Hypothesis Generation and Testing in Wason's 2-4-6 Task

J. Edward Russo
Cornell University

Margaret G. Meloy
Penn State

The authors wish to thank Josh Klayman and Kevyn Yong for their comments on an earlier draft. J. Edward Russo is the S.C. Johnson Family Chair of Management, Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853-6201 (e-mail: jer9@cornell.edu). Meg Meloy is Associate Professor of Marketing, Smeal College of Business, The Pennsylvania State University, University Park, PA 16802 (email: mgm16@psu.edu). Correspondence regarding this article should be addressed to J. Edward Russo. The preparation of this article was supported by Grants SBR-9617873 and SES-00112039 from the National Science Foundation.

Hypothesis Generation and Testing in Wason's 2-4-6 Task

Abstract

The roles of hypothesis generation and testing in Wason's 2-4-6 rule discovery task were investigated by tracing the discovery process. This process consisted of an initial phase of major hypothesis shifts to identify the essence of the correct rule and a second phase of smaller hypothesis refinements. The familiar claim that both negative tests and falsifications lead to more successful rule discovery was supported, but only during the initial phase of hypothesis shifts preceding the generation of the essence of the correct rule. Neither negative tests nor falsifications facilitated success in the refinement phase that ended when the rule was announced. We tested two tactics for generating more hypotheses, delaying feedback and working in groups. Both tactics increased success in rule discovery and yielded a pattern of results compatible with that of individuals in both phases of the process.

Keywords: 2-4-6 task; confirmation bias; hypothesis generation; hypothesis testing; rule discovery

In 1960, Wason introduced the 2-4-6 task to study rule discovery. He adopted Popper's (1959) view that successful discovery depends on eliminating invalid hypotheses,[1] and that such elimination depends on a systematic effort to falsify hypotheses (see also Platt, 1964). Wason found, as expected, that the solution rate for rule discovery in his 2-4-6 task was quite low, just 21 percent. Compatible with the Popperian view, he attributed this result to a bias toward seeking confirming feedback from a test of the current hypothesis, typically a "yes" response. Instead, falsifying feedback was more efficacious because a single falsification could eliminate an incorrect rule. Klayman and Ha (1987) modified Wason's position by noting that individuals have only limited control over the feedback they receive, either a falsification or a verification of the tested hypothesis. In contrast, individuals fully control the kind of test that they choose to propose. Thus, according to Klayman and Ha, adopting a negative test strategy, one designed to disconfirm the current hypothesis and, therefore, to yield more of the superior falsifications, should be more efficacious than merely taking advantage of whatever negative feedback is encountered.

The present work aims to investigate the relative value of negative tests and falsifications by tracking the process of rule discovery and revealing their roles in that process. Further, we employ two tactics designed to increase the number of unique hypotheses that are generated and available for testing. These tactics serve two purposes. First, if they succeed in yielding more hypotheses, they should also provide more cases to track the process of testing those hypotheses. Second, they draw attention to the role of hypothesis generation in the overall success of rule discovery. Although some researchers have focused on hypothesis generation as fundamental to rule discovery (e.g., Cherubini, Castelvecchio, & Cherubini, 2005; Farris & Revlin, 1989a; and

Kareev & Avrahami, 1995), the tendency since Popper has been to emphasize hypothesis testing as the key to success (e.g., Dunbar, 1993; Koehler, 1994; McGuire, 1997; Poletiek, 2001).

The 2-4-6 Task

In the 2-4-6 task, the goal is to discover an underlying rule that specifies the relationship among three numbers. The discovery method is to propose triples of numbers, as many as desired, as tests of candidate rules. After each triple is proposed, the experimenter provides feedback, either "Yes," meaning that the participant's triple conforms to the rule, or "No," signaling that the proposed triple violates the rule. Participants are encouraged to continue testing their hypotheses about the identity of the underlying rule by proposing sets of three numbers until they are "highly confident" that they have discovered it. The announcement of the participant's guess of the rule concludes the task. However, in some studies, including Wason's original experiment, participants are allowed to continue to propose triples that they want to test and to announce new guesses of the rule until they have identified the correct one.

The task is deceptively simple. Brief instructions suffice; no practice is needed. The rule itself, any three monotonically increasing real numbers, is not complex.[2] Yet participants typically stop testing and conjecture the wrong rule a majority of the time. Indeed, as will be seen shortly, participants often fail for the reason Wason gave. That is, a pattern of seeking and receiving confirming feedback generates increased confidence in the current hypothesis and leads to an incorrect guess of the true rule.

For the purposes of our work, the 2-4-6 task has the essential property of requiring not only hypothesis testing, but also the generation of hypotheses that are tested. (This contrasts with Wason's (1968) more popular 4-card problem.) It is this joint task of hypothesis generation and

testing that qualifies the 2-4-6 task as an example of rule discovery and not merely of hypothesis testing.

At the same time that the 2-4-6 task is a genuine exemplar of rule discovery, it is quite particular in other ways. First, the feedback provided after each hypothesis test is accurate and precise.[3] In contrast, Dunbar (1995; see also Dunbar, 1999) examined the reasoning of molecular biologists as they evaluated the results of their experiments. He found that their initial reaction to disconfirming evidence was to question the validity of the evidence itself. Only when the biologists could not dismiss the empirical finding on methodological or other grounds did they relinquish their current hypothesis and seek a new one that could account for the contrary evidence.

A second special feature of the 2-4-6 task is that feedback from the experimenter is immediate, received within a few seconds of proposing a triple of numbers. Third, that feedback is also unlimited, or at least limited not by the experimenter but only by a participant's willingness to keep working at the task. Finally, participants receive only a simple yes/no response, lean feedback that provides no information that might help them diagnose the cause of the feedback or point toward a revision of the tested hypothesis. Again, contrasting the 2-4-6 task with most scientific rule discovery, the feedback from research studies is rarely immediate. It is costly as well and, therefore, not unlimited. Scientific rule discovery also offers the advantage of usually revealing something more about the underlying causal mechanism than a simple success or failure of the experiment.

Because these four feedback characteristics (viz., accuracy, immediacy, low cost, and leanness) differ from more naturally occurring situations, any one of them might cause the findings from the 2-4-6 task to generalize poorly to natural rule discovery. Despite these

particularities of the 2-4-6 task, it does preserve the main elements of the process of rule discovery: the generation of hypotheses based on existing evidence, the freedom to devise tests of those generated hypotheses, and the application of the results of those tests for further generation and testing until the underlying rule is identified. In all three respects, it mimics most naturally occurring environments of rule discovery.

The Confirmation Bias

One of the most robust outcomes of the 2-4-6 task is the apparent presence of a "confirmation bias"**.** This is the tendency to seek confirming evidence, which leads to feedback that contains too many "Yes" responses. In turn, this leads to "a failure to eliminate hypotheses".[4] Consider the following sequence of tests from the rule discovery trial of an experimental participant. Three pieces of information are reported for each round: the current hypothesis (i.e., the participant's best guess of the underlying rule), the proposed triple (i.e., the three numbers that constitute a test in the 2-4-6 task), and the experimenter's feedback (i.e., either "yes," a confirmation that the test conformed to the true rule, or "no," a falsification that should cause the current hypothesis to be rejected). The complete trial, five such three-part rounds, was: (i) add 2 [to the preceding number], {8, 10, 12}, yes [conforms to the rule]; (ii) every 2 numbers, {14, 16, 18}, yes; (iii) every 2 numbers, {20, 22, 24}, yes; (iv) every 2 numbers, {26, 28, 30}, yes; and (v) every 2 numbers, {32, 34, 36}, yes. After the last "Yes," the participant incorrectly conjectured as the underlying rule "every 2 numbers." In this trial, every tested triple accorded with the current hypothesis, only "Yes" responses were received as feedback, and not a single hypothesis was ever eliminated. This pattern of hypothesis-test-feedback illustrates Wason's argument about the danger of seeking only confirming feedback.

Feedback versus Test Strategy

A bias toward seeking confirming evidence has been reported in numerous forms and in numerous settings (Klayman, 1995; Nickerson, 1998). It has been shown to degrade task performance (e.g., Biais et al., 2000) and to increase confidence in the current hypothesis (e.g., Croskerry, 2002). Especially interesting are the reports of a confirmation bias in rule discovery for consequential tasks performed by professionals (e.g., Cloyd & Spilker, 1999; Croskerry, 2002; Mitroff, 1974; Muthard & Wickens, 2003).

Although the existence of a confirmation bias is well accepted, an important open question concerns the source of that bias. Does it reside in the feedback (i.e., too many "Yes" responses) or in the testing strategy (i.e., too many tests designed to yield "Yes" responses)?

Klayman and Ha (1987) argued that the source of the confirmation bias lies in the test strategy. They also simultaneously resolved confusions of terminology. First, a confirming or positive test is a triple that conforms to the currently held hypothesis or exhibits “confirmatory intent.” [5] Thus, if the currently held hypothesis is "add 2," the triple {8, 10, 12} would be a confirming or positive test. A negative test of this hypothesis would be any triple that violates it, such as {8, 10, 11} or {8, 10, 9}. Confirming feedback or a verification, which is based on the true underlying rule (in this case, "three increasing numbers"), is a response that supports the tested hypothesis. Thus, the hypothesis "add 2" would be verified either by the triple {8, 10, 12} followed by the feedback "Yes" (the triple conforms to the rule), or by the negative test {8, 10, 9} followed by a “No” (the triple violates the rule). In the latter case, “No” is a negative response but positive or verifying feedback. Similarly, a falsification or disconfirmation is feedback indicating a violation of the tested hypothesis. Again, if the hypothesis is “add 2,” the triple

{8, 10, 11} would receive the falsification "Yes." For a discussion of the various uses of the term confirmation bias, see Klayman (1995).

Klayman and Ha (1987) pointed out that a "No" response indicates a falsification only for positive tests. When a negative test is used, a "No" is expected and, therefore, constitutes a verification. Thus, the test strategy, not the feedback from the experimenter, determines whether there is a confirmation bias because it is the test strategy that participants choose or control**.** According to this line of reasoning, a confirmation bias involves the overuse of positive tests of the current hypothesis (or, more generally, tests reflecting confirmatory intent).

Klayman and Ha's (1987) explanation for why so many participants fail to discover the 2-4-6 task's underlying rule lies in the combination of a general tendency to use positive tests, and Wason's choice of an underlying rule (any three increasing numbers) to which a confirming approach is ill-suited. As the sample hypothesis "add 2" illustrates, participants' initial hypothesis or guess of the underlying rule is usually too specific, comprising a narrow subset of the actual rule (Kareev, Halberstadt, & Shafir, 1993). As a result, all positive tests of the participant's initial hypothesis are verified, since they are also instances of the true hypothesis (i.e., contained within the set of three increasing numbers). Such confirming feedback not only fails to prompt the consideration of alternative hypotheses, but also continually builds confidence in the current one. The tendency to use positive tests is far less damaging in tasks where the underlying rule is not so very much larger than the natural initial hypotheses as it is in the 2-4-6 task.

Research Strategy

Our empirical approach is to trace the rule discovery process, specifically by asking participants to state their current hypothesis, as well as the triple to be tested, for each round of the entire trial. This enables the identification of each test as positive or negative, and the resulting feedback as a verification or a falsification. The hypothesis-test-feedback rounds are then used to identify the transitions from one hypothesis to the next until the participant is satisfied that the correct rule has been identified. Our particular focus is on the role of negative tests and falsifications in prompting those crucial transitions. We construct a path to rule discovery that begins with the first test and proceeds through feedback and the transition to the next hypothesis, round by round, ending when a rule is announced.

Plan of the Studies

Both to increase the number of hypotheses available for testing and to verify the value of hypothesis generation (as well as hypothesis testing) to rule discovery, we test two tactics designed to help participants generate more candidate rules (i.e., hypotheses). These are delaying feedback and solving the 2-4-6 problem as a group. Both are described in detail below.

Two studies were conducted, one for each tactic. Both studies included a control group in which the 2-4-6 task was performed by individuals under Wason's (1960) original conditions. Rather than present each control group separately with its corresponding study, they are combined into a single "standard" condition. This grouping of individual participants increases sample size and, therefore, improves the stability of the estimates and the power of statistical

tests. It also allows us to begin with a unified baseline against which the results of each tactic can be contrasted.

Participants

For efficiency, we describe all of the participants in both studies. A total of 257 volunteers participated in the two studies. They were recruited via advertisements and paid for their time, or alternatively received course extra credit. The majority were undergraduate students (85%), with the remainder drawn from graduate students and staff. The latter were randomly assigned to conditions. Because there were no reliable differences in solution rates or other noteworthy variations across the subsamples (all $p > .50$), their results were combined. Of the 257 participants, 27 did not produce complete data (5 gave up prior to conjecturing a rule and 22 incompletely recorded their current hypotheses), leaving 230 participants in the final sample.

Standard Condition: Individuals

Our main goal from the analysis of individuals performing the standard Wason task is to identify the value of negative tests and falsifications during the process of rule discovery. A second goal is to establish a baseline against which to compare both the success rate and path of discovery of the two experimental tactics we employ for increasing hypotheses.

Method

The standard Wason (1960) rule, "three increasing numbers," was used in the current work. Participants read written instructions that were then reviewed orally by the experimenter about how the discovery task would proceed. They were told (a) that 2-4-6 was a triple of

numbers that conformed to the rule, (b) to propose both a hypothesis to be tested and a triple to test it, and (c) to continue to do so until they felt "highly confident" that they had identified the underlying rule. This last phrase was taken from Wason's (1960) original instructions. The tradeoff between accuracy and efficiency in discovering the rule was also accentuated by instructions that read, “There is no time limit, but you should try to discover the rule by proposing the minimum number of triples." All participants recorded on a tally sheet their current hypothesis about the underlying rule, the triple of numbers that they wanted to test, and the feedback received from the experimenter about whether the triple conformed to the rule. No feedback was provided as to the accuracy of the current hypothesis. A total of 49 participants were run individually in 30-minute sessions.

Results

Negative Tests and Falsifications

Of the 49 individual participants, 19 (39%) discovered the correct rule. However, the essential question was whether negative tests or falsifications contributed to the process of discovery. To arrive at an initial answer to this question, we computed, separately for the successful and unsuccessful participants, the absolute number of negative tests and falsifications along with their corresponding proportions (i.e., the proportion of negative relative to positive tests and of falsifications relative to verifications).

Successful participants used significantly more negative tests, both in absolute number and proportion (2.63 and .292), than did unsuccessful participants (0.23 and .037). For number, $t(47) = 6.12$, SE = .392 , $p < .001$; for proportion, $t(47) = 6.09$, SE = .0417, $p < .001$. (All tests are two-sided unless otherwise noted.) Thus, negative tests (relative to positive tests) had a

beneficial impact for successful rule discovery. Likewise, for falsifications both their absolute number and the proportion (relative to verifications) were significantly greater for successful (2.05 and .208) than unsuccessful (0.77 and .072) participants. For number, $t(47) = 2.46$, SE = .520, $p < .05$; for proportion, $t(47) = 3.17$, SE = .0429, $p < .01$. Thus, as has been claimed repeatedly, both the use of negative tests and feedback that falsifies the current hypothesis contributed to success in discovering the 2-4-6 rule.

Rule Discovery Process

Although the analyses above verified that negative tests and falsifications were beneficial to success in rule discovery, we were interested in tracing the process at a deeper level. Our specific focus was to trace the links from the type of test conducted, through the type of feedback received, to its consequent effect on hypothesis revision. We refer to these linking relationships as a "causal path" and begin by noting that crucial to successful rule discovery was the abandonment of incorrect hypotheses and the consideration of different ones. Successful participants listed significantly more distinct hypotheses (5.84) than did those who were unsuccessful (3.80) ($t(47) = 2.14$, SE = .953, $p < .05$). However, the essential issue was how both the test strategy and feedback influenced the transition away from one hypothesis toward a new one.

Hypothesis Transitions

To classify the transitions from one hypothesis to the next, our scheme followed the logic of Dunbar (1993). He distinguished between major shifts from one hypothesis to the next — "set[ting] a new goal to generate a new and/or alternative hypothesis" — as distinct from smaller

refinements of the current hypothesis — "small changes to their hypothesis" (p. 427; see also Schunn & Dunbar, 1996). We identified four types of hypothesis transitions, starting with Dunbar's distinction between major shifts and minor refinements. The first, hypothesis shifts, meant that $Hypothesis_{n+1}$ retained few, if any, of the elements of $Hypothesis_n$. These major transitions amounted to the rejection or abandonment of $Hypothesis_n$. Examples are the transitions "[consecutive] even numbers" to "numbers under 10" and "add up to < 100" to "get bigger by some amount." The second category, minor hypothesis changes or refinements, preserved a substantial part of the prior hypothesis. For instance, the hypothesis "numbers are increasing" was refined to "increasing negative numbers." Here the participant was testing whether the hypothesis "numbers are increasing" included negative as well as positive numbers. Refinements seemed to be used to determine the boundaries of the current hypothesis, or what is called "limit testing" by Klayman and Ha (1989, p. 603; also Tweney, 1990, p. 474). Both shifts and refinements were a source of alternative hypotheses and a part of a process of "progressive refinement" of the current hypothesis (e.g., Klahr & Dunbar, 1988). The third type of transition occurred when no specific hypothesis was stated. Instead participants wrote something equivalent to "just testing," without any indication of what they were testing. We termed these explorations. The fourth and final category was null transitions or no change between H_n and H_{n+1}. These were called retentions. Whenever a retention occurred, two different triples of numbers were used to test the same hypothesis on $Round_n$ and $Round_{n+1}$.

Two Phases of Rule Discovery

Until the idea of order was generated, participants should have exhibited at least some major hypothesis shifts as their current guesses of the rule were disqualified and they moved on

to different ones. Once they realized that the rule involved the order of the three numbers, participants may have made only the “small changes to their hypothesis” that Dunbar describes. As such, we divided each trial at the point where the participant first tested a hypothesis containing the idea of order, which we termed “generic order.” Examples included such hypotheses as "numbers go up in size," "assuming order of value," and "not descending." [6, 7]

The frequencies of each type of transition, partitioned by phase, are shown in Table 1. The pattern of frequencies was clear in two respects. First, hypothesis abandonment was largely confined to Phase 1. Of the 52 shifts, 47 occurred in this initial phase. (The five major hypothesis shifts in Phase 2 came from two individuals only.) Second, refinements were the main work of Phase 2, accounting for 71% of its hypothesis transitions. Although the partition into two phases yielded distinct differences between the frequencies of hypothesis shifts and refinements, these results needed to be interpreted within the context of the causal path described in more detail below.

Insert Table 1 Here

Path of Rule Discovery

The path of rule discovery is structured in Figure 1. The initial question is whether positive or negative tests were better at yielding the more useful falsifications (i.e., a “No” in response to a positive test or a “Yes” in response to a negative test). In Phase 1, the mean proportions of falsifications following positive and negative tests, respectively were .05 and .58, reflecting the often claimed superiority of negative tests. This difference was highly reliable, $t(63) = 5.60$, SE = .0947, $p < .001$. The same relative value of negative tests did not hold in Phase 2, where the mean proportion of falsifications following positive tests (.20) actually

exceeded that for negative tests (.14). (The difference did not approach significance in the reverse direction, $t(26) = 0.56$, SE = .107, $p > .50$.)

Insert Figure 1 Here

The next question was whether falsifications yielded more hypothesis changes than did verifications. In Phase 1, the proportion of verifications that led to hypothesis abandonment was .13, while the proportion of falsifications that led to a major hypothesis change was .37, a difference that was again statistically reliable ($t(63) = 2.34$, SE = .101, $p < .05$). In Phase 2, falsifications (.91) also proved superior to verifications (.81) in leading to hypothesis refinements, though this difference was not statistically reliable ($t(25) = 1.08$, SE = .093, $p > .25$). Thus, replicating the results for negative versus positive tests, we found that the predicted benefit of falsifications over verifications held in Phase 1 but not reliably in Phase 2.

Finally, how important were the major hypothesis shifts to discovering the approximate rule in Phase 1, and how important were the minor refinements to identifying the precise rule in Phase 2? Of the 28 participants who exhibited no shifts in Phase 1, only 4 (.14) correctly realized that generic order was the essence of the correct rule. In contrast, 15 of the 21 participants (.71) who exhibited at least one hypothesis shift achieved generic order. This difference was reliable (Fisher's exact test, $p < .05$). Turning to Phase 2, we examined the role of hypothesis refinements in discovering the exact rule. Of those 8 participants who exhibited not a single hypothesis refinement, only 1 identified "increasing" as the correct rule. In contrast, 14 of the 16 participants with at least one refinement achieved success. This difference was highly reliable (Fisher's exact test, $p < .001$). Successful rule discovery relied on abandoning hypotheses in Phase 1 and, assuming that Phase 1 had been successfully completed, refining the approximate hypothesis in Phase 2.

One of the complications of Phase 2, a phase characterized by refinements, was recognition that the same hypothesis could be framed both positively and negatively, and the choice of a frame seemed to be more arbitrary than strategic. The refinements that dominated Phase 2 all had well-defined complements that provided natural alternative hypotheses. This facilitated stating a refining hypothesis in either a positive way, such as "fractions are OK," or as its negative complement, "fractions not permitted." To participants, this choice must have seemed inconsequential, a matter of convenience not efficacy. As a result, we could find little, if any, systematic rationale to the positive or negative statement of a refined hypothesis. Indeed, the actual representation of the refinement in the participant's mind may have been a question ("Are fractions allowed?"). If so, the choice of a positive versus a negative test was more or less arbitrary and inconsequential. Note that we did not allow participants to state their current hypothesis as a question but instead required a proposed hypothesis so that all tests could be coded as positive or negative.

We concluded from these analyses that in Phase 2, negative tests were not more valuable than positive ones (and, therefore, falsifications conferred no special benefits over verifications). Indeed, the test strategy, important because it is what participants' control, may have been chosen arbitrarily in many cases. In contrast, negative tests in Phase 1 seemed to play a valuable role and did so because negative tests led to more falsifications which, in turn, led to greater hypothesis abandonment and ultimately, to the discovery of generic order.

Nature of Errors

The errors made by the 30 individuals who failed to conjecture the correct rule fell into three classes. The majority, 17, were made by "pure confirmers." These participants used only a

positive test strategy and received only verifying feedback. They stuck to a hypothesis, at least in part because they employed only positive tests and received only confirming feedback. (It was one of these pure confirmers who provided the example of Wason's "confirmation bias" that was used at the beginning of this report.) For 13 of the 17 pure confirmers, the announced hypothesis was also the participant's initial hypothesis. The remaining 4 engaged in some hypothesis refinement, but never shifted far from their original hypothesis. These 17, and especially the subset of 13, exhibited what has often been described in the literature as the "typical path to failure" (e.g., Klayman & Ha, 1989, p. 596).

The incorrect conjectures of the rule by eight other individuals were attributed to poor reasoning.[8] For instance, one participant, after eight rounds of positive, verified tests of "successive evens" or "successive integers," proposed the hypothesis "greater than 1000 and not successive." The "Yes" feedback led this participant to stop immediately and conjecture "numbers divisible by 1." All eight of the errors of these poor reasoners seemed to reflect some combination of confusion, frustration, and a commitment to solving the task that was either exhausted by the discovery process or limited in the first place.

The remaining five errors were distinctly different. In all cases, the test-feedback combinations could have been interpreted in multiple ways and a wrong way was chosen. For instance, one participant conjectured the conjunctive hypothesis "increasing and no more than one negative number." Three of these five misinterpreters generated generic order but either chose to conjecture an alternative hypothesis or, as in the above example, made "increasing" incorrect by conjoining it with another condition. These last five mistakes were more sophisticated than the confirming and poor reasoning of the other 25 errors made by individuals. At a minimum, these five did not seem to grasp at the first hopeful straw or stop prematurely.

Discussion

The analysis of the rounds of hypothesis-test-feedback that ended with the announced rule led us to partition the rule discovery process into two phases. During Phase 1, both negative tests and falsifications proved instrumental to success in generating generic order, something akin to arriving at a hypothesis that was "in the right ballpark" or what Dunbar (1995) calls the "overarching" hypothesis. Clearly reaching such a state of understanding requires abandonment of those generated hypotheses that are not in the right ballpark because they do not contain the essence of the rule. After generic order was generated and Phase 2 began, the refinement process seemed robust. It did not appear to depend on whether tests were positive or negative and, partly as a consequence, did not seem to depend on whether feedback was verifying or falsifying. Given that generic order had been generated (i.e., Phase 1 had been successful so that Phase 2 was reached), failure to discover the precise rule was uncommon and was attributed largely to blunders or confusion.

While examination of the path of rule discovery for individuals verified the importance of using negative tests and encountering falsifications, it also shifted our focus to a deeper appreciation for hypothesis generation and abandonment. When participants reject a hypothesis as "not in the right ballpark," they must then generate a new and substantially different candidate rule. How might we capitalize on what we understand about the discovery process to encourage more hypothesis shifts? One possibility for improving the success rate in the 2-4-6 task might be achieved by tactics that encourage individuals to generate and consider more hypotheses, especially more unique hypotheses that are different enough to eventually capture the essence of

the correct rule. Thus, it is to the hypothesis generation component of the rule discovery process that we turn.

Hypothesis Generation

Although much of the research involving rule discovery in general, and the 2-4-6 task in particular, has examined "hypothesis testing" (e.g., Dunbar, 1993; Klahr & Simon, 1999; Koehler, 1994; Oaksford & Chater, 1994; Poletiek, 2001), we now want to focus on hypothesis generation. The value of hypothesis generation in rule discovery has been explicitly acknowledged by several researchers (e.g., Adsit & London, 1997; Cherubini et al., 2005; Farris & Revlin, 1989a, 1989b; and Kareev & Avrahami, 1995).

In the context of Popperian falsification, Farris and Revlin (1989a, 1989b) advocate a "counterfactual strategy." Such a strategy focuses on the continual introduction of new hypotheses that are inconsistent with all prior ones. Using the 2-4-6 task, they found that successful rule discoverers could be distinguished from unsuccessful ones by the use of counterfactual hypotheses (Farris & Revlin, 1989a). In contrast, the proportion of tests that were negative (i.e., were not expected to conform to the rule) did not discriminate among participants.

Some researchers have taken the concept of new hypotheses a step further by requiring that one specific alternative hypothesis be directly contrasted with the current best guess via a test that discriminates between them (e.g., Platt, 1964). One useful source of alternative hypotheses is to "indicat[e] how the best guess might be changed" (Klayman & Ha, 1989, p. 600). Another appreciation of hypothesis generation, as opposed to hypothesis testing, is expressed by Cherubini et al. (2005) who focused on the initial hypotheses. "Only a few studies have focused on the content of discovery… and they [have] mostly addressed the revision of

existing hypotheses in the light of new data rather than the generation of initial hypotheses...” (p. 311). In sum, the need for considering multiple hypotheses, though not generally highlighted in the 2-4-6 task, has been increasingly recognized alongside hypothesis testing as essential to task success.

Increasing Hypothesis Generation

The most obvious tactic for increasing the number of hypotheses generated is to explicitly instruct participants to do so. Unfortunately, the direct method has yielded mixed results. Tweney et al. (1980, Experiment 3) found that instructing participants to generate two hypotheses for the 2-4-6 rule, to test those hypotheses separately, and to maintain two hypotheses throughout (i.e., always replacing an eliminated hypothesis with a new one) did not improve participants’ first guess of the rule and actually hindered their ability to discover it eventually. However, rather than generating genuinely different additional hypotheses, their participants often resorted to using "dummy" hypotheses that merely reiterated the initial one. Freedman (1992a, 1992b) also found that requiring participants to generate and record two hypotheses did not increase the solution rate (although a reanalysis of his data for individuals reveals a marginally significant advantage of the multiple hypothesis condition, $p < .10$, Fisher’s exact test). Freedman instructed both individuals and groups to generate either single or multiple hypotheses. Contrary to expectation, the multiple hypothesis condition did <u>not</u> exhibit greater accuracy than did single hypothesis instructions. We wonder, however, whether the multiple individuals in a group could suppress having multiple hypotheses available, even under single-hypothesis instructions.

Penner and Klahr (1996, Experiment 2) instructed participants to begin the 2-4-6 task by "list[ing] as many rules as they could that described this set of numbers," that is, that described the exemplar {2, 4, 6}. In contrast to Wason's "increasing numbers" rule, however, the target rule in this task was "sequential, even, numbers between 2 and 100, inclusive." As such, their results may not be fully comparable to Wason's task. Nonetheless, although their participants listed, on average, 3.8 hypotheses prior to the first triple tested, their solution rate (.29) was no higher than that of a standard group that was not instructed to begin by listing as many hypotheses as they could.

In contrast to the work cited above, other studies have reported success with instructing the generation of multiple hypotheses to increase the solution rate in the 2-4-6 task (Adsit & London 1997; Freedman & Endicott, 1997; Freedman & Jayaraman, 1993). Explanations for the variability in the usefulness of multiple hypotheses in rule discovery have focused on distorted memory processes and high cognitive load (Mynatt, Doherty, & Dragan, 1993; see also Freedman & Endicott, 1997). Whatever ultimately explains the mixed results of multiple hypotheses, it does seem clear that to succeed, the additional hypotheses generated should be genuinely different from each other and be accepted as serious competitors for the conjectured rule. Of course, this may be easier said than done (Gnepp & Klayman, 1992).

Finally, more hypotheses might be generated as a result of requiring participants to work longer. When Klayman and Ha (1989) required that 18 triples be proposed, a number larger than typically tested when participants can control when they stop, they found that accuracy (.52) was higher than normal. When Vallée-Tourangeau, Austin, and Rankin (1995) required that 15 triples be proposed as tests, accuracy was also a relatively high, .44. Finally, Wharton, Cheng, and Wickens (1993; Experiment 3) made their participants generate 5 tests as the minimum and

observed a success rate of .25. This last value fits the range of rule discovery rates usually reported in the standard 2-4-6 task. However, maybe more telling is the monotone increase of accuracy with the required number of tests. Because not all of the above studies reported the number of different hypotheses generated, we can only presume that the requirement of more tested triples induced more hypotheses.

Manipulations to Increase Hypothesis Generation

The inconsistent impact on accuracy of direct instructions to increase the number and quality of generated hypotheses prompted us to search for other manipulations. We found two. First, feedback can be withheld until multiple triples have been proposed (Vallée-Tourangeau et al., 1995). If, say, three tests must be proposed before any feedback is received, it might well appear inefficient and counterproductive to test the same rule three times. Instead, we expected that participants would generate three different hypotheses, one for each test. Vallée-Tourangeau et al. (1995) suggested that "delayed feedback may lead subjects to explore a greater variety of possible sequences of numbers" (p. 897). They did not find that such a delay led to greater success in rule discovery, but this null result may have been accounted for by the requirement that all their participants generate 15 tested triples before guessing the rule, whether they were assigned to the delayed or normal feedback conditions. Unfortunately, neither the number of hypotheses nor the number of different triples was reported.

Second, the 2-4-6 task can be solved by groups, with different individuals likely bringing different hypotheses to the discussion (Crott, Giesel, & Hoffmann, 1998; Freedman, 1992a; Hacker et al., 1990; Hoffmann & Crott, 2004; Miwa, 2004). In both cases, we sought a manipulation that would yield more hypotheses than typically generated by individuals under

standard conditions. As a result, in both cases, we had to change the task to some extent. The first tactic, delaying feedback, preserved the discovery by individuals. For this reason, it seemed more comparable to the standard task and was tested first.

Experiment 1: Delayed Feedback

In Experiment 1, individuals were required to propose multiple triples of numbers to be tested before any feedback was provided. We note that delay in feedback was first introduced as a manipulation in the 2-4-6 task by Vallée-Tourangeau et al. (1995), though without any benefit to accuracy.[9] Laughlin, Magley, and Shupe (1997) used this same tactic in a different rule induction task, and claimed that not receiving feedback after each test is more natural (see p. 273). In delaying feedback, we hoped that participants would see that testing one hypothesis three times (i.e., setting $\text{Hypothesis}_i = \text{Hypothesis}_{i+1} = \text{Hypothesis}_{i+2}$, but $\text{Test}_i \neq \text{Test}_{i+1} \neq \text{Test}_{i+2}$) would not be as informative as testing three different hypotheses. We expected that as feedback delay was lengthened to, say, a block of five tests before receiving feedback on any of them, the proportion of different hypotheses tested would also increase, leading to greater solution rates.

Method

Task

All participants performed the standard 2-4-6 task with the sole modification that in the experimental conditions, feedback was delayed and blocked. Participants received feedback about whether their proposed triples conformed to the underlying rule only after they had proposed either 2, 3 or 5 triples.

Procedure

The participants were 94 volunteers from the University community who were run following the basic procedure described above in the standard condition. They read the same written instructions, reviewed orally by the experimenter, and were required to record their current hypothesis, the triple they wished to test, and the feedback received from the experimenter. The only difference was that feedback was delayed until all of the hypothesis-test pairs in a block had been listed. Three different block sizes were used: 2, 3, and 5. We adopted a maximum block size of five because we believed that five would be the maximum number of hypotheses that a typical participant could generate without experiencing counterproductive frustration. However, this was an assumption without empirical evidence. The 94 participants were distributed as follows: 34 with block size of 2, 34 with block size of 3, and 26 with block size of 5. The last was lowest because more participants failed to provide a complete listing of hypotheses in this condition.

Results

Number of Different Hypotheses Tested

Although delaying feedback forced participants to generate multiple simultaneous tests, was there a corresponding increase in the total number of different hypotheses tested over the entire trial? For each block size, 1 (the standard condition, which is included for comparison), 2, 3 and 5, the mean total number of different hypotheses tested over the course of the discovery task was 4.59, 5.74, 6.71 and 9.42, respectively. The increase was statistically reliable as revealed by a linear regression of number of hypotheses on block size, both when the individuals were included as block size = 1 (est. $\beta = 1.19$, $t(141) = 5.36$, MSE = 14.31, $p < .001$) and when

only the three delay conditions were included (est. $\beta = 1.24$, $t(92) = 3.84$, MSE = 14.04, $p < .001$).

Accuracy

The solution rates revealed a similar increasing pattern. For block sizes 1, 2, 3, and 5, in order, the solution rates were: .39, .50, .53 and .62. This monotone increase was marginally reliable (Bartholomew's test of ordered proportions (Fleiss, Levin, & Paik, 2004), $\chi^2 = 3.91$, $p < .10$). We tested the assertion that participants who generated more hypotheses would also be more likely to solve the 2-4-6 problem. A logistic regression of accuracy on hypotheses yielded a significant effect (est. $\beta = .28$, $z = 4.87$, SE = .057, $p < .001$). The effect of unique hypotheses also obtained when block size was controlled for (est. $\beta = .29$, $z = 4.66$, SE = .063, $p < .001$). The mean number of different hypotheses was 8.11 for the 70 correct participants and 4.45 for the 73 incorrect ones, or 82% more hypotheses tested by correct participants. Just as for individuals solving the standard 2-4-6 problem, hypothesis generation was found to reliably predict success in rule discovery.

Blocking and the Two Phases

The blocking of feedback altered the structure of the process of rule discovery. In the standard condition, individuals engaged in multiple rounds of the same hypothesis-test-feedback cycle. When feedback was delayed, however, that tripartite structure was altered, and altered differently for each block size. Nonetheless, for completeness we attempted to analyze the data in Experiment 1 just as had been done for the standard condition. Thus, we partitioned the entire trial into Phase 1 and Phase 2 at the first mention of a hypothesis involving generic order. To do

so, we ignored blocking and recorded the specific position where the first mention of generic order occurred. For instance, if generic order first appeared as the second hypothesis in a block of three, the last hypothesis in that block became the first one in Phase 2. Once a continuous rather than blocked series of tests was assumed, the analyses of negative tests, falsifications, hypothesis transitions and sources of error could parallel those in the standard condition. We caution, however, that the structure of the blocked version of the 2-4-6 task was different, and that the following results can be only imperfectly contrasted to those who received feedback after every proposed triple.

Hypothesis Transitions

The frequency of each hypothesis transition is reported in Table 1, separated by both phase and block size. Several aspects of the data should be noted. First, hypothesis abandonment was more common when feedback was delayed than in the standard condition. In Phase 1, the relative frequency of these major shifts increased with block size, from .14 for blocks of 2, to .30 for blocks of 3, to .37 for blocks of 5. They also appeared in Phase 2, comprising .11 of transitions in the three delay conditions, a value similar to the .15 observed for individuals in the standard condition. Second, refinements were the dominant transition in Phase 2 (.70), just as for individuals (.71). Third and last, when feedback was delayed, it was not always easy to generate more candidate rules, especially to fill out large blocks. This may help explain the 31 explorations in the delay condition and that most of them (19) appeared in blocks of size 5.

The Role of Falsifications and Negative Tests

Accuracy. Although our data analyses emphasize the full path of rule discovery, we continued to examine the overall impact of negative tests and falsifications on overall accuracy. The proportion of negative tests was .13 for accurate participants and .03 for inaccurate ones ($t(92) = 3.51$, SE = .028, $p < .001$). A logistic regression of accuracy on the proportion of negative tests revealed a significant relation, holding block size constant (est. $\beta = 6.41$, $z = 2.79$, SE = 2.30, $p < .01$). For falsifications, the mean proportions for successful and unsuccessful participants were .21 and .10, respectively, also a reliable difference ($t(92) = 3.99$, SE = .028, $p < .001$). A logistic regression of accuracy on the proportion of falsifications, holding block size constant, again revealed a significant relation (est. $\beta = 6.31$, $z = 3.54$, SE = 1.78, $p < .01$). Thus, just as for the individuals in the standard condition, both negative tests and falsifications were systematically associated with the successful completion of the 2-4-6 task.

Path of Rule Discovery

We now tested the links in the causal path from negative tests to falsifications to hypothesis shifts to accuracy, beginning with the ability of negative tests to generate falsifications. Recall that with individuals in the standard 2-4-6 task, negative tests were far more effective than positive tests in yielding falsifications, but only in Phase 1. In Phase 2, their relative efficacy was actually reversed, though not significantly so.

As can be seen in Figure 2, the mean proportion of falsifications following positive and negative tests, respectively, were, .08 and .27 in Phase 1, and .34 and .26 in Phase 2. The predicted superiority of negative over positive tests in leading to falsifications was found in

Phase 1 ($t(113) = 2.25$, SE = .084, $p < .05$), but not in Phase 2 ($t(72) = 0.99$, SE = .081, $p > .30$). Both results duplicated those observed for individuals in the baseline condition.

Insert Figure 2 Here

We next tested for the power of falsifications (versus verifications) to lead to hypothesis abandonment. Recall, however, that because feedback was blocked, a falsification could not help to produce a new hypothesis until the next block. In Phase 1, the mean proportion of hypothesis shifts following verifications versus falsifications were, in order, .32 and .44 ($t(131) = 1.49$, SE = .081, $p > .10$). For Phase 2, the mean proportion of hypothesis refinements following verifications versus falsifications were, in order, .79 and .88 ($t(89) = 1.58$, SE = .057, $p > .10$). Though neither difference was statistically reliable, both exhibited the predicted direction. We acknowledge that the lack of a statistically significant difference may have been caused, in part, by the blurring of the identification of feedback from one test to its impact on the next hypothesis, something that was an inevitable consequence of blocking feedback.

What impact did hypothesis abandonment have on overall accuracy? In Phase 1, the presence of hypothesis shifts was significantly related to accuracy. Of the 39 participants who exhibited none of these major changes, only 15 (.38) discovered generic order. In contrast, 40 of the 55 participants (.73) who made at least one hypothesis shift arrived at generic order, a reliable difference (Fisher's exact test, $p < .01$).

In Phase 2, if an individual discovered generic order but failed to engage in hypothesis refinement, what was the probability of success in accurately discovering the rule? Of the 55 participants who discovered generic order, 2 failed to make at least one refinement. In sharp contrast to the standard condition, both of these individuals discovered the exact rule. As expected, 92% of those who proposed at least one refinement (N = 53) accurately discovered the

2-4-6 rule. Fisher's exact test revealed the improbability that the difference between the 1.00 and 0.92 was reliable (one-sided $p >$. 85).

Nature of Errors

Recall that errors fell into three categories: confirmation, poor reasoning, and misinterpretation. The 43 errors in the three delay conditions of Experiment 1 were categorized into the same three categories. The frequencies of each type of error were: pure confirmation,[10] 22; poor reasoning, 13; and misinterpretation, 8. With the exception of this last error, the other categories of error seemed unsophisticated and, especially for poor reasoning, avoidable. Interestingly, the incidence of the unsophisticated errors declined only at the largest block. For block sizes 2, 3, and 5, the number of pure confirmers and poor reasoners totaled 13, 14, and 8 respectively.

Individual Differences in Generation of Multiple Hypotheses

Because of the nature of the delayed feedback manipulation, it became important to understand how quickly individuals caught on to the inefficiency of testing the same hypothesis multiple times within a block. As such, we created an individual difference measure that asked whether participants perceived the value of generating multiple hypotheses immediately (i.e., in the first block of tests), learned it later in the trial, or never used multiple tests at all. We defined a <u>high hypothesizer</u> as a participant who tested the maximum number of different hypotheses in a given block of triples (e.g., 3 different hypotheses when the block size was 3).[11] All participants were classified into three groups. The <u>initial high</u> hypothesizers tested a high number of unique hypotheses in the first block of triples. The <u>later high</u> hypothesizers did not test a high number of

unique hypotheses until some block after the first one. The remaining group, never high hypothesizers, did not test a high number of unique hypotheses in any block of tests.

There were 65 initial high, 11 later high and 18 never high hypothesizers. Over two-thirds of the participants whose feedback was delayed and blocked seemed to realize quickly the value of generating multiple hypotheses to accompany their enforced multiple tests. Did the 11 later high hypothesizers attain the same success rate as those participants who tested multiple hypotheses right from the start? This is a "better late than never" question. Alternatively, the participants who were slow to generate multiple hypotheses may have adopted this tactic only half-heartedly and benefited from it at a level below that of the initial high group. The accuracy rates for initial, later, and never high hypothesizers were .63, .73, and .11, respectively. Thus, the later high hypothesizers fully achieved the solution success of those participants who seemed to perceive the value of testing multiple hypotheses right from the beginning. The accuracy rates for both initial and later high hypothesizers exceeded that of the never high participants (Fisher's exact test for initial high, $p < .001$; and for later high, $p < .01$).

Efficiency of Hypothesis Abandonment

As a final parallel to the individuals in the standard condition, we checked whether the increase in the total number of hypotheses tested reflected more efficient testing as signaled by fewer contiguous tests of the same hypothesis, that is, by fewer consecutive retentions. The mean numbers of consecutive retentions were 2.42 for the standard (no delay) condition and for the three delay conditions, in order, 2.08, 1.77 and 1.32. Using a simple linear regression to test for a linear decline, est. $\beta = -.279$, $t(141) = 3.03$, SE = .092, $p < .01$. Thus, blocked feedback

succeeded in part by leading participants to work more efficiently, as evidenced by fewer consecutive retentions.

Discussion

The simple technique of delaying feedback succeeded in facilitating rule discovery, and seemed to do so through the predicted generation and testing of more hypotheses. Unfortunately, it is not possible to determine whether delaying feedback's impact on accuracy was caused solely by superior hypothesis generation or by the joint impact of superior hypothesis generation with superior hypothesis testing. For instance, were there fewer consecutive retentions with increasing block size because participants felt pressure to generate a different hypothesis for each test in a block? Or, alternatively, was testing superior because all of the feedback from the previous block was brought to bear on the evaluation of each hypothesis in the new block, even if a particular test had not been intended to test every hypothesis?

Because Experiment 2 altered the structure of the 2-4-6 task, details have to be interpreted with caution. Nonetheless, many results conformed to expectations, and those that did not were consistent with the special circumstance of delaying feedback. Like the standard condition, the same distinction in process between Phases 1 and 2 was found in Experiment 1. For instance, the power of negative tests to yield falsifications was clear in Phase 1 but not apparent in Phase 2. Similarly, the sources of errors and their relative frequencies were similar to those reported in the standard condition. Even the few differences in results between the two studies seemed well explained by the presence of immediate versus delayed feedback. For instance, there were more explorations and hypothesis shifts as most participants (especially the initial high hypothesizers) struggled not to test the same hypothesis more than once in a block.

Similarly, the mean number of consecutive retentions was low, approaching the minimum of 1 when block size equaled 5, presumably because of the presence of competing hypotheses.

Experiment 2: Groups

The second technique that we employed for generating more hypotheses was to ask multiple individuals to solve the 2-4-6 problem. Working in groups should achieve our goal of generating meaningful multiple hypotheses so long as different individuals generate different hypotheses (Crott et al., 1998; Laughlin et al., 1997; Okada & Simon, 1997), and they share those hypotheses through discussion.

Multiple hypotheses should appear especially at the start of a group's rule discovery process. Crott et al. (1998) showed that the number of new hypotheses was likely to be significantly higher in the first half of a 2-4-6 solution trial. Furthermore, this partition into halves was the only factor that explained the number of new hypotheses generated. In the current work, groups were expected to surpass individual accuracy because if even one member generated the correct hypothesis, the subsequent discussion and testing should enable the group as a whole to adopt it as their conjecture (Laughlin & Hollingshead, 1995). If this claim holds true, the superiority of groups should improve with group size. For this reason we tested the ability to solve the 2-4-6 problem of both small (3-person) and large groups (5-6 individuals).

Method

Task

The standard Wason (1960) rule discovery task was used, with a procedure that, as closely as possible, matched that for the individual participants. For instance, substituting "your

group" for "you," the instructions read "There is no time limit, but your group should try to discover the rule by proposing the minimum number of triples." All participants were instructed to record on a private tally sheet their current (private) hypothesis about the underlying rule and the triple of numbers that they would have tested if completing the task individually.[12] Each group also had its own large tally sheet visible to all on which the experimenter recorded the consensus hypothesis that the group had arrived at through discussion and the triple they selected for testing. The group tally sheet also posted the feedback the group received about whether the triple of numbers conformed to the rule. The task for a group ended when the spokesperson announced the group's conjecture of the rule. The designated spokesperson for the group was the individual whose next birthday was soonest.

Procedure

Of the 87 participants who participated in Experiment 2, 33 were run in 11 small (N = 3) groups and 54 in 10 large (N = 5-6) groups. In scheduling participants, an effort was made to control for possible gender effects. All groups of 3 were single-gender. Large groups were either single-gender or comprised of approximately equal numbers of males and females, that is, always between 2 and 4 of the same gender. Groups were allocated 1-hour time slots. Members were seated around a small circular conference table. A consensus of all group members had to be reached about both the hypothesis and triple/test before either could be proposed by the spokesperson to the experimenter (and feedback received). After the triple had been proposed by the group's spokesperson, each individual was asked to write down "your own personal hypothesis" about the rule. This latter record enabled us to determine the number of additional hypotheses that were active at any given time.

Results

Numbers of Hypotheses Generated

Fundamental to all that follows was whether groups generated and tested more different hypotheses than individuals. The mean total number of different hypotheses tested during a trial was 6.09 for small groups and 7.20 for large groups, a difference that was not statistically reliable ($t(19) = 0.75$, SE = 1.48, $p > .40$). Because a group discussion preceded agreement on the hypothesis to be tested, it was highly likely that more than a single hypothesis was discussed. For this reason each group's verbal exchange was audio taped. Unfortunately, the entire trial of one large group was not sufficiently audible for the number of hypotheses to be identified and had to be dropped from this analysis. However, we calculated the mean number of hypotheses discussed for each triple proposed across all remaining rounds for the remaining groups whose tapes we could understand. The mean numbers of discussed hypotheses per round were 2.58 for small and 2.94 for large groups, respectively. Their difference, though in the expected direction, did not approach statistical significance, t(18) = 0.76, SE = 0,74, p > .40. However, the grand mean of 2.74 hypotheses discussed per round was reliably greater than the minimum of 1 hypothesis (t919) = 6.99, SE = 0.25, $p < .001$) and suggested that groups generated more hypotheses than individuals.

Accuracy

Accuracy increased with group size. The proportions of successful participants were, for group sizes 3 and 5-6, in order: .82 (9 of 11) and 1.00 (10 of 10). The difference between the small and large groups was not significant (Fisher's exact test, one-sided $p > .20$)[13]. Indeed, what

seemed more important about the two levels of accuracy was how high they both were. For example, compared to individuals in the standard condition (.39), the likelihood of either small or large groups discovering the rule more than doubled.

The relation between accuracy and the number of hypotheses was weak at best because both small and large groups were so accurate. However, for descriptive purposes we can include the individuals in a logistic regression, with the result that the number of tested hypotheses is found to be a statistically reliable predictor of accuracy (est. $\beta = .259$, $z = 2.77$, $SE = .094$, $p < .01$). Descriptively, one additional hypothesis tested (in the observed range) was estimated to cause the solution rate to increase .234 ([1.00 - .388]/[7.20 – 4.59]).

Nature of Errors

Because there were so few errors, it will prove useful to reveal them before moving to the results for negative tests, falsifications, and the path of rule discovery. All 21 groups succeeded in generating generic order, and only two small groups failed to discover the rule. Both committed errors of misinterpretation, choosing the wrong interpretation and failing to test the other possibility. For example, one group interpreted the "No" responses to two tests, {-5, -313, -28} and {-1, -52, -17}, as indicating "only positive numbers allowed" instead of indicating order. Thus, groups made fewer and more sophisticated errors than individuals.

A sense of the process that yielded the high success rate for groups can be obtained from a sample transcript of the discussions of one of the small groups. We report for each round not only the tested hypothesis, the tested triple, and the feedback received, but also the verbatim discussion (in italics) that led to the selection of the stated hypothesis and its test. The seven rounds of Phase 1 were: (i) "*Could be anything. All even. Test even numbers high up. Numbers in*

the 50's," even numbers, {50, 52, 54}, yes; (ii) "*Try odd numbers. Random numbers not going up by 2*", even numbers, {11, 27 33}, yes; (iii) "*Numbers below 100?*," numbers below 100, {101, 102, 103}, yes; (iv) "*Non-whole numbers. Three random numbers. Leave some whole, some non-whole,*" whole numbers, {3.5, 51, 418}, yes; (v) "*Maybe it's positive numbers; try negative numbers. Try zero*," positive numbers, {-.5, -318, -28}, no; (vi) "*Try zero. Try just one negative. Try a really big number*," positive numbers, {0, 25, 3427}, yes; (vii) "*They're all positive. Try a fraction. Least to greatest. Going up and down,*" positive numbers in ascending order, {3427, 0 25}, no. Note that "least to greatest" was an expression of generic order and represented the successful completion of Phase 1.

The stated hypothesis seemed sometimes to be as much a matter of convenience as strategy. Thus, in the first round, the discussion focused on two potential characteristics of the rule, "even numbers" and "numbers high up -- numbers in the 50's." Yet the stated hypothesis was only "even numbers," with no mention of numbers in the 50's. However, this second characteristic was captured in the tested triple, {50, 52, 54}. Similarly, in the other rounds the consensus hypothesis and its complement seemed essentially equivalent and the choice between them arbitrary. Thus, in Round ii "Try odd numbers" led to the consensus hypothesis "even numbers" with the test {11, 27, 33}, where "odd numbers" could just as easily been the tested hypothesis. This arbitrary choice of the consensus hypothesis over its complement undermined the analyses of negative tests, falsifications, and the path of rule discovery. Nonetheless, we report these analyses for completeness.

Path of Rule Discovery

Neither the proportion of negative tests (.20 for accurate; .35 for inaccurate) nor the proportion of falsifications (.27 for accurate; .44 for inaccurate) had, by themselves, any impact

on accuracy. A more complete picture of the impact of falsifications and negative tests was obtained by tracing their path of influence, as shown in Figure 3.

Insert Figure 3 Here

Did negative tests produce more falsifications than did positive tests? For Phase 1, the mean proportions exhibited the predicted relation, .12 for positive tests and .54 for negative tests, a significant difference ($t(26) = 2.60$, SE = .162, $p < .05$). In Phase 2, the effect was reversed (.25 for positive tests and .15 for negative tests), though not significantly so ($t(32) = 0.91$, SE = .110, $p > .35$).

Did falsifications lead to more changes of hypothesis than did verifications? The two mean proportions in Phase 1 were .19 for verifications and .15 for falsifications, a small and non-significant reversal of the predicted direction ($t(29) = 0.36$, SE = .110, $p > .70$). This is the only case where falsifications did not lead to more hypothesis shifts than did verifications. Also the proportions were relatively low, a result of the low frequency of hypothesis shifts and, especially, the high frequency of hypothesis refinements (see Table 1). In Phase 2, the proportions of verifications and falsifications were .88 and .96, respectively. Their difference was directionally predicted but not statistically reliable ($t(31) = 1.21$, SE = .066, $p > .20$).

Finally, did more hypothesis abandonment yield greater discovery of generic order, which was the successful conclusion of Phase 1? No test of the effect of hypothesis shifts was possible because <u>all</u> groups successfully discovered the approximate hypothesis of generic order, whether they exhibited major hypothesis shifts (n = 11) or not (n = 10). A similar difficulty occurred in Phase 2. Of the 21 groups, 19 had at least one hypothesis refinement, with an average of 5.47 refinements in Phase 2. All 19 of these discovered the precise rule. The two groups that

failed to discover the rule had, on average, just 1.50 refinements in Phase 2, a difference that was not reliable ($t(19) = 2.46$, SE = 1.61, $p > .20$).

The results from the various links in the path of rule discovery for groups differed somewhat from those in Figures 1 and 2. As we asserted earlier regarding the sample transcript of one small group's discussion, the presence of multiple discussed hypotheses seemed to blur the distinction between negative and positive tests and between falsifications and verifications. In the presence of the multiple hypotheses that individuals discussed simultaneously, a negative test of one may have been a positive test of another and what verified one may have falsified another. Thus, the link between a tested hypothesis and its feedback may have been too loose to enable a definitive test of the relative effectiveness of negative versus positive tests and of falsifications versus verifications.

What Made Groups Different from Individuals?

Why were groups so successful relative to individuals (accuracies of .90 and .39, respectively)? In light of the value of more hypotheses to accuracy observed for individuals in both the standard and delayed feedback conditions, the answer seemed to lie in the greater number of hypotheses that groups discussed in preparation for each test. The mean of 2.74 hypotheses almost certainly exceeded the number that individuals privately considered in each round. Of course, no comparison to individuals could be made, because individuals were not instructed to speak their thoughts aloud.

Did groups also test hypotheses more effectively? One indicator of the effectiveness of hypothesis testing was the ability to discard a hypothesis quickly and consider a new one. Just as earlier, we computed the number of consecutive retentions, that is, the mean number of tests of

the same hypothesis conducted without interruption. For individuals, this mean was 2.42 consecutive tests. The corresponding values for small and large groups, respectively, were 1.57 and 1.22. The differences among these three values were statistically significant, $F(2, 67) = 3.45$, MSE = 2.32, $p < .05$. We note that the number of consecutive retentions for groups may have been shorter because with group members advocating different hypotheses, there may have been social pressure to take turns in selecting a hypothesis to be tested. However, though the source of groups' lower number of consecutive retentions may have reflected group dynamics as much as a superior hypothesis testing strategy, it was still a salient characteristic of groups relative to individuals.

Discussion

Groups succeeded in discovering the 2-4-6 rule both because of superior hypothesis generation and superior hypothesis testing. Relative to individuals, groups produced and discussed more different hypotheses and they abandoned the invalid ones more quickly. This more efficient abandonment was facilitated by negative tests, but not by falsifications. However, just as for the standard and delayed feedback conditions, the superiority of negative over positive tests was confined to a first phase defined by the search for generic order. A second phase devoted to refining generic order did not consistently reveal the same value of a negative test strategy or of receiving falsifications. Finally, there were no errors of the purely confirming kind, or of the poor reasoning that occurred for individuals and delayed feedback participants. Instead, the only two errors made by groups reflected relatively sophisticated misinterpretations of the test-feedback combinations.

Groups versus Delayed Feedback

Working in groups seemed to be a more successful intervention than delaying feedback. From individuals to large groups, the number of different hypotheses tested increased from 4.59 to 7.20, which was not as large as the impact of delaying feedback (from 4.59 hypotheses for individuals to 9.42 for blocks of size 5). However, the impact on accuracy was substantially greater for groups. All large groups discovered the rule, whereas the highest success rate in Experiment 1 was only .62 for blocks of 5.

The sensitivity of accuracy to the number of different hypotheses can be captured by the ratio of the difference in accuracy to the difference in the number of hypotheses tested. As computed earlier for groups, this sensitivity or impact measure was a .234 increase in accuracy for every additional hypothesis tested. In contrast, for delayed feedback the corresponding value was only .047 ([.615 - .388]/[9.42 – 4.59]). Thus, each additional hypothesis that was tested in Experiment 2 yielded about five times the increase in solution rate as did an additional hypothesis in Experiment 1. Of course, again it must be noted that groups generated and discussed many more hypotheses than they tested. Thus, the advantage of groups over delayed feedback, though quite real in accuracy, cannot be assessed so precisely by the sensitivity of accuracy to number of hypotheses tested.

Further support for this comparison was provided by the pattern of consecutive retentions, or how quickly participants discarded a hypothesis and moved on to a different one. Again, the impact of groups was greater, with the number of consecutive retentions always approaching the minimum of 1 (1.57 for small groups and 1.22 for large ones). Only block size = 5 yielded a mean (1.32) at this level. Groups clearly made better use of each hypothesis test.

General Discussion

We consider two issues. First, the joint emphasis on hypothesis generation and testing raises the larger topic of the paradigm for investigating rule discovery. The second is generality, namely how much of what has been learned through studying the 2-4-6 task generalizes to more realistic settings. We address each of these issues in turn.

Hypothesis Generation

The philosophy of science tradition in rule discovery has focused primarily on hypothesis testing rather than on the generation of hypotheses as candidates for testing (Platt, 1964; Popper, 1959). The study of the 2-4-6 task has been closely linked to this tradition. Indeed, the task was designed by Wason (1960, 1962) to highlight the need to follow the Popperian dictum to seek falsification. It must be acknowledged that much progress has been made in understanding how people solve (and fail to solve) the 2-4-6 problem by analyzing hypothesis testing alone. The Klayman and Ha (1987) work on test strategy, especially when to employ positive versus negative tests, is a prominent example.

In spite of this tradition's concentration of effort on hypothesis testing in general and falsifications in particular, the need for generating multiple hypotheses has been increasingly recognized as just as important to task success as is hypothesis testing (e.g., Adsit & London, 1997; Farris & Revlin, 1989a, 1989b; Kareev & Avrahami, 1995). Our two tactics for enhancing the generation of multiple hypotheses substantially improved rule discovery. In this regard, we join the stream of researchers who value both hypothesis generation and testing.

A focus on hypothesis generation also points toward the more complete account of rule discovery provided in the literature by cognitive psychologists (e.g., Dunbar, 1993; Klahr, 2005). Among the latter, we note the compatibility between our results and the problem-solving perspective of Dunbar (1993, 1995). He claims that different goals motivate different strategies and behaviors, distinguishing particularly between the "goal to generate a new and/or alternative hypothesis" and the goal of testing a current "specific (or 'local') exemplar hypothesis" (Dunbar, 1995, p. 427). Tests of the former correspond closely to our major hypothesis shifts, while the latter seems to be achieved by the refinements of our Phase 2. Further, Dunbar's core claim is that before progress can be made, the accumulation of inconsistent evidence must lead to a change of goals from refinement to finding a new overarching type. Although our data do not enable a test of this goal-based view of the rule discovery process, they are consistent with it.

The goal-based perspective of rule discovery has not yet been empirically tested in the standard 2-4-6 task. However, Vollmeyer, Burns, and Holyoak (1996) have suggested how this framework might explain the success of Tweney et al.'s (1980) DAX/MED manipulation. To shift participants' focus from confirmation, Tweney et al. labeled feedback "DAX" when the tested triple conformed to the rule and "MED" when it did not. Vollmeyer et al. explain the resulting improvement in performance in terms of goals. Specifically, the DAX/MED instructions may shift the goal from confirming the current rule to the more general goal of "finding a rule that distinguishes the two labels" (i.e., DAX and MED; Vollmeyer et al., 1996, p. 98). These authors go on to link the differences between specific versus general goals to a shift from searching the space of experiments (or instances) to searching the space of rules (Klahr & Dunbar, 1988). Holding aside this larger issue of dual spaces, Vollmeyer et al. (1996) explicitly

link the goal-based framework of cognitive science to the 2-4-6 task, albeit speculatively and not empirically.

For our purposes, the most salient aspect of the goal-based view is that it enables hypothesis generation to receive as much attention as hypothesis testing. In our case, the two-part goal structure advocated by Dunbar seems to map well into our partition of the observed rule discovery process into two phases. At the same time, we note that the goal-based view's respect for hypothesis generation in no way diminishes its sensitivity to the value of hypothesis testing. For instance, it fully accepts the goal of optimizing the value of feedback obtained by choosing a positive versus a negative test (Klayman & Ha, 1987).

Generality

How generalizable is the 2-4-6 task and the findings derived from studying it? The doubt conveyed by this question was first raised by Wetherick (1962) who "criticized [Wason's results] on the grounds that the task set is in important respects untypical of problem solving situations in general (p. 246)." [14] Recall that we characterized its feedback as accurate, immediate, low cost, and lean. None of these characteristics seem common in the real world tasks of rule discovery. We do not dispute that sometimes feedback is accurate, that sometimes it is immediate, and that sometimes it is even low cost. However, rarely is it lean. That is, the 2-4-6 task is "knowledge-lean" (Okada & Simon, 1997, p. 113) or content-independent (e.g., Chronicle, MacGregor, & Ormerod, 2004). An experimental participant learns what happened, that is a verification or falsification, but nothing about why it happened. And there is nothing in the context of the situation to offer further insight. Such tasks may need to be approached in fundamentally different ways than context-rich ones (e.g., Lovett, 2002).

Besides the unusual nature of the feedback, Wason's choice of a broad rule ("very general" is his phrase; 1962, p. 250) and the 2-4-6 exemplar particularize it further. As Klayman and Ha (1987) point out, this exemplar elicits initial hypotheses that are almost always much narrower than the rule that participants must eventually uncover. Thus, both the task's content-independence and further particularity suggest that it is certainly not straightforward to generalize what has been revealed in the above studies to natural situations of scientific discovery or problem solving in which uncertainty and richness of context are typical.

Yet even considering the special characteristics of the 2-4-6 task, we see two candidates for generalization to other tasks. First, we found that the value of both negative tests and falsifying feedback was limited to the first phase of the rule discovery process, that is, to the discovery of the overarching or generic rule. The implication is that these two widely accepted elements of good hypothesis testing may be limited to tasks where the initial hypothesis is relatively far from the correct rule. Whenever the beginning guess contains the essence of the underlying rule, hypothesis refinements alone may prove sufficient and there may be no advantage of negative over positive tests or of falsifying over verifying feedback. Thus, the partition of the rule discovery process into two phases may be general and so may be the limitation of the value of negative tests and feedback to the first phase. Both claims, of course, are subject to empirical verification.

Second, there may be a substantial set of tasks where little progress can be made until the "overarching" hypothesis has been generated. Even the best hypothesis testing may offer limited assistance. In such cases, techniques for aiding hypothesis generation may prove necessary. Both delaying feedback and working in groups may contribute to the essential task of hypothesis generation. These two tactics have the merit of generality, being relatively easy to adapt to a

variety of rule discovery or other problem-solving tasks. Of course, how they succeed in aiding hypothesis generation in any particular task remains an open question until tested.

If there is an overarching conclusion from the above results, it is that generating hypotheses is as important to rule discovery as testing those hypotheses. It would be ironic if Wason's devising of the 2-4-6 task to demonstrate the value of a particular form of testing, namely falsification, were to point equally to the need to generate candidate hypotheses.

Footnotes

[1] In our work, hypothesis means a candidate rule. However, hypothesis is usually more broadly defined. For instance, Sanbonmatsu, Posovac, Kardes, and Mantel (1998, p. 197) state, “a hypothesis is a perceived possible estimation, interpretation, evaluation, exploration, rule, or solution. More specifically, hypotheses are beliefs about the possible relation between variables, values, objects, or events.”

[2] Most studies follow Wason and set the rule to be any three ascending numbers. However, the 2-4-6 task has been altered to include other target rules. For example, Penner and Klahr (1996) use the narrow rule, "sequential even numbers between 2 and 100, inclusive."

[3] Some researchers have injected error into the 2-4-6 task by providing incorrect feedback (Gorman, 1989; Penner & Klahr, 1996). Although this may make the task more realistic, we preserve the accurate feedback of Wason's original task.

[4] The 2-4-6 task has, rightly or wrongly, become a popular demonstration of the dangers of a bias toward seeking confirmation (e.g., Russo & Schoemaker, 2002) and of why it is important for a smart person to try “to prove himself wrong” (Brandenberg & Nalebuff, 1996).

[5] Some authors include under positive tests, not only cases where the test matches the hypothesis or condition (called positive hypothesis tests), but also those where a target event occurs that enables a check of whether the hypothesized condition was met (called positive target tests; see, e.g., Wharton, Cheng, & Wickens, 1993). That is, one can start with the focal condition, a positive hypothesis test, or with the predicted result of that condition, a positive target test. In the 2-4-6 task, the hypothesis $X_1 + X_2 = X_3$ suggests such positive tests as 3+6=9 and 10+20=30. Both would receive the verifying feedback, “Yes.” Alternatively, a participant could gather all

positive targets and try to induce hypotheses that might have generated them. In the above example, the positive targets are {2, 4, 6}, {3, 6, 9}, and {10, 20, 30}. Various hypotheses can then be tested for consistency with the positive targets, so $X_1 + X_2 = X_3$ would pass that test, but {X, 2X, 3X} would fail. In the present work positive tests refer only to positive hypothesis tests.

[6] The essence of generic order was an ordinal relation among the numbers. This was not always indicated by the word "increasing" itself. For instance, the common hypothesis "add 2" (i.e., an arithmetic series with an increment of 2) was sometimes stated as "increasing even numbers." It was frequently apparent from the sequence of tests that participants meant "increasing consecutive even numbers." This hypothesis possessed the properties of an interval rather than an ordinal scale. Thus, even though it included the word "increasing," this hypothesis did not qualify as an example of generic order.

[7] Recall that participants were allowed to test any hypothesis, not necessarily the one that they currently believed was most likely. Thus, the observed changes from H_n to H_{n+1} might not have reflected corresponding changes in the participants' current best guess of the rule. We did not require our participants to list both the hypothesis they were testing and also their best guess of the rule, as did, for example, Klayman and Ha (1989).

[8] For example, one of these participants proposed 12 consecutive positive tests of "increasing by 2." The next test was negative, {80, 90, 100}, and received the falsifying feedback, "Yes." After two more positive tests of "increasing by 10," this participant stopped proposing triples and conjectured as the rule "increments of 2 or 10." The total time of this trial (including the time required to read the instructions) was 22 minutes, well above average, so this participant was not simply indolent. The other cases of poor reasoning differed in detail but not in quality.

[9] Vallée-Tourangeau et al. (1995) used only one block size, 15. That is, they required participants

to generate 15 triples without feedback. In their study, feedback was provided after the 15 tests, and participants were required to conjecture the rule whether they were highly confident or not. Compared to a standard condition with feedback after each of the 15 triples, the delay had no significant impact on the solution rate (.44 without delay versus .45 with delay). Nonetheless, delay of feedback has been found to improve behavior in some tasks. See, for example, Kudadjie-Gyamfi and Rachlin, 1996.

[10] Not all confirming participants tested only a single hypothesis. Instead, 2 of the 22 confirmers abandoned an earlier hypothesis for one they adopted later, and then exhibited a confirmer's pattern of only positive tests, only verifying feedback, and then conjectured this later hypothesis.

[11]Two adjustments were made for comparability across block size: (a) for block size = 2, the maximum of 2 hypotheses had to be tested in two successive blocks for a participant to be classified as a high hypothesizer; and (b) for block size = 5, 4 different hypotheses were sufficient to qualify a participant as a high hypothesizer.

[12] Despite being instructed to record their own hypothesis, we found that most participants either provided incomplete data or merely adopted the hypothesis being testing by the group. As such, these data could not be analyzed.

[13] If the individuals' accuracy in the standard condition is included (.39), Bartholomew's test of ordered proportions revealed that the observed increase in accuracy was statistically reliable, $\chi^2 = 16.53$, $p < .01$.

[14] Wason (1962) replied that his "task was not intended to be typical of problem solving situations in general. His aim was to investigate the acquisition of evidence on inductive reasoning. The task used was deliberately biased… The point of the experiment was not *whether* the subjects eventually discovered this rule (which was necessarily very general), but *how* they

set about trying to discover it" (p. 250; italics in original).

References

Adsit, D. J. & London, M. (1997). Effects of hypothesis generation on hypothesis testing in rule-discovery tasks. *Journal of General Psychology, 124* (1), 19-34.

Brandenberg, A. M. & Nalebuff, B. J. (1996). *Co-Opetition: A revolution mindset that combines competition and cooperation: The game theory strategy that's changing the game of business*. New York (Doubleday).

Biais, B. R., Hilton, D., Mazurier, K., & Pouget, S. (2002). Psychological traits and tracking strategies. Working paper, Center for Studies of Economics and Finance, University of Salerno, April.

Cherubini, P., Castelvecchio, 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*, (2), 309-332.

Chronicle, E. P., MacGregor, J. N. & Ormerod, T. C. (2004). What makes an insight problem? The roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30,* 14-27.

Cloyd, C. B. & Spilker, B. C. (1999). The influence of client preferences on tax professionals' search for judicial precedents, subsequent judgments, and recommendations. *The Accounting Review*, July, 299-322.

Croskerry, P. (2002). Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Academic Emergency Medicine*, *9* (11), 1184-1204.

Crott, H. W., Giesel, M. & Hoffmann, C. (1998). The process of inductive inference in groups: The use of positive and negative hypothesis and target testing in sequential rule-discovery tasks. *Journal of Personality and social Psychology, 75* (4), 938-952.

Dunbar, K. (1993). Scientific reasoning strategies for concept discovery in a complex domain. *Cognitive Science, 17,* 397-434.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *The Nature of Insight* (pp. 365-395). Cambridge, MA (MIT Press).

Dunbar, K. (1999). How scientists build models: In vivo science as a window on the scientific mind. In L. Magnani, N. Nersessian & P. Thagard (Eds.), *Model-based Reasoning in Scientific Discovery* (pp. 89-98), (Plenum Press).

Farris, H. & Revlin, R. (1989a). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory & Cognition, 17,* 221-232.

Farris, H. & Revlin, R. (1989b). The discovery process: A counterfactual strategy. *Social Studies of Science, 19,* 497-513.

Fleiss, J. L., B. Levin, & M. C. Paik. (2004). *Statistical methods for rates and proportions* (third edition). New York (John Wiley and Sons).

Freedman, E. (1992a). Scientific induction: Individual versus group processes and multiple hypotheses. In J.K. Kruschke (Ed.), *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 183-188). Hillsdale, NJ (Lawrence Erlbaum).

Freedman, E. (1992b). The effects of possible error and multiple hypotheses on scientific induction. Paper presented at the 33rd Annual Meeting of the Psychonomic Society, St. Louis, MO.

Freedman, E. & Endicott, S.A. (1997). Individual differences in working memory and testing of single versus multiple hypotheses. Paper presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia, PA.

Freedman, E. & Jayaraman, R. (1993). Evaluating multiple hypotheses with a diagnostic strategy and maximally different hypotheses. Paper presented at the 34th Annual Meeting of the Psychonomic Society, Washington, D.C.

Gnepp, J. & Klayman, J. (1992). Recognition of uncertainty in emotional inferences: Reasoning about emotionally equivocal situations. *Developmental Psychology, 28,* 145-158.

Gorman, M. E. (1989). Error falsification and scientific inference: An experimental investigation. *Quarterly Journal of Experimental Psychology, 41A (2),* 385-412.

Hacker, K. L., Freedman, E. G., Gorman, M. E. & Isaacson, R. (1990). The emergence of task representations in small-group simulations of scientific reasoning. *Journal of Social Behavior and Personality, 5* (3), 175-186.

Hoffman, C. & Crott, H. W. (2004). Effects of amount of evidence and range of rule on the use of hypothesis and target tests by groups in rule-discovery tasks. *Thinking and Reasoning, 10* (4), 321-354.

Kareev, Y. & Avrahami, J. (1995). Teaching by example: The case of number series. *British Journal of Psychology, 86,* 41-54.

Kareev, Y. & Halberstadt, N. (1993). Evaluating negative tests and refutations in a rule discovery task. *Quarterly Journal of Experimental Psychology, 46A,* 715-727.

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., (2005). A framework for cognitive studies of science and technology. In Gorman, M., Tweney, R., Gooding, D. and Kincannon, M.E. (Eds.), *Scientific and Technological Thinking.* London (Routledge).

Klahr, D. & Dunbar, K. (1988). Dual space search during scientific experimentation: A developmental study. *Cognitive Psychology, 29,* 111-146.

Klahr, D. & Simon, H. A. (1999) Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 5*, 524-543.

Klayman, J. (1995). Varieties of confirmation bias. In J.R. Busemeyer, R. Hastie & D.L. Medin (Eds.), *Decision Making from the Perspective of Cognitive Psychology*. New York (Academic 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.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 461-469.

Kudadjie-Gyamfi, E. & Rachlin, H. (1996). Temporal patterning in choice among delayed options. *Organizational Behavior and Human Decision Processes, 65* (1), 61-67.

Laughlin, P. R. & Hollingshead, A. B. (1995). A theory of collective induction. *Organizational Behavior and Human Decision Processes, 61,* 94-107.

Laughlin, P. R., Magley, V. J. & Shupe, E. I. (1997). Positive and negative hypothesis testing by cooperative groups. *Organizational Behavior and Human Decision Processes, 69* (3), 265-275.

Lovett, M. (2002). Problem solving. In D.L. Medin (Ed.), *Stevens' Handbook of Experimental Psychology, Vol. 3*. New York (Wiley).

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology, 48,* 1-30.

Miwa, K. (2004). Collaborative discovery in a simple reasoning task. *Cognitive Systems Research,* 5, 41-62.

Mitroff, I. I. (1974). *The subjective side of science*. Amsterdam (Elsevier).

Muthard, E. K., & Wickens, C. D. (2003). Factors that mediate flight plan monitoring and errors in planning revision: Planning under automated and high workload conditions. 12th International Symposium on Aviation Psychology, Dayton OH.

Mynatt, C. R., Dougherty, M. E., & Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. *Quarterly Journal of Experimental Psychology Section A, 46*, 759-778.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608-631.

Okada, T. & Simon, H.A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science, 21* (2), 109-146.

Penner, D. E. & Klahr, D. (1996). When to trust the data: Further investigations of system error in a scientific task. *Memory & Cognition, 24* (5), 655-668.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347-353.

Poletiek, F. H. (2001). *Hypothesis-testing behavior*. Philadelphia (Psychology Press).

Popper, K. R. (1959). *The logic of scientific discovery*. London (Hutchinson).

Russo, J. E. & Schoemaker, P. J. H. (2002). *Winning decisions: Getting it right the first time.* New York (Doubleday).

Sanbonmatsu, D. M., Posovac, S. S., Kardes, F. R. & Mantel, S. P. (1998). Selective hypothesis testing. *Psychonomic Bulletin & Review, 5* (2), 197-220.

Schunn, C. D. & Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning. *Memory and Cognition, 24* (3), 271-284.

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

Tweney, R. D. (1990). Five questions for computationalists. In J. Shrager and P. Langley (Eds.). *Computational Models of Scientific Discovery and Theory Formation.* San Mateo CA (Morgan Kaufmann).

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.

Vollmeyer, R., Burns, B. D. & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75-100.

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

Wason, P. C. (1962). Reply to Wetherick. *Quarterly Journal of Experimental Psychology, 14*, 250

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

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

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.

Table 1

Frequency of Hypothesis Transitions by Phase

Phase	Type of Transition	Individuals[a]	Block Size			Group Size	
			2	3	5	Small	Large
1	Shift	47	33	70	99	9	7
	Refinement	118	112	82	105	38	33
	Retention	129	82	73	44	8	5
	Exploration	6	3	7	19	0	0
	Total	300	230	232	267	55	45
2	Shift	5	7	18	5	0	0
	Refinement	65	56	82	45	45	62
	Retention	21	9	15	23	17	12
	Exploration	0	1	1	0	0	0
	Total	91	73	116	73	62	74

[a] Samples sizes were: 49 individuals; 34, 34, and 26 for block sizes 2, 3, and 5, respectively; and 11 small groups and 10 large groups.

Figure 1

Path of Discovery—Standard 2-4-6 Task

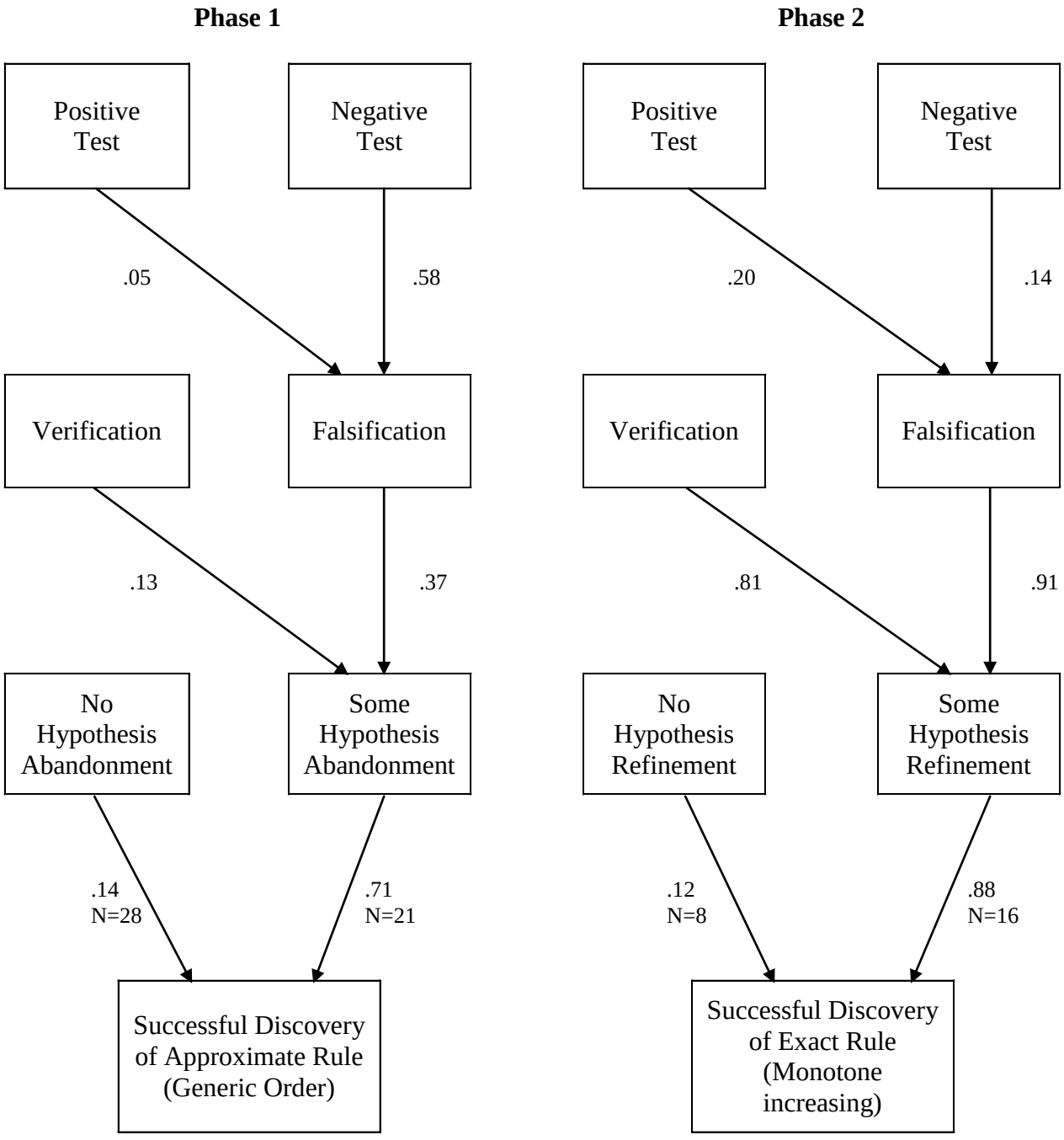

Figure 2

Path of Discovery—Delayed Feedback

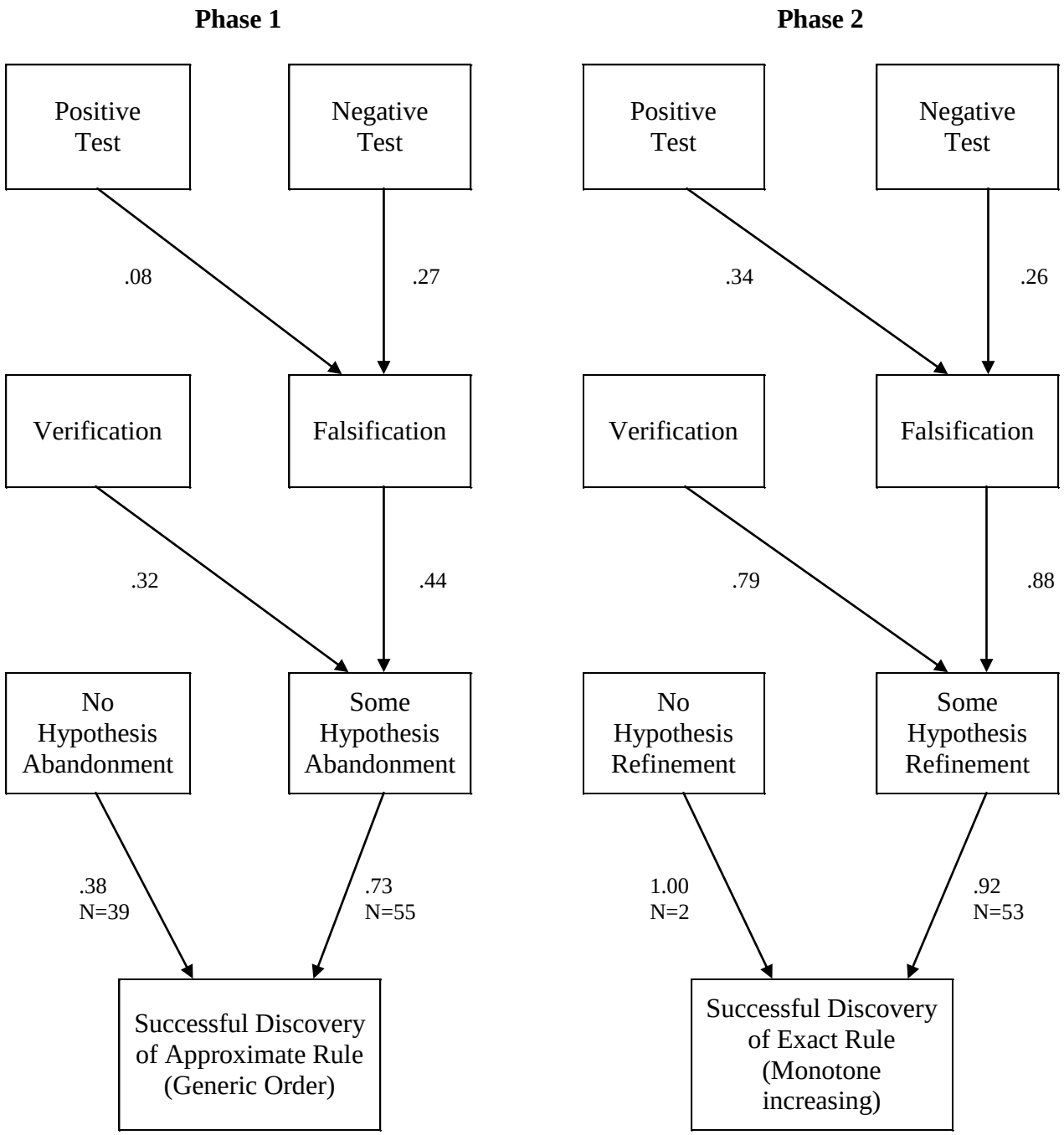

Figure 3

Path of Discovery—Groups