Memory & Cognition
1996, 24 (5), 655-668

When to trust the data: Further investigations of system error in a scientific reasoning task

DAVID E. PENNER
University of Wisconsin, Madison, Wisconsin

and

DAVID KLAHR
Carnegie Mellon University, Pittsburgh, Pennsylvania

When evaluating experimental evidence, how do people deal with the possibility that some of the feedback is erroneous? The potential for error means that evidence evaluation must include decisions about when to "trust the data." In this paper we present two studies that focus on subjects' responses to erroneous feedback in a hypothesis testing situation—a variant of Wason's (1960) 2–4–6 rule discovery task in which some feedback was subject to *system error*: "hits" were reported as "misses" and vice versa. Our results show that, in contrast to previous research, people are equally adept at identifying false negatives and false positives; further, successful subjects were less likely to use a positive test strategy (Klayman & Ha, 1987) than were unsuccessful subjects. Finally, although others have found that generating possible hypotheses prior to experimentation increases success and task efficiency, such a manipulation did little to mitigate the effects of system error.

Understanding our world often takes the form of hypothesis testing. We formulate hypotheses about why the car will not start and about the origins of the universe. We test these hypotheses with data from observations and experiments. However, like the Hubble Space Telescope, the observational instruments may be flawed, or the experimental equipment may be faulty. Indeed, one of the principle difficulties we face in both everyday and scientific reasoning is that of dealing with data that are not necessarily veridical.

Two types of errors can degrade evidence: *measurement error* and *system error*. Measurement error is usually characterized as random noise added to a continuous variable. In contrast, system error is characterized as categorical error in which a signal is changed so that it indicates a different category than the one from which it actually came. For example, in a nuclear reactor, measurement error in a thermocouple reading would result in a distribution of temperature readings around the true temperature, whereas system error would result in a signal light indicating that a valve was open when it was, in fact, closed.

Although the study of measurement error has long been a source of interest to researchers (e.g., Brehmer, 1979, 1980, 1987; Castellan, 1977; Einhorn & Hogarth, 1981; Klayman, 1984, 1988; Slovic, Fischhoff, & Lichtenstein, 1971; York, Doherty, & Kamouri, 1987), the effects of system error on reasoning processes have only relatively recently attracted attention (e.g., Doherty & Tweney, 1988; Gorman, 1986, 1989; Kern, 1982). The possibility of system error suggests that the evidence interpretation process includes a phase in which the veridicality of the evidence must be evaluated. Usually, such decisions are implicit. However, there are important circumstances when one must make an explicit determination about whether or not system error has occurred. Moreover, this determination may interact with the consequences of accepting or rejecting the evidence. In this paper, we report on two experiments in which we investigated the cognitive strategies that people use to deal with the possibility of system error. We employed a laboratory task that has been widely used in investigations of the psychology of scientific reasoning: the Wason (1960) rule discovery task.

In the Wason task, subjects are presented with a number triple, [2–4–6], and are told that it is a positive instance of a general rule which they are to discover. Subjects generate a hypothesis and an "experiment" (a number triple) to test their hypothesis. After each experiment, subjects are told whether or not their *triple* conforms to the rule. Subjects continue in this manner until they are sure they know the rule. The usual measures of interest are the number, type, and pattern of subjects' experiments and hypotheses.

One consistent finding from studies using the Wason task is that people overwhelmingly prefer to use what Klayman and Ha (1987) called a *positive test strategy*. Subjects using this strategy generate instances that are posi-

This research was originally conducted as part of the first author's doctoral dissertation. Support was provided in part by an American Psychological Association Dissertation Research Award to the first author, and a grant to the second author from the National Institute of Child Health and Human Development (R01-HD25211). We thank the reviewers for comments on an earlier draft. Correspondence concerning this article should be addressed to D. E. Penner at WCER, University of Wisconsin, 1025 W. Johnson St., Madison, WI 53706 (e-mail: dpenner@macc.wisc.edu).

tive exemplars of the currently hypothesized rule (+Htests) rather than instances that are negative exemplars of the hypothesized rule (−Htests). For example, if one's current hypothesis is "even numbers," then [4–10–8] would be a +Htest, while [3–4–11] would be a −Htest.

Nearly all previous investigations of the Wason task (e.g., Gorman, 1986, 1989; Wason, 1960) concluded that subjects approach this rule discovery task with a strong "confirmation bias": a desire to select instances that confirm rather than disconfirm the current hypothesis. The conclusion is based on the consistent tendency for subjects to choose triples that are positive instances of the current hypothesis. However, as Klayman and Ha's (1987) analysis clearly demonstrated, there is no logical basis for interpreting +Htests as attempts to confirm, nor −Htests as attempts to disconfirm. Depending on the relation between the hypothesized rule and the true rule, both +Htests and −Htests can provide either *conclusive falsification* or *ambiguous verification* of the current hypothesis. Conclusive falsification occurs when a +Htest receives "no" feedback, or a −Htest receives "yes" feedback. Ambiguous verification occurs when a +Htest receives "yes" feedback, or a −Htest receives "no" feedback. Thus, it is impossible to determine whether subjects are attempting to confirm or disconfirm simply by noting whether or not their triples are instances of the current hypothesis.

Klayman and Ha (1987) argued that a major problem in testing hypotheses is deciding whether, on average, conducting +Htests or −Htests will be most informative. They suggested that people tend to use a simple approach: Select the strategy that is likely to have the greatest impact on your belief in the current hypothesis. Moreover, Klayman and Ha argued that in the majority of real-world situations a +Htest strategy is just such a strategy. That is, even in nondeterministic environments (which could include system error) a +H strategy remains appropriate, since, in such situations

> falsifications are not conclusive but merely constitute some evidence against the hypothesis, and verifications must also be considered informative, despite their logical ambiguity. Ultimately, it can never be known with certainty that any given hypothesis is or is not the best possible. One can only form a belief about the probability that a given hypothesis is correct, in light of the collected evidence. (p. 219)

Consequently, in the real world, where error is always possible, scientists must decide not only how a particular datum bears on their hypotheses, but also how their hypotheses, plus all the accumulated evidence, bear on the reliability of the datum. That is, they must decide when to "trust the data."

STUDIES OF SYSTEM ERROR

The experiments reported in this paper extend earlier work on the effects of system error during scientific reasoning tasks (e.g., Doherty & Tweney, 1988; Gorman, 1986, 1989, 1992; Kern, 1982; O'Connor, Doherty, & Tweney, 1989; Tweney et al., 1980). To place our work in context, we will briefly summarize the principal results from this body of literature.[1]

Kern (1982) investigated system error by using a microworld task in which subjects were told that errors would occur on approximately 25% of the trials. Kern found that subjects in the error condition were more likely to challenge the validity of the feedback following falsification than following verification. This lead Kern to conclude that subjects were biased to believe that falsification trials were system errors, and that they used this bias to justify retaining their hypotheses. That is, they used the possibility of error to *immunize*, or preserve, their current hypothesis.

Gorman (1986) also explored the effects of system error, particularly how such errors affect efforts to falsify hypotheses. He presented three groups of subjects with a variant of the Wason task and instructed each group to use a different experimental strategy, one of which was to attempt to falsify, rather than verify, the active hypothesis. In addition, all subjects were told that between 0% and 20% of the feedback that they received would be erroneous. In fact, all feedback was veridical. Gorman's results showed that strategy instructions affected neither task success nor the number of trials considered to be errors. However, subjects were more likely to consider falsification, rather than verification, trials to be errors, even though they never received false feedback. These results led Gorman to concur with Kern's (1982) conclusion about hypothesis preservation; people use the possibility of system error to label +Htest falsification trials as errors, and thus retain their current hypothesis.

In a later study, Gorman (1989) replicated his earlier work, but with the addition of an *actual error* condition. He found that actual error subjects often retained a hypothesis following falsification, labeling such trials as errors. However, in both this and his earlier study, Gorman used the traditional Wason rule "Ascending numbers." Klayman and Ha (1987) point out that this rule is almost always more general than subjects' initial hypotheses. Thus, since most people prefer a +Htest strategy, their tests tend to be positive instances of the true rule. Consequently, veridical falsification is rare, and error trials are overwhelmingly false negatives.

Gorman (1989) did not report the distribution of verification and falsification trials. However, given his use of Wason's (1960) traditional rule, it is likely that most, if not all, errors were in the form of false negatives (i.e., reporting that a +Htest was not an exemplar of the rule when, in fact, it was). This suggests at least one alternative explanation for Kern's (1982) and Gorman's conclusion that subjects have an immunization bias. Subjects might simply note the low frequency of falsified trials and interpret them as similarly infrequent system errors. That is, both immunization bias and a pattern-matching heuristic would lead to +Htest falsification trials being marked as errors. Therefore, it is possible that Gorman's results reflect a methodological artifact, rather than a bias toward immunization.

In summary, although the inclusion of system error into the classic Wason (1960) task represents an important step toward increasing the face validity of laboratory explorations of the scientific discovery process, the procedures used thus far confound rare errors with rare +Htest falsification. In the studies described below, we disentangled these two factors; further, we investigated additional questions about the role of a +Htest strategy in cases where system errors occur. Specifically, we addressed the following questions:

1. Are people really biased to label falsification trials as errors, or is this conclusion based on the confounding of rare +Htest falsification feedback with rare error trials? One way to determine the existence of an immunization bias is to contrast performance when the rule to be discovered is very broad (as in the typical Wason task) with performance when the rule is structured so as to generate a balanced distribution of verifications and falsifications. Gorman (1986, 1989) suggested that subjects' immunization bias will preclude them from suspecting that false positives (i.e., "yes" feedback when in fact the instance does not match the rule) are errors, since such feedback provides support for the current hypothesis. However, if subjects are equally able to identify false positives and false negatives, the existence of an immunization bias would be called into question.

2. Is the effectiveness of a +Htest strategy maintained in the context of system error, as Klayman and Ha (1987) have argued?

3. How does system error interact with the unique qualities of different content domains? Klayman and Ha (1989) found no differences in rule discovery performance in a study in which they used both a numerical domain and a geography domain. However, this work was based on tasks without system error. Since different content areas support different types of relationships between members, it is possible that people will not find it equally easy to identify errors in different domains. For example, the seed [2–4–6] encompasses mathematical relationships such as "even numbers." In contrast, the seed [mouse–cat–elephant] entails a set of relationships, such as "alive," that do not exist in the [2–4–6] domain. The addition of system error may interact with the unique qualities of each domain to differentially impair success.

EXPERIMENT 1

Experiment 1 incorporated system error into the basic paradigm of the Wason rule discovery task. In addition we included a second content domain, along with a second rule that has been shown to provide subjects with a more even distribution of verification and falsification than does the traditional form of the Wason rule.

Method

Subjects

One hundred and twenty psychology undergraduates participated in this study for partial course credit. The subjects were run either singly or in groups of 2 or 3.

Design

A domain (numerical/animal) × rule (broad/narrow) × error (presence/absence) between-subjects design was used, with 15 subjects in each of the eight cells. Specific rules (broad or narrow) were nested within domain. In the numerical domain, the broad rule was "Numbers in ascending order," and the narrow rule was "Sequential, even numbers between 2 and 100, inclusive." For both rules, the initial instance was [2–4–6]. In the animal domain, we used rules developed by Farris (1992): the broad rule was "Living things," and the narrow rule was "Mammals in increasing order of size." The initial instance for both was [mouse–cow–elephant]. In both the numerical and the animal domains, the broad rule is likely to include subjects' initial hypotheses. We expected that the preference for +Htests would lead broad-rule subjects to receive predominantly verification of their test triples, while narrow-rule subjects would receive a more even distribution of verification and falsification (cf. Klayman & Ha, 1989).

Procedure

The subjects were presented with a pencil-and-paper task and were told that we were interested in studying scientific reasoning. At the top of the paper was the initial instance for their condition. Below the initial instance were columns labeled "Hypothesis," and "Experiment," and two response columns: "Conforms" and "Does Not Conform." The subjects were told that the initial instance was a positive example of a rule that they were to discover (see the Appendix for the full text of the instructions). They were told that they were to generate and write down a hypothesis and a test of that hypothesis, such as [6–8–10] or [cat–dog–horse]. It was emphasized that they must state their current hypothesis on each trial, though they were free to repeat hypotheses and tests throughout the study. After each test, the experimenter indicated whether or not the *test* conformed to the rule for the assigned condition, by placing a check mark in the appropriate response column. The subjects could terminate the study at any time by writing out their proposed rule. The subjects' record sheets were available to them throughout the study.

In addition to receiving the basic instructions, the subjects in the error conditions were told that in order to simulate real-world science, there might be some "noise" or random error in the feedback they received. That is, if their experiment conformed to the rule, they might be told that it did not conform, and vice versa. The subjects were told that on 0%–20% of their trials they would receive false feedback (cf. Gorman, 1989). It was emphasized that there might be no errors, but that, if there were, the errors would occur on no more than one in five trials, on the average. In fact, all error subjects received errors on the same 20% of their trials (e.g., Trials 5, 7, 12, etc.), determined in advance by using a random number table to select 5 trials out of 25. However, while the trials on which the subjects received error feedback were determined in advance, the *type* of error (i.e., false positive or false negative) depended on the experiments that the subjects conducted on each of the error trials.

Although a 20% error rate may seem so high as to reduce the face validity of the task as a laboratory analogue of real-world science, there are several arguments for using such a rate. First, it is the error rate used in earlier studies, and thus it provides continuity with previous research (e.g., Gorman, 1986, 1989). Second, we wanted to affect subjects' reasoning processes without making the task impossible to solve. Using very low error rates would require a task with many more trials, in order to have sufficient instances of error to get an effect. Third, the rate is not so high as to hopelessly confuse subjects: Gorman's (1989) research showed that some subjects could generate the correct rule even when 20% of the feedback was erroneous. Finally, a 20% system error rate is not unheard of in real-world situations, such as medical diagnosis.[2]

Subjects were instructed to indicate suspect trials by placing an "X" next to them. Subjects could change their mind about the status of a suspected trial by placing a "√" next to any "X." At the conclusion of the study, the subjects were asked to add any final "Xs"

or "√s" they thought necessary. This allowed us to determine both the types of trials subjects associated with error, and whether or not they correctly identified the errors they received.

Results

Rule Discovery

We used both a strict and a lenient criterion for determining whether or not subjects discovered the correct rule. For the strict criterion, subjects were scored as successful only if they stated the complete rule for their condition. For the lenient criterion, subjects had to discover the core of the rule (e.g., sequential, even numbers), but not the boundary conditions (e.g., between 2 and 100, inclusive). In the following analyses, *success* refers to the strict criterion; *near success*, to the lenient criterion.

The mean success rate for no-error subjects was 48%, whereas for error subjects it was 17% [$\chi^2(1, N = 120) = 13.71, p < .001$]. The no-error success rates (see Figure 1) were generally similar to those reported by previous researchers (e.g., Farris, 1992; Freedman, 1992a; Gorman, 1989; Klayman & Ha, 1987). A chi-square analysis revealed a significant difference between conditions [$\chi^2(7, N = 120) = 26.4, p < .001$]. Inspection of post hoc cell contributions revealed that in the numerical domain more broad-rule no-error subjects and fewer narrow-rule error subjects were successful than expected by chance; within the animal domain, fewer narrow-rule error subjects were successful than expected by chance ($p < .05$ in all cases).

Number of Trials

As Table 1 shows, error subjects, on the average, generated more than twice as many trials as did their no-error counterparts [$F(1,112) = 72.3, p < .001$]. These results partially replicate those of previous investigations (Kern, 1982; Gorman, 1989) with respect to the overall effect of error on number of trials. However, the effect comes from the numerical domain, for there is a significant interaction between error and rule nested within domain [$F(3,112) = 4.38, p < .01$]. Scheffé post hoc analyses were conducted to compare error and no-error conditions for the same rule within each domain, yielding four comparisons. Within the numerical domain, error subjects conducted significantly more trials than did no-error subjects for both broad and narrow rule comparisons ($p < .05$). In the animal domain, although there were more trials by error subjects than by no-error subjects, the difference was nonsignificant. Because subjects differed in the number of trials that they conducted, the following analyses, except where noted, are based on proportions and not absolute values.

Hypothesis Categories

We suggested above that different domains support categorically different types of hypotheses. In order to investigate this premise, we categorized subjects' hypotheses. All hypotheses in the numerical domain fell into one category: arithmetical principles (e.g., even numbers, numbers in increasing order, $n + 2$, etc.). Hypotheses in the animal domain belonged to five unique categories: 41% were taxonomic (e.g., mammals); 17% were idiosyncratic (e.g., number of letters, must have thick skin); 13% were based on relative size (e.g., small to large); 11% were based on appearance (e.g., all have legs); 3% referred to location (e.g., found on land); 2% referred to function (e.g., they all walk). Some hypotheses involved combinations of two categories: 9% combined taxonomic and size categories (e.g., mammals in order of increasing size); 4% combined taxonomic and appearance categories (e.g., mammals with four legs). Thus, although subjects were free to generate any type of hypothesis they wished, all subjects in the numerical do-

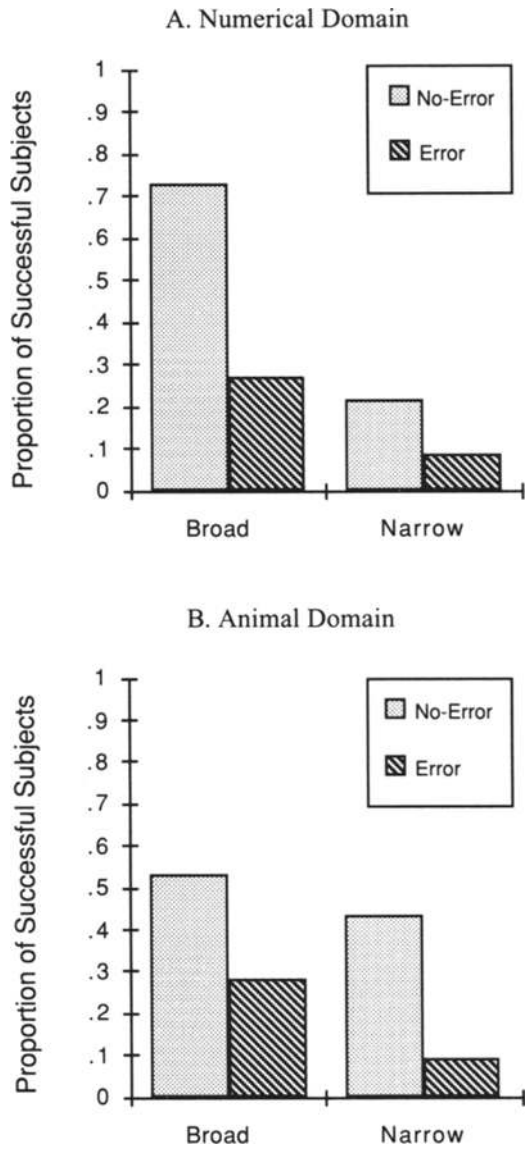

Figure 1. Proportion of subjects discovering (strict success) the rule for each rule type and error condition. (A) numerical domain (broad rule, numbers in ascending order; narrow rule, sequential, even numbers between 2 and 100, inclusive). (B) animal domain (broad rule, living things; narrow rule, mammals in increasing order of size).

Table 1
Mean Number (and Standard Deviation) of Trials

	Numerical				Animal					
	Broad Rule		Narrow Rule		Broad Rule		Narrow Rule		Combined	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
No error	9.7	5.8	12.3	9.3	8.7	3.2	11.9	5.3	10.6	6.3
Error	27.6	10.9	26.5	8.3	15.8	5.4	18.3	7.5	22.1	9.6
M	18.6	12.5	19.4	11.3	12.3	5.7	15.1	7.2		

main appear to have assumed that the rule must be arithmetically based. In contrast, subjects in the animal domain had few a priori constraints on the types of hypotheses that might prove useful.

+Htest Strategy

One of the motivations for this study was to evaluate Klayman and Ha's (1987) suggestion that a +Htest strategy can be effective in cases where system error occurs. Test strategy was scored by comparing the current hypothesis with the test triple for that trial. For example, a trial with the hypothesis "even numbers" and the triple [8–10–12] would be scored as a +Htest. If the triple had been [7–9–11], the trial would be scored as a −Htest.

We found that the already high base rate of +Htests in no-error conditions was not increased in error conditions: In both conditions, more than 80% of the experiments were +Htests (see Table 2). Across conditions, between 8 and 14 subjects generated +Htests on at least 75% of their trials. There was a main effect only for rule nested within domain [$F(3,112) = 3.58, p < .05$]. In the numerical domain, rule type did affect the proportion of +Htests: 76% of the broad-rule subjects' experiments were +Htests; this proportion was 91% for the narrow-rule subjects ($p < .05$). There was no effect of rule for the animal domain.

Experiment Feedback

The observation that subjects predominantly use a +Htest strategy does not provide any information on the type of feedback that they received. Gorman (1986, 1989) argued that the proportion of falsification feedback received is positively correlated with success. However, Klayman and Ha (1987) have claimed that verification can be as useful as falsification: +Htest verification highlights hypotheses for further investigation. Since feedback is independent of test type (Klayman & Ha, 1987), we analyzed +Htests and −Htests separately, using subjects' *perceived* data sets.[3]

+Htest feedback. Virtually all subjects received verification on more than 50% of their +Htests, with broad-rule no-error subjects in both domains receiving almost all of their feedback in this form. As noted earlier, the narrow-rule condition was designed to increase the frequency with which subjects would generate triples that were not instances of the correct rule. Table 3 lists the ratio of +Htest verification to falsification for broad-rule conditions and the ratio of falsification to verification for narrow-rule conditions.

The manipulation worked: broad-rule subjects received from 1.5 to 7.3 times as much verification as falsification. For narrow-rule subjects, the ratio of verification to falsification feedback was reversed and attenuated: narrow-rule subjects received from 1.1 to 1.9 times as much falsification as verification to their +Htests. Analysis revealed a main effect for rule nested within domain [$F(3,112) = 43.44, p < .001$]. Scheffé post hoc comparisons revealed that, in both domains, broad-rule subjects had proportionally more +Htest verification (71% and 76%, for numerical and animal domains, respectively) than did narrow-rule subjects (41% and 39%, for numerical and animal domains, respectively) ($p < .05$ for all comparisons). There was also an effect of error: 65% of no-error subjects' +Htests were verified, whereas 49% of error subjects' +Htests were verified [$F(1,112) = 33.25, p < .001$].

−Htest feedback. −Htests were much less frequent than were +Htests. Moreover, there was considerable disparity in the type of feedback −Htests received: between 50% and 100% of the narrow-rule subjects received only −Htest verification; however, virtually all broad-rule subjects received falsification on at least one −Htest.

Analysis revealed a main effect only for rule nested within domain [$F(3,63) = 17.61, p < .001$]. In the numerical domain, 36% of broad-rule subjects' −Htests were verified, whereas 78% of narrow-rule subjects' −Htests

Table 2
Mean Proportion (and Standard Deviation) of +Htests

	Numerical				Animal					
	Broad Rule		Narrow Rule		Broad Rule		Narrow Rule		Combined	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
No error	.81	.24	.86	.13	.86	.22	.90	.16	.86	.19
Error	.72	.21	.96	.07	.75	.24	.85	.18	.82	.21
M	.76	.22	.91	.12	.80	.23	.83	.17		

Table 3
Ratio of Mean Proportion of Verified +Htests to Falsified +Htests (Bold) and Falsified +Htests to Verified +Htests

	Numerical		Animal	
Condition	Broad Rule	Narrow Rule	Broad Rule	Narrow Rule
No error	**4.9**	1.1	**7.3**	1.2
Error	**1.5**	1.8	**1.8**	1.9

were verified; the results were similar in the animal domain (25% and 91% for the two rule conditions). Scheffé post hoc comparisons within both domains were significant ($p < .05$ for both comparisons).

Hypothesis Change

An immunization bias predicts that subjects in error conditions should be less likely to change their hypothesis after receiving falsification than no-error subjects, since they can treat such feedback as a system error. Overall, from 13 to 15 subjects in each condition changed a hypothesis at least once following falsification. Error subjects changed their hypothesis following falsification on 60% of such trials, whereas the no-error subjects did so on 73% [$F(1,107) = 7.06, p < .01$].

Response to Error Trials

If subjects have an immunization bias, they should be more likely to attribute errors to falsified rather than verified trials. However, as noted above, previous studies of system error utilized a rule that may have lead to the majority of error trials occurring as false negatives (i.e., "no" feedback); our inclusion of a narrow rule was designed to counter this potential problem.

Collapsing over rule and domain, subjects marked 44% of their verification trials and 56% of their falsification trials as errors. The difference was nonsignificant. However, given the qualitative differences between narrow and broad rule, separate analyses were conducted to see the effect of rule type on error ascription.

Broad-rule subjects labeled significantly more falsification trials, 64%, than verification trials, 36%, as errors [$t(30) = 2.49, p < .05$]. This result is consistent with Gorman's (1986, 1989) results. In contrast, narrow-rule subjects were almost evenly split in labeling verification, 52%, and falsification, 48%, trials as errors.

Virtually all error subjects identified at least one of the false negative errors they received. All numerical, but only 50% of the animal, subjects identified at least one false positive. Overall, subjects did not differ in the proportion of false negatives or false positives correctly identified (see Table 4).

Experiment Replication

Replicating an experiment—that is, using a test triple identical to one used earlier—can be a useful strategy for identifying error. If the replication feedback is discrepant from the original feedback, one of the two must have been an error trial. In contrast, replication is of no benefit in the absence of errors. Consequently, it is not surprising that only 2 no-error subjects ever replicated an experiment. Error subjects varied widely in their use of a replication strategy. Fifty percent of the error subjects never replicated a trial; the remaining subjects replicated between 1 and 15 trials. Overall, error subjects replicated an average of 8% of their trials. There was no effect of rule nested within domain on the proportion of trials replicated.

A replication strategy is most useful when subjects correctly identify error trials as the ones to replicate. Of those subjects replicating a trial, approximately 50%–75% replicated at least one error trial.

Overall, 48% of the subjects' replications involved trials on which they had received false feedback. Testing the proportion of error trial replications against the actual error rate of 20% showed that the subjects who did replicate were more likely to replicate an error trial than would be expected by chance alone [$t(29) = 3.73, p < .001$].

Task Success

Strict success. As discussed above, strict success required subjects to state the complete rule for their condition. The following reports the dependent measures with respect to this criterion.

As stated above, error subjects changed their current hypothesis following falsification less often than did no-error subjects. However, there was no association between rule discovery and this measure: both *successful* and *unsuccessful* subjects changed their hypothesis following falsification on approximately 70% of such trials.

In order to investigate the efficacy of a +Htest strategy, we analyzed the proportion of +Htests conducted with respect to success. Successful subjects had *lower* proportions of +Htests (76%) than did unsuccessful subjects (88%) [$F(1,118) = 10.36, p < .01$].

While successful subjects had lower proportions of +Htests, they received more verification (65%) than did unsuccessful subjects (53%) to such tests [$F(1,118) = 6.28$,

Table 4
Mean Proportion (and Standard Deviation) of Errors Correctly Identified by Condition

	Numerical						Animal							
	Broad Rule			Narrow Rule			Broad Rule			Narrow Rule			Combined	
Errors	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
False positives	.83	.27	9	.60	.38	15	.50	.55	6	.44	.43	15	.58	.41
False negatives	.62	.31	15	.64	.42	14	.68	.42	15	.60	.46	10	.64	.39

$p < .05$]. Successful and unsuccessful subjects received verification on approximately half of their −Htests.

Klayman and Ha (1987) argued for the importance of conclusive falsification, and show how it can result from either +Htests or −Htests. In the present study, conclusive falsification played an important role in task success. Although successful and unsuccessful subjects received about the same proportions of conclusive falsification (39% and 46%, respectively; see Table 5), the two groups differed in the *source* of that feedback. Of the conclusive falsifications received by successful subjects, 30% were in the form of falsified −Htests, while for unsuccessfuls, only 11% of the conclusive falsification came from falsified −Htests. This pattern was generally maintained for individual rule × domain × error conditions: in six of eight conditions, −Htest falsifications were between 2 and 10 times more frequent for successful than for unsuccessful subjects.

Task success was also associated with the correct identification of false negative errors; successful subjects identified 100% of their false negative errors and 78% of their false positives. In contrast, unsuccessful subjects identified only 56% of their false negatives and 53% of their false positives. The difference between successful and unsuccessful subjects was significant for only false negatives [$F(1,52) = 11.15, p < .01$]. Thus, although subjects exhibited no bias toward ignoring false positives, only the correct identification of false negatives was associated with successful rule discovery.

Although replication potentially provides a powerful tool for identifying error trials, our results show that successful subjects replicated approximately as often as did unsuccessful subjects (12% and 7%, respectively). Moreover, there was no difference in proportion of error trial replications by successful (57%) and unsuccessful (46%) subjects.

Near success. Traditionally, success on the Wason task has been defined as a complete statement of the rule. This perspective was used in the analyses above. However, the narrow rule can be considered as a two-part rule: a core (e.g., sequential even numbers), plus a range condition (e.g., between 2 and 100, inclusive). Thus failure can occur in two qualitatively distinct manners. For example, a subject may have no idea about either the core or the range condition, and consequently be classified as unsuccessful. However, under the traditional paradigm, a subject would also be classified as incorrect even if he/she discovered the core rule (e.g., sequential even numbers), but missed the range condition (e.g., between 2 and 100, inclusive).

Table 5
Proportion of Ambiguous Verification (AV) and Conