, as predicted under the contrast class account. Moreover, the lack of any differences between solvers and nonsolvers on any other indices of triple type or triple variety appears to run counter to other extant theories of DG facilitation, such as triple-heterogeneity theory (Vallée-Tourangeau et al., 1995) and goal complementarity theory (Wharton et al., 1993). We suggest, therefore, that the contrast class account provides a compelling basis for understanding the nature of DG facilitation in Wason's (1960) rule discovery paradigm.

It remains, however, for us to flesh out a comprehensive theoretical framework that can capture the mechanisms that give rise to the contrast class effects observed. In doing this, we appeal to what we believe is the most psychologically plausible account of behavior on the SG version of the 2–4–6 task—that is, Oaksford and Chater's (1994) *iterative counterfactual model* (ICM), which is a development of Farris and Revlin's (1989a, b) counterfactual strategy. The ICM focuses mainly on how hypotheses are created rather than on how they are tested, with one key aspect of the model relating to how hypotheses are revised when falsifying evidence is obtained.

Summarizing the ICM is best achieved by means of an illustrative example. Imagine a reasoner attempting an SG version of the 2–4–6 task. When confronted with the 2–4–6 seed triple, it is assumed that this reasoner will formulate a working hypothesis (H) such as "even numbers ascending by two," since he or she will strive to retain the maximum amount of relational information contained within the seed triple (Cherubini et al., 2005). The reasoner then generates an alternative "counterfactual" hypothesis (H') that varies from H on just one dimension, such as "odd numbers ascending by two" (cf. Tschirgi's, 1980, "vary one thing at a time" strategy). The reasoner next produces an example triple of this type (e.g., 3–5–7) in order to test H'. Feedback from this test would be positive, such that the reasoner would then add this triple to the set of triples that conform to the to-be-discovered rule and search for a new H that would encompass this additional information (e.g., "any numbers ascending by two").

Using this previous method, our imaginary reasoner would be able to explore the boundaries of the working hypothesis in a way that would eventually lead to successful rule discovery. The ICM is, therefore, a normative model that describes the strategy used by successful hypothesis testers. How, though, does this model relate to the contrast class account of the DG facilitation effect that we are proposing, which claims that task success is dependent on the production of an $a > b > c$ triple that matches the H' "descending numbers"? First, we suggest that DG instructions impact on the ICM by extending and focusing the participant's exploration of alternative hypotheses for both DAX and MED. We further propose that in the "standard" DG task, the reasoner will soon latch onto a DAX hypothesis such as "numbers ascending by twos" and will then turn attention to the MED rule. Since the reasoner has no information on which to base an exploration of the MED rule, but he or she does know that the DAX and MED rules are different, we propose that the reasoner will, at this point, generate a MED hypothesis that is perceived as being the "opposite" of the DAX rule (i.e., "numbers descending by twos"), and following Klayman and Ha's (1987, 1989) positive test strategy, will produce an example

of a descending triple that fits this rule. When given MED feedback for this descending triple, the reasoner will now have examples of both DAX and MED triples as well as working hypotheses for both rules: “ascending by twos” for DAX, and “descending by twos” for MED.

It is at this point in the reasoning process that we propose that reasoners will explore the nature of the boundaries of their contrasting DAX and MED hypotheses by again applying “counterfactual” hypothesis testing, which varies hypotheses by just one dimension at a time. In seeking such a counterfactual hypothesis for either DAX or MED, it will become apparent to the reasoner that a salient dimension to relax would be the “by two” constraint that is common to both the DAX and MED cases. Relaxing this restriction (again in a “vary one thing at a time” mode) while keeping in mind the dominant ascending-versus-descending DAX/MED contrast class, we argue, soon leads to successful rule discovery in many participants.

Although we acknowledge that the aforementioned proposals regarding the role of iterative counterfactual testing in the DG paradigm are somewhat speculative, we note that it is possible to examine aspects of the data deriving from Experiment 3 in order to assess their validity. Recall that Experiment 3 involved a useful 6–4–2 MED triple, together with the standard 2–4–6 DAX exemplar. In addition, participants were asked to record their best guess at the DAX rule after the experimenter’s delivery of DAX/MED feedback for each generated triple. This procedure permits the examination of some specific predictions concerning the way in which iterative counterfactual testing is used to explore the boundaries of the DAX and MED hypotheses. First, participants should be observed to rapidly relax the salient “by two” constraint that is common to the DAX and MED examples in order to drive forward their counterfactual testing. Our analyses indicated that this was exactly what occurred, with 80% of solvers stating a DAX hypothesis that included a “by two” component after presentation of the exemplar triples, before they relaxed this constraint and generated triples that had constant intervals other than twos (e.g., 5–10–15). For all solvers, this process of relaxing the “by two” restriction arose within the first three triple-generation trials.

