

Dual-goal facilitation in Wason's 2–4–6 task: What mediates successful rule discovery?

Maggie Gale

University of Derby, Derby, UK

Linden J. Ball

Lancaster University, Lancaster, UK

The standard 2–4–6 task requires discovery of a single rule and produces success rates of about 20%, whereas the dual-goal (DG) version requests discovery of two complementary rules and elevates success to over 60%. The experiment examined two explanations of DG superiority: Evans' (1989) positivity-bias account, and Wharton, Cheng, and Wickens' (1993) goal-complementarity theory. Two DG conditions were employed that varied the linguistic labelling of rules (either positively labelled Dax vs. Med, or mixed-valence "fits" vs. "does not fit"). Solution-success results supported the goal-complementarity theory since facilitation arose in both DG conditions relative to single-goal tasks, irrespective of the linguistic labelling of hypotheses. DG instructions also altered quantitative and qualitative aspects of hypothesis-testing behaviour, and analyses revealed the novel result that the production of at least a single descending triple mediates between DG instructions and task success. We propose that the identification of an appropriate contrast class that delimits the scope of complementary rules may be facilitated through the generation of a descending instance. Overall, our findings can best be accommodated by Oaksford and Chater's (1994) iterative counterfactual model of hypotheses testing, which can readily subsume key elements of the goal-complementarity theory.

Hypothesis testing is a fundamental mode of mental functioning that involves a comparison between internal thoughts and external facts in order to facilitate interaction with the environment and other people (e.g., Klahr, 2000; Poletiek, 2001). One paradigm that has been used to study hypothesis-testing behaviour is

Wason's (1960) 2–4–6 task, which is a rule-discovery problem that Wason devised to investigate people's conformity to the contemporary scientific norm of empirical falsification when testing hypotheses (Popper, 1959). This norm hinges on the principle that whilst a single disconfirming instance can reveal that a hypothesis is

Correspondence should be addressed to Linden J. Ball, Psychology Department, Lancaster University, Lancaster, LA1 4YF, UK.
Email: L.Ball@lancaster.ac.uk

A subset of the data reported here formed part of a poster presented at the Twenty-Fourth Annual Conference of the Cognitive Science Society, Fairfax, Virginia, USA, August 2002. We are grateful to Jeremy Miles for valuable advice on statistical analyses. We also thank Mike Oaksford, Kevin Paterson, Miles Richardson, and two anonymous reviewers for their helpful comments on previous versions of this article. We acknowledge the University of Derby for its financial support.

incorrect, any number of confirming instances can never prove that the hypothesis is true.

In the standard 2–4–6 task participants have to discover a rule that generates sequences of three numbers (*triples*). Participants are initially given an example of a conforming triple (“2–4–6”) and then produce further triples that the experimenter classifies as conforming or not conforming to the rule. Participants generate triples until they are confident they know the rule, at which point they announce it. The to-be-discovered rule is “three ascending numbers”. Despite the apparent simplicity of task people perform poorly, with only around 20% announcing the correct rule at their first attempt (e.g., Tukey, 1986; Wason, 1960). Many incorrect announcements are restricted versions of the target rule, such as “numbers increasing by two” (Kareev, Halberstadt, & Shafir, 1993). Wason (1960) also noted that solvers and nonsolvers could be differentiated in terms of the number of triples produced, with solvers generating more, and the type of triples produced, with solvers generating a higher proportion of triples receiving negative feedback (cf. Klayman & Ha, 1989). Wason interpreted the nonsolvers’ strategy of testing positive instances of hypotheses as a failure to appreciate the benefits of falsification, labelling this deficit *confirmation bias*. Other hypothesis-testing paradigms have provided further evidence for confirmation bias (e.g., Gorman & Gorman, 1984; Mynatt, Doherty, & Tweney, 1977), although Mynatt et al. note that people seem to appreciate the value of falsifying information if they come upon it.

Although first rule-announcement success on the standard 2–4–6 paradigm is poor, Tweney et al. (1980) introduced a manipulation that improved solution rates to over 60%. Tweney et al. asked participants to discover two rules, one producing “Dax” triples (“any ascending number sequence”), the other producing “Med” triples (“any other number sequence”). Evidence for superior performance with these *dual-goal* (DG) instructions is remarkably robust (e.g., Farris & Revlin, 1989a, 1989b; Gorman, Stafford, & Gorman, 1987; Tukey, 1986;

Wharton et al., 1993), but explanations for the effect are inconclusive. Two prominent accounts are Evans’ (1989) positivity-bias theory and Wharton et al.’s (1993) goal-complementarity theory, and in this paper we report an experiment that aimed to arbitrate between them.