A second prediction from the iterative counterfactual model is that participants should, having relaxed the “by two” constraint, soon go on to identify a more encompassing DAX hypothesis that generalizes over the properties of the various DAX triples that have been generated. Our analysis indicated that 80% of solvers stated a “numbers increasing with constant intervals” DAX hypothesis at some point prior to correct DAX identification. Furthermore, all of these individuals were subsequently seen to test either ascending or descending triples that involved “variable” intervals prior to DAX discovery. Again, such

findings are fully commensurate with the dominance of an iterative counterfactual testing strategy in the DG paradigm.

A final question that we were keen to address concerned the nature of the triples that were generated immediately prior to DAX rule discovery. In Experiment 3, we found that these triples were overwhelmingly ones involving numbers that were either ascending with variable intervals (50%) or descending with variable intervals (30%). This suggests that just prior to identifying the correct DAX rule, the crucial insight was attained by solvers that ascending or descending numbers with variable intervals are still categorizable as DAX and MED, respectively. Overall, these analyses corroborate core assumptions concerning counterfactual testing in the DG paradigm.

We finally examine the issue of how our findings and theoretical claims might relate to a more general understanding of the nature of hypothesis testing beyond the rather artificial confines of the 2–4–6 paradigm. In addressing this issue, we draw heavily on Evans’s (e.g., 2006; 2007) “hypothetical-thinking theory,” since we believe that it has the capacity to subsume evidence for contrast class effects as well as iterative counterfactual testing as key aspects of hypothetical reasoning. Evans’s hypothetical-thinking theory embraces a “dual-process” perspective on reasoning, whereby reasoning is considered to involve the operation of both “heuristic” processes (i.e., those that are relatively fast, automatic, low-effort, and high-capacity) and “analytic” processes (i.e., those that are relatively slow, controlled, high-effort, and resource-limited). Evans’s theory also embodies three core principles. First, the *singularity principle* claims that when people think hypothetically, they tend to consider only one possibility at a time because of the highly capacity-limited nature of analytic thought. Second, the *relevance principle* proposes that people tend to consider a possibility that is the most plausible or probable in relation to the prevailing context. Third, the *satisficing principle* maintains that possibilities are evaluated with reference to the reasoner’s current goals and are accepted if they are considered to be satisfactory for achieving such goals.

According to Evans’s hypothetical-thinking framework, poor performance on the traditional, SG version of the 2–4–6 task arises from the combined effects of both heuristic and analytic processes (cf. Evans, 2007, p.37). Thus, the participant’s initial, overly restricted hypothesis (e.g., “even numbers ascending by twos”) occurs because of a heuristic relevance effect, whereby information that is salient in the 2–4–6 exemplar is assumed to be relevant to the task (cf. Cherubini et al., 2005; see also Van der Henst et al., 2002). However, the analytic system does not readily intervene to revise the initial hypothesis generated by the heuristic system, since the singularity principle ensures that attention is closely focused on this hypothesis. In addition, the

hypothesis will be evaluated by the analytic system in a satisficing-oriented manner, which is equivalent to the reasoner getting caught up within the “positive subloop” of the iterative counterfactual model: repeatedly testing a plausible hypothesis with confirming instances (see Oaksford & Chater, 1994). The use of DG instructions, however, appears to help the reasoner to overcome the singularity principle as well as satisficing biases by encouraging iterative testing of two hypotheses, with tests of the MED rule providing useful diagnostic information that can serve to extend the boundaries of the DAX rule. We anticipate that future studies of hypothesis testing in both laboratory and real-world contexts will further clarify how contrast class identification and iterative counterfactual testing may underpin the formulation and evaluation of hypotheses, while also validating the assumptions of Evans’s (2006, 2007) hypothetical-thinking framework.

In conclusion, we believe that these reported experiments have advanced our theoretical understanding of the basis of DG facilitation effects in Wason’s (1960) rule discovery task in terms of the role that is played by contrast class cues in leading to the successful identification of the critical dimensions on which opposing rules differ. In addition, we contend that Oaksford and Chater’s (1994) iterative counterfactual model provides a compelling foundation for explicating the detailed processes involved in hypothesis generation and testing with both SG and DG versions of the 2–4–6 paradigm. Moreover, we have shown how the iterative counterfactual model may itself be subsumed within the broader hypothetical-thinking framework of Evans (e.g., 2007).

References

- Cherubini, P., Castelvich, E., & Cherubini, A. M. (2005). Generation of hypotheses in Wason’s 2–4–6 task: An information theory approach. *Quarterly Journal of Experimental Psychology*, 58A, 309–332. doi:10.1080/02724980343000891
- Evans, J. St B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, U.K.: Erlbaum.
- Evans, J. St B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, 13, 378–395. doi:10.3758/BF03193858
- Evans, J. St B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, U.K.: Psychology Press.
- Farris, H., & Revlin, R. (1989a). The discovery process: A counterfactual strategy. *Social Studies of Science*, 19, 497–513.
- Farris, H., & Revlin, R. (1989b). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory & Cognition*, 17, 221–232.
- Gale, M., & Ball, L. J. (2003). Facilitated rule discovery in Wason’s 2–4–6 task: The role negative triples. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the twenty-fifth annual conference of the cognitive science society* (pp. 438–443). Boston: Cognitive Science Society.
- Gale, M., & Ball, L. J. (2005). Contrast class cues and dual goal facilitation in Wason’s 2–4–6 task: Evidence for an extended iterative counterfactual model. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the cognitive science society* (pp. 738–743). Mahwah, NJ: Erlbaum.
- Gale, M., & Ball, L. J. (2006). Dual-goal facilitation in Wason’s 2–4–6 task: What mediates successful rule discovery? *Quarterly Journal of Experimental Psychology*, 59, 873–885. doi:10.1080/02724980543000051
- Gale, M., & Ball, L. J. (2009). Exploring the determinants of dual goal facilitation in a rule discovery task. *Thinking and Reasoning*, 15, 294–315. doi:10.1080/13546780903040666
- 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, U.K.: Harvester Wheatsheaf.
- Gorman, M. E., & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory and a control strategy on Wason’s 2–4–6 task. *Quarterly Journal of Experimental Psychology*, 36A, 629–648.
- Gorman, M. E., Gorman, M. E., Latta, R. M., & Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem solving. *British Journal of Psychology*, 75, 65–74.
- 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. doi:10.1080/02724988743000006
- 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.
- Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211–228. doi:10.1037/0033-295X.94.2.211
- 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.
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29, 85–95. doi:10.1080/00335557743000053
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, 30, 395–406. doi:10.1080/00335557843000007
- Oaksford, M. (2002). Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking and Reasoning*, 8, 135–151. doi:10.1080/13546780143000170
- Oaksford, M., & Chater, N. (1994). Another look at eliminative and enumerative behaviour in a conceptual task. *European Journal of Cognitive Psychology*, 6, 149–169. doi:10.1080/09541449408520141
- Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 835–854. doi:10.1037/0278-7393.18.4.835
- Poletiek, F. H. (1996). Paradoxes of falsification. *Quarterly Journal of Experimental Psychology*, 49A, 447–462. doi:10.1080/1713755628
- Poletiek, F. H. (2001). *Hypothesis-testing behaviour*. Hove, U.K.: Psychology Press.
- Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

- Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57, 31–95.
- Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51, 1–10.
- 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. doi:[10.1080/14640748608401583](https://doi.org/10.1080/14640748608401583)
- Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., et al. (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. doi:[10.1080/14640749508401422](https://doi.org/10.1080/14640749508401422)
- 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.
- Van der Henst, 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 & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 902–907). Mahwah, NJ: Erlbaum.
- 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. doi:[10.1080/14640749308401037](https://doi.org/10.1080/14640749308401037)

Author note

Aspects of the data reported in Experiment 1 formed part of a poster presented at the 27th Annual Conference of the Cognitive Science Society, Stresa, Italy, from the 21st to the 23rd of July, 2005; these results were published as Gale and Ball (2005). We thank Klaus Oberauer, Fenna Poletiek, and an anonymous reviewer for many valuable comments on an earlier version of this article.