Positivity-bias and goal-complementarity theories of DG facilitation

Evans (1983, 1989) proposed that poor performance on the standard 2–4–6 task can be attributed to the operation of a general *positivity bias*, a form of selective processing whereby people see positive information as relevant to a task, whereas negative information is considered to be irrelevant. On the standard *single-goal* (SG) 2–4–6 task, people are viewed by Evans as testing positive cases that “match” their overly restricted hypothesis (e.g., “ascending with equal intervals is right”), but not negative cases (triples such as “2–4–10”) that do not match such hypotheses. As Evans (1983) puts it, “a ‘wrong’ hypothesis is semantically negative and thus overlooked” (p. 144).

Evans (1989) argues that when Dax and Med are exchanged for the terms “right” and “wrong” on the DG task, people succeed not because they attempt to disconfirm their Dax hypothesis, but, instead, because they test the Med hypothesis (“not ascending with equal intervals”) with positive triples (e.g., “2–4–10”). Because the structure of the DG task means that Med is the complement of Dax, a positive test of Med is effectively a negative test of Dax, which is precisely what is required to eliminate the overspecific Dax hypothesis. Evans’ account, however, still gives rise to the question of why people in the DG paradigm should concern themselves with testing the Med hypothesis. In addressing this, Evans (1989) clarifies that it is the creation of a *positive label* (Med) for the negative hypothesis (“not ascending with equal intervals”) that changes people’s conception of the problem, thereby promoting the testing of Med hypotheses. As Evans (1989) explains, “When Dax and Med are substituted for ‘right’ and ‘wrong’, the two hypotheses appear to have equal standing. . . . Creating a positive label for

the negative hypothesis entirely changes subjects' representation of the task" (p. 52).

The goal-complementarity theory (Wharton et al., 1993) emphasizes three factors that promote DG success: (a) the embedding of initial hypotheses within the target rule (cf. Wetherick, 1962); (b) the complementarity of the Dax and Med rules; and (c) the tendency for people to adopt a *positive test strategy* (Klayman & Ha, 1987, 1989) involving the generation of triples that match hypotheses. The positive test strategy is identical in its functional consequences to Evans' (1989) notion of positivity bias: Both promote the testing of positive exemplars of a single, overly restricted hypothesis in the SG paradigm, and of the Dax and Med hypotheses in the DG paradigm. Where the positivity-bias and goal-complementarity theories differ is in their accounts of why people test Med hypotheses at all. Whilst Evans (1989) argues that it is the positive labelling of a negated Dax rule that makes Med hypotheses appear relevant and worth testing, the goal-complementarity theory holds that people test Med hypotheses because they are directly requested to discover the two rules pertaining to the task.

Information-quantity and triple-heterogeneity theories of DG facilitation

The information-quantity theory of DG facilitation (Wharton et al., 1993) centres around two pieces of evidence: (a) that prior to initial rule announcement SG solvers test more triples than do nonsolvers (Farris & Revlin, 1989a; Wason, 1960); and (b) that DG instructions invoke the testing of more triples than do SG instructions (e.g., Gorman et al., 1987). These observations suggest that the quantity of triples produced, irrespective of their characteristics, may mediate between DG instructions and task success. However, this theory lacks clarity as to the specific mechanisms that promote increased triple testing in DG conditions; it is also questionable why increased testing should be associated with task success. Tweney et al. (1980) and Gorman et al. (1987) have gone some way toward addressing

these issues, suggesting that switching between the testing of two different rules in the DG paradigm may counter the rigid testing of a single hypothesis, thereby increasing the number of hypotheses examined, the quantity of triples generated, and the likelihood of success.

The information-quantity view has some merit, but seems mostly to be a useful description of one aspect of the data surrounding DG superiority (i.e., that DG instructions lead to increased triple production) and is certainly compatible with either the goal-complementarity or the positivity-bias theories. Moreover, these latter accounts have the advantage of providing a clear explanation of why people test more triples in the first place with DG instructions: People do this because they are explicitly requested to discover two rules (the goal-complementarity position) or because the positive labelling of negated hypotheses encourages a broader exploration of the hypothesis space (the positivity-bias position).

Vallée-Tourangeau, Austin, and Rankin (1995, Exp. 1) explored the link between DG instructions and information quantity in a study where participants had to test 15 triples before rule announcement. This led to 44% solvers with SG instructions and 69% solvers with DG instructions. The fact that DG solvers outstripped SG solvers even though the number of tests was standardized affirms that there is more to the DG effect than the mere quantity of triples tested. Vallée-Tourangeau et al. (1995) proposed that the crucial factor underpinning DG success is not so much triple quantity as triple heterogeneity (the extent to which triples discriminate between multiple hypotheses), with DG requirements fostering a more flexible and creative exploration of the triple space. This idea provides a valuable development of the insights concerning multiple-hypothesis testing and task success presented by Tweney et al. (1980) and Gorman et al. (1987). It also links with Oaksford and Chater's (1994) argument that, *prima facie*, the more hypotheses a person tests the higher will be the probability of success.

Triple-heterogeneity theory also gains support from Vartanian, Martindale, and Kwiatkowski's

(2003) study, which demonstrated that successful participants on SG 2-4-6 tasks had higher fluency scores on a measure of creative thinking than did unsuccessful participants and also generated more hypotheses. Intriguingly, however, Vartanian et al. noted that whilst the creativity factor was predictive of task success in a stepwise multiple-regression analysis, the number of generated hypotheses did not prove to be a reliable predictor. In fact, the most important factor in rule discovery was the frequency of disconfirming triples generated (i.e., triples that contradicted a participant's working hypothesis such as stating "6-5-4" to test the hypothesis "numbers ascending by twos"). The value of testing disconfirming triples was highlighted in Wason's (1960) original analysis of successful 2-4-6 performance, where people who used disconfirmation as their dominant strategy were more likely to announce correct rules on their first guess. However, other data suggest that disconfirmation may not always be the most efficient strategy, especially in the initial stages of hypothesis testing (e.g., Tweney et al., 1980).

Comparing the positivity-bias and goal-complementary theories

In relation to our experiment we were conscious of the need to determine whether the number and characteristics of triples produced mediate successful rule discovery. We anticipated that assessing such issues in variants of the 2-4-6 task would facilitate the development of a clearer account of the links between triple generation and correct rule announcement. Our primary aim, however, was to develop a crucial test to arbitrate between the positivity-bias and goal-complementarity views of the 2-4-6 task as measured in terms of successful initial rule announcement.

Despite the conceptual overlap between these accounts we believed that a crucial test is possible (cf. Wharton et al., 1993). As noted earlier, the key difference between these theories is that whilst for Evans (1989) DG facilitation is tied to the use of a positive label (Med) to denote a negative hypothesis ("not ascending with equal intervals

is wrong"), for Wharton et al. (1993), what is critical is the explicit instruction for participants to discover the Med as well as the Dax rule in the DG paradigm. One consequence of the positivity-bias theory is that it would predict that participants given positively labelled DG hypotheses (Dax and Med) should perform better than participants given mixed-valence DG hypotheses of the form "fits" (positively labelled) and "does not fit" (negatively labelled). In contrast, the goal-complementarity theory would predict that participants given the task of discovering two complementary hypotheses should be more successful than those seeking a single hypothesis, regardless of the way in which these rules are labelled.

Our experiment set out to test these alternative predictions by manipulating the linguistic labelling of hypotheses in the DG paradigm, such that the sought-for hypotheses were described as being either Dax and Med rules (the positive-label condition) or "fits" and "does not fit" rules (the mixed-label condition). In addition, we included two SG task variants to provide a baseline measure of rule discovery success. One SG condition asked participants to discover a rule labelled Dax, with positively labelled Dax and Med feedback being given for generated triples; the other asked people to discover a rule labelled "fits", with mixed-valence "fits" and "does not fit" feedback being provided. These two forms of SG task therefore involved equivalent feedback as arose in the DG conditions. This linguistic balancing of feedback provided a manipulation check to ensure that any effects associated with the DG conditions could not be attributed to the nature of the available feedback per se.

EXPERIMENT

Method

Participants

A total of 60 undergraduates from Derby University participated in the experiment for course credit. They had not received any prior teaching on the psychology of reasoning.

Design

The experiment involved random allocation of 15 participants to each of four conditions. Two conditions presented DG instructions and included a linguistic labelling manipulation whereby participants were asked to discover either Dax and Med rules (positive-label condition) or “fits” and “does not fit” rules (mixed-label condition). The remaining two conditions involved SG instructions and simply acted as controls for the presented feedback in the DG conditions (i.e., feedback was either of the form Dax–Med or of the form fits–does not fit).

Procedure

Participants were tested in groups of up to four in a quiet laboratory. The following safeguards avoided social contamination of individual results: (a) participants sat well apart in screened cubicles; (b) participants wrote down triples and rules and were provided with written feedback; (c) no spoken communication was permitted during the experiment (see Gorman & Gorman, 1984; Gorman et al., 1987, for effective use of group testing with the 2-4-6 task). The SG instructions referred to a unique rule that needed to be discovered and stated, “I have in mind a rule that specifies how to make up sequences of three numbers (triples), and your task is to discover this rule”. In what we subsequently refer to as the SG–fits condition, participants were asked to discover the target rule by generating triples that they would then be told either fitted or did not fit the rule that the experimenter had in mind. In the SG–Dax condition, participants were told that triples that fitted the rule were called Dax triples, and those that did not fit the rule were called Med triples. It was explained to participants that on generating a triple they would be informed as to whether it was a Dax or a Med type.

The DG instructions emphasized that there were two rules to be discovered, “Your task is discover this rule, and also a second rule for categorizing the triples that do not fit my rule”. In the DG–fits condition participants were asked to generate triples that would be classified in terms of whether they fitted or did not fit the rule. In the standard DG task (DG–Dax) participants

were informed that triples that fitted the rule were called Dax, and those that did not fit the rule were called Med. They were instructed to produce further triples that the experimenter would classify as being Dax or Med.

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

Results and discussion

Solution success across conditions

An alpha level of .05 was set for all statistical tests. In scoring participants’ success we were interested only in whether initial rule announcement correctly mapped onto the Dax or “fits” rules, and not whether announcements for the Med or “does not fit” rules were also correct in the DG conditions. This scoring procedure is standard practice in research with DG task variants (e.g., Vallée-Tourangeau et al., 1995; Wharton et al., 1993). Table 1 shows the frequency of correct and incorrect initial rule announcements in each of the experimental conditions. A number of findings are apparent. First, nearly four times the number of participants in the DG conditions announced the correct rule when compared with the SG baseline conditions. Second, the linguistic-labelling manipulation associated with the DG rules had little impact on rule discovery, with nearly equal numbers of participants finding the sought-for rule in the positive-label and mixed-label conditions. Third, the distribution of initial solvers and nonsolvers was identical across the SG conditions, indicating that the labelling of feedback (Dax–Med vs. fits–does not fit) had no effect on the likelihood of successful rule announcement.

Table 1. Frequency of correct initial rule announcements, mean numbers of total triples produced, types of triples produced, and triples receiving negative feedback, by condition

| Condition | n | Solvers | Nonsolvers | Total triples | | Variable positives | | Negative types | | Negative feedback | |
|-----------|----|---------|------------|---------------|------|--------------------|------|----------------|------|-------------------|------|
| | | | | M | SD | M | SD | M | SD | M | SD |
| SG—Dax | 15 | 3 | 12 | 7.60 | 6.25 | 0.33 | 0.62 | 0.33 | 0.62 | 0.73 | 1.28 |
| SG—fits | 15 | 3 | 12 | 5.87 | 2.75 | 0.53 | 1.36 | 0.80 | 1.42 | 0.93 | 1.58 |
| DG—Dax | 15 | 12 | 3 | 10.27 | 6.30 | 1.13 | 1.06 | 1.20 | 0.86 | 2.67 | 1.95 |
| DG—fits | 15 | 11 | 4 | 8.33 | 3.22 | 1.07 | 1.10 | 0.93 | 0.80 | 1.40 | 1.24 |

Note: SG = single goal; DG = dual goal.

A contingency table chi-square analysis was performed on the frequencies of correct and incorrect announcements, collapsing across the two SG conditions and the two DG conditions. This revealed a highly reliable effect of goal requirement (SG vs. DG), $\chi^2(1, N = 60) = 19.29, p < .001$. In terms of correct rule announcements, then, the results arbitrate in favour of the predictions of the goal-complementarity account of DG facilitation (e.g., Wharton et al., 1993) and against the predictions of the positivity-bias account (e.g., Evans, 1989).

Quantity and variety of triples generated across conditions

It is straightforward to obtain measures of *total triple quantity* (i.e., the cumulative count of a participant's generated triples), and of *total triples receiving negative feedback* (i.e., the cumulative count of a participant's generated triples receiving "Med" or "does not fit" responses from the experimenter). Such measures were computed from the dataset on a by-condition basis. Obtaining a measure of the variety of triples generated by participants is more complex and involved recourse to techniques pioneered by Vallée-Tourangeau et al. (1995), which focus on two main classes of triple referred to as *variable positives* and *negative types*.

Variable positives are triples such as 2–8–20 that receive positive feedback (Dax or fits) but that do not increase by a constant number as in the 2–4–6 exemplar (described as having a "constant positive" form). Thus, if the numbers

that make up a triple are denoted by the letters a, b, and c, then a variable positive is any triple in which $(b - a) \neq (c - b)$, whereas a constant positive is any triple where $(b - a) = (c - b)$. We computed a variable positives score for each participant, which was simply a cumulative count of the number of such triples produced.

Negative types reflects a measure of the heterogeneity of negative triples. There are eight possible types of triple that could receive negative feedback, such as descending triples and identical-number triples. The possible set of negative types is captured by the following rules: (a) $a > b > c$; (b) $a = b = c$; (c) $a > b < c$; (d) $a < b > c$; (e) $a = b < c$; (f) $a = b > c$; (g) $a > b = c$; (h) $a < b = c$. To obtain a negative types score for each participant we counted the number of distinct types of negative triple that they produced. Thus, if a participant generated five negative triples of the same kind (say the decreasing triples of the $a > b > c$ form), then their negative types score equalled one. If the participant generated three negatives—two decreasing and one $a < b > c$ "hill" kind—then their negative types score would be two. Using these indices of triple heterogeneity, Vallée-Tourangeau et al. (1995) demonstrated that the DG manipulation leads to increased production of both variable positives and negative types compared to SG instructions—a finding they interpret as indicating that people consider a wider range of hypotheses with DG instructions.

Table 1 presents mean scores by condition for the number of variable positives and negative

types generated in the present study, as well as means for the total quantity of triples generated and the number of triples receiving negative feedback. To examine whether our experimental manipulations had an effect on any of these indices we undertook a series of two-way analysis of variance (ANOVA) tests where the factors were goal requirement (SG vs. DG) and linguistic labelling of feedback (i.e., our experimental controls involving the linguistic balancing of feedback across the SG and DG tasks allowed for a comparison between conditions where Dax-Med feedback had been given vs. conditions where fits-does not fit feedback had been presented).

Consistent with the task-success analyses already reported, linguistic labelling of feedback produced no reliable differences on any of our measures of triple quantity or type. With regard to goal requirement, however, there were significant main effects on total number of triples produced, $F(1, 56) = 4.09, p < .05$, number of triples receiving negative feedback, $F(1, 56) = 9.10, p < .01$, and number of variable positives, $F(1, 56) = 5.86, p < .05$. The difference in the number of negative types produced across SG and DG conditions also approached significance, $F(1, 56) = 3.96, p = .052$. There were no significant interactions between factors for any of the measures (all $F_s < 1$). These results underline the importance of DG instructions as a determining factor in engendering quantitative and qualitative changes in triple generation on the 2-4-6 task. As such, the findings again favour the goal-complementarity view of DG facilitation.

Presence of triple types and solution success

Although the previous analyses indicate an effect of DG instructions on measures of triple quantity and type, they do not reveal the importance of generating specific types of triple for actual task success. We therefore pursued further analyses in which the production of either at least one variable positive or at least one negative type was crossed with success. Such analyses are important for a simple but critical reason. The point is, both the goal-complementarity and the positivity-bias accounts place a central emphasis on the role of

variable positives in engendering successful rule announcement—that is, both accounts claim that once people generate a variable positive (e.g., “2-4-10”) that receives Dax or fits feedback they should be able to ascertain immediately that the Dax or fits rule is broader than the initially hypothesized form. On the other hand, both theories are silent as to the value of negative testing in facilitating task success.

To examine the association between generation of variable positives and success a contingency table was produced (see Table 2) in which the production by a participant of at least one variable positive was crossed with successful rule discovery. Table 2 reveals that such an association is indeed evident, with substantially more participants who produced a variable positive making a correct rule announcement than those who did not produce a variable positive. A chi-square analysis confirmed the reliability of this observation, $\chi^2(1, N = 60) = 13.14, p < .001$.

Table 2 also shows a contingency table in which production of at least one negative triple was crossed with success. Here the association is even more striking than in the case of variable positive production, with there being only one instance of a participant correctly announcing the rule but not producing a negative triple. In contrast, of the 34 participants who did produce a negative triple, 28 solved the task. A chi-square analysis indicated that these differences were highly significant, $\chi^2(1, N = 60) = 36.36, p < .001$. These latter findings—that production of at least a single triple receiving negative feedback is more closely associated with success than production of

Table 2. Frequency of correct rule announcements by presence versus absence of variable positives and negative triples

| | <i>Solvers</i> | <i>Nonsolvers</i> | <i>Total</i> |
|---------------------------|----------------|-------------------|--------------|
| Variable positive present | 19 | 6 | 25 |
| Variable positive absent | 10 | 25 | 35 |
| Total | 29 | 31 | 60 |
| Negative triple present | 28 | 6 | 34 |
| Negative triple absent | 1 | 25 | 26 |
| Total | 29 | 31 | 60 |

a single variable positive—seem paradoxical. Since most incorrect rule announcements are of the “numbers ascending by equal intervals” type it would appear that only production of variable positives could lead to the falsification of overly restricted hypotheses, whereas production of negative triples would seem of little obvious value for rule discovery.

To clarify the association between the triple-type variables and task success we modelled the dataset using logistic regression. An initial model using negative triple (present vs. absent) as the predictor variable, and success (solver vs. nonsolver) as the outcome variable, revealed that negative triple was a highly reliable predictor of task success, $B = 4.76$, $Wald = 18.23$, $p < .001$. A second model using variable positive (present vs. absent) as the predictor, and success (solver vs. nonsolver) as the outcome variable again revealed this predictor to be reliable, $B = 2.07$, $Wald = 11.91$, $p < .001$. A final model was assessed in which negative triple and variable positive were regressed onto success in a hierarchical manner to examine whether the variable positive predictor had an effect additional to the negative triple predictor. Negative triple was selected as the first predictor due to the considerably higher overall chi-square value obtained for Model 1, $\chi^2 = 42.95$, $p < .001$, than for Model 2, $\chi^2 = 13.68$, $p < .001$. This third model showed that negative triple continued to be a highly reliable predictor of task success, $B = 4.58$, $Wald = 15.71$, $p < .001$, but with negative triple controlled for the variable positive predictor failed to achieve significance, $B = 1.72$, $Wald = 3.55$, $p > .05$. This finding suggests that the production of variable positive triples may have a limited association with successful rule discovery on the 2–4–6 task in comparison to the more striking influence that the generation of negative triples seems to have.

Since production of negative triples appeared to have such a strong association with task success we examined what specific property of negative triples might underpin this phenomenon. It was clear upon scrutinizing the negative triples that participants generated that the majority were

“descending” in nature, and that most other types of negative triple were produced infrequently (only 10 participants produced a negative type distinct from the descending $a > b > c$ form). For this reason we collapsed all negative triples apart from those of the $a > b > c$ type into a single pool. We then pursued separate analyses comparing (a) the effect of producing versus not producing at least one descending triple on task success, and (b) the effect of producing versus not producing at least one other type of negative triple on task success. Logistic regression revealed that participants generating a descending triple were 142 times more likely to solve the task than those not producing a descending triple, $B = 4.98$, $Wald = 19.47$, $p < .001$. Production versus nonproduction of at least one other type of negative triple was also predictive of success, $B = 2.60$, $Wald = 5.67$, $p = .017$, although clearly not to such a marked degree as production versus nonproduction of descending triples.

Path analysis of mediator effect

The results of our logistic-regression analyses are instructive, but the role of production of a descending triple in mediating between goal requirement (SG vs. DG) and task success could better be illustrated using path analysis (Baron & Kenny, 1986). However, because the quantitative logic of path analysis does not work effectively with logistic regression, it was necessary to use linear regression for this mediation analysis. This is not ideal given the dichotomous nature of the relevant variables, but we present comparisons of obtained p values for logistic and linear regressions (Table 3) to illustrate the high degree of similarity in the statistical outcomes of these two approaches and thereby to validate the use of the linear-regression procedure with the present dataset. Table 3 also summarizes the results of this path analysis of the relationship between goal requirement (SG vs. DG), descending triple (present vs. absent), and success (solver vs. nonsolver) in the 2–4–6 task. The amount of mediation relating to the production of a descending triple was large (0.352). A Sobel test revealed that this mediating effect was highly reliable, Goodman (I) = 3.87, $p < .001$.

Table 3. Summary of mediation analysis of the relationship between goal, production of at least one descending triple, and success, using linear regression

| | | | | <i>p</i> values obtained | |
|--|------------|----------|-------------|--------------------------|----------------------------|
| | | | | <i>Linear regression</i> | <i>Logistic regression</i> |
| | | <i>B</i> | <i>Beta</i> | <i>SE(B)</i> | |
| Goal onto solver | | .567 | .567 | .108 | <.001 |
| Goal onto descending triple | | .500 | .503 | .113 | <.001 |
| Goal and descending triple onto solver | Goal | .215 | .215 | .086 | .015 |
| | Descending | .703 | .700 | .086 | <.001 |

Note: Includes a comparison of the *p* values obtained using linear and logistic regression.

GENERAL DISCUSSION

This study's task-success findings indicate that the DG superiority effect in the 2-4-6 task cannot be attributed to labelling a negatively-valenced "does not fit" hypothesis as a positively-valenced "Med" hypothesis: Participants instructed to discover two rules performed significantly better than those in SG conditions, regardless of whether the two rules were defined as Dax-Med or as fits-does not fit. Such findings support key elements of Wharton et al.'s (1993) goal-complementarity account of the facilitatory effect of DG instructions and run counter to Evans' (e.g., Evans, 1989) positivity-bias account, which proposes that people selectively attend to positively labelled information at the expense of attending to potentially useful information that is negatively labelled. Although we find the idea of positivity bias affecting hypothesis testing appealing (cf. Ball, Lucas, Miles, & Gale, 2003), it seems that the concept of a generalized positivity bias cannot easily extend to an explanation of behaviour on Wason's 2-4-6 task.

Despite our solution-success evidence for goal-complementarity theory we are aware that a strong version of this theory is undermined by evidence that strict rule complementarity is unnecessary for rule discovery. For example, Vallée-Tourangeau et al. (1995) ran DG conditions that explicitly suggested a noncomplementary representation of the Dax and Med rules. In one condition participants were told that triples could be Dax, Med, or neither, and in another they were told that triples could be Dax, Med, or both. With these manipulations 80% of people still

discovered the Dax rule on initial announcement. One weakness of Vallée-Tourangeau et al.'s (1995) study was that although the apparent relationship between the Dax and Med rules was manipulated such that they were not represented as complementary, the reality was that (unknownst to participants) the two rules actually remained logically complementary, and feedback other than Dax or Med was never given. Recently, however, Gale and Ball (2003) have replicated Vallée-Tourangeau et al.'s (1995) findings when genuinely noncomplementary rules were used in conjunction with appropriate feedback.

As for information-quantity theory (Wharton et al., 1993) our results support the view that DG instructions promote increased test generation when compared with SG instructions. Likewise, in relation to triple-heterogeneity theory (Vallée-Tourangeau et al., 1995), we observed increased generation of triples receiving negative feedback and increased variable positive triples in DG conditions relative to SG ones. It thus seems that DG facilitation is mediated by both quantitative and qualitative changes in triple-testing behaviour. As we noted previously, however, information-quantity and triple-heterogeneity theories are perhaps more descriptive than explanatory in emphasis, and they may be better subsumed within the process-oriented perspective afforded by other contemporary accounts of the 2-4-6 task discussed below.

Our final set of analyses revealed a hitherto unnoticed phenomenon: It is the production of at least a single descending triple that is most

closely associated with task success, rather than other factors linked to triple-generation behaviour. The observation that negative-triple testing in general and descending-triple testing in particular are so closely linked to success appears to challenge a central assumption of the goal-complementarity theory of DG facilitation. This is because according to this theory it is the production of discriminatory variable positives that should determine rule discovery as it is only variable positives that can falsify overly restricted hypotheses. How, then, might we explain the importance of descending-triple production as a predictor of task success and, in particular, its central role as a mediator between DG instructions and rule discovery? To progress toward an account of these findings we turn to what we believe is the most psychologically plausible contemporary account of behaviour on the standard SG 2-4-6 task—that is, Oaksford and Chater's (1994) iterative counterfactual model, a development of Farris and Revlin's (1989a, 1989b) counterfactual strategy.

The iterative counterfactual model (ICM) emphasizes how hypotheses are created by participants in the first place (the "context of discovery"), and not just on how such hypotheses are tested (the "context of justification"). One particularly important aspect of the ICM concerns how hypotheses are revised in the light of falsifying evidence. The operation of the ICM can best be illustrated with reference to an example. Imagine that a participant's working hypotheses, H , based on the initial 2-4-6 triple, is "even numbers ascending by two". According to the ICM, the participant generates an alternative hypothesis, H' , that is complementary to H for one property. For example, with the triple 2-4-6 and an H of "even numbers ascending by two", H' could be "odd numbers ascending by two". H' is then tested using a positive test strategy; in the present example a positive test triple for H' is 3-5-7. If this test obtains confirmation then both H' and H must be false and can be rejected. A common property of the triples conforming to H and H' (i.e., 2-4-6 and 3-5-7) is now selected, and a new H is generated. This H might be "numbers ascending by two". The whole process is then

iterated with the testing of a new H' complementary to H (e.g., "numbers descending by two"), with the accumulation of new test triples, and with the production of revised hypotheses that are closely informed by the identification of perceivable common properties that hold across all triples that have received positive feedback so far.

Oaksford and Chater (1994) acknowledge that the ICM generates inputs to the hypothesis generation process but does not explain why one common property shared by a set of triples is selected over another common property. For example, given 2-4-6 and 3-5-7, why should "numbers ascending by two" rather than "ascending numbers" be selected? Cherubini, Castelvechio, and Cherubini (2005) have proposed that the former hypothesis is preferred because it conveys more information—that is, people are sensitive to perceiving "regularities" in triples that have maximal information relevance and then generate hypotheses by abstracting those regularities. Support for these ideas is provided by a series of experiments presented by Cherubini et al. that varied the type and number of perceivable relationships in example triples, with the dependent measure being the informational structure of the first hypothesis participants generated. For example, one experiment used two example triples and revealed that the presence of "high-information" regularities in these triples affected the information in the initial hypothesis more than did the presence of "low-information" regularities.

Cherubini et al.'s research goes far in clarifying the process of common-feature extraction that drives the formulation of hypotheses in the ICM. Although these proposals are currently restricted to the first hypothesis generated in SG versions of the 2-4-6 task, similar mechanisms may well underpin hypothesis revision in the SG paradigm and hypothesis generation processes in DG task variants. This brings us back to some final reflections on how the ICM and the notion of common-feature extraction may link with our observation of the importance of descending triples in mediating DG facilitation effects. First, we need to reconsider the illustrative example of

the ICM described above and imagine what might transpire in the SG paradigm at the second iteration where the participant's working hypothesis, H , is "numbers ascending by two", and the complementary H' is "numbers descending by two". According to the ICM a triple congruent with H' would now be proposed (e.g., 6-4-2) that would receive "no" feedback, indicating that this triple is not an instance of the target rule (T). Oaksford and Chater (1994) propose that if "no" feedback is obtained then H' is rejected, and H is a possible candidate for T . Before announcing H as T , however, people may generate a few positive exemplars of H to test whether they are instances of T , entering a so-called "positive subloop" of the ICM.

We note, however, that the ICM as it is currently formulated does not deal with the DG paradigm where two complementary rules (e.g., Dax and Med) are being sought. Our proposal is that in the DG task people are sensitive not only to the regularities in triples that are categorized as Dax and Med but also to the contrast class that such regularities invoke. Thus, at this second iteration outlined above, the participant in the DG paradigm would be given the feedback that 6-4-2 is an example of the to-be-discovered Med hypothesis. H and H' are, of course, complementary for the properties *ascending* versus *descending*; indeed it was the ascending versus descending dimension that was being varied by our imaginary participant at the second iteration of the ICM, with all other triple features being kept constant. We claim that the fact that H and H' are complementary for the properties ascending and descending serves to establish a salient contrast class that promotes a participant's insight into the potential scope of the Dax rule as reflecting "ascending" numbers and the Med rule as reflecting "descending" numbers (thereby facilitating effective rule discovery).

Our ideas concerning the role of contrast-class identification within the ICM and its link to DG facilitation are clearly speculative, and closer investigation of the role of descending triples in facilitating task success would seem an important line for future research. To achieve this a finer

grained system of codifying triples may be required as well as closer assessment of the links between generated triples and working hypotheses. It is encouraging, however, that Oaksford (2002) views the contrast class concept as fundamental to understanding behaviour in Wason's other famous hypothesis-testing task (i.e., the four-card selection task), and it may likewise be relevant to explaining behaviour in Wason's 2-4-6 paradigm. Oaksford (2002) argues that a contrast class is best viewed as a psychological rather than a logical concept in that it does not simply refer to the complement of a set, but is instead "made up of the most likely or *relevant* members of the complement set" (p. 140). Oaksford and Stenning (1992) also emphasize that computing the most relevant contrast class can exploit many sources of information, including semantic, syntactic, and pragmatic cues. We similarly suggest that the properties "descending" and "ascending" when identified as capturing salient features of Med and Dax triples establish a powerful contrast class that may promote insightful rule discovery in DG variants of the 2-4-6 task.

Overall, we believe that our study has progressed an understanding of DG facilitation effects in the 2-4-6 task. Our basic task-success measures provide little support for a positivity-bias view of the DG effect (e.g., Evans, 1989), as increased levels of task success can arise even with negatively labelled rules, so long as participants are still instructed to discover two rules. Although this latter result provides support for the goal-complementarity view that direct requests to discover two rules are central to DG facilitation (Wharton et al., 1993), we have also argued that there is other evidence that calls into question the necessity of having logically complementary rules within DG manipulations. Indeed, the idea that DG instructions may promote identification of psychologically salient contrast sets seems to be more pivotal to explaining the enhanced performance on DG variants of the 2-4-6 task, and it may also pave the way toward linking Oaksford and Chater's (1994) iterative counterfactual model of the standard task to our novel observation that the production of descending

triples mediates between DG instructions and rule discovery. In addition, the iterative counterfactual model can neatly capture the importance of multiple-hypothesis testing and creative search of the hypothesis space described by Vallée-Tourangeau et al. (1995) as being associated with task success, which is also supported in our present dataset in relation to the basic heterogeneity of triple generation that was observed with DG instructions. Further development of the iterative counterfactual model in terms of its capacity to explicate DG effects is clearly necessary (cf. Oaksford & Chater, 1994), and we view our present work as a first step in this direction. Direct recording and analysis of the actual hypotheses that people are testing with each generated triple is likely to be needed in future research in order to advance an understanding of the complex route that people traverse between hypotheses, triple instances, and rule discovery.

Original manuscript received 10 July 2003

Accepted revision received 20 January 2005

PrEview proof published online 25 July 2005

REFERENCES

- Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, *56A*, 1053–1077.
- Baron R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.
- Cherubini, P., Castelvécchio, E., & Cherubini, A. M. (2004). Generation of hypotheses in Wason's 2–4–6-task: An information theory approach. *Quarterly Journal of Experimental Psychology*, *58A*, 309–332.
- Evans, J. St. B. T. (1983). Selective processes in reasoning. In J. St. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 135–163). London: Routledge and Kegan Paul.
- Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, UK: Lawrence Erlbaum Associates, Inc.
- Farris, H., & Revlin, R. (1989a). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory and Cognition*, *17*, 221–232.
- Farris, H., & Revlin, R. (1989b). The discovery process: A counterfactual strategy. *Social Studies of Science*, *19*, 497–513.
- 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, MA: Cognitive Science Society.
- 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, UK: 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., 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. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.
- 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.
- 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.
- Oaksford, M. (2002). Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking and Reasoning*, *8*, 135–151.

- Oaksford, M., & Chater, N. (1994). Another look at eliminative and enumerative behaviour in a conceptual task. *European Journal of Cognitive Psychology*, 6, 149-169.
- Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 835-854.
- Poletiek, F. H. (2001). *Hypothesis-testing behaviour*. Hove, UK: Psychology Press.
- Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.
- Tukey, D. D. (1986). A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, 38A, 5-33.
- Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., 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.
- Vartanian, O., Martindale, C., & Kwiatkowski, J. (2003). Creativity and inductive reasoning: The relationship between divergent thinking and performance on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, 56A, 641-655.
- 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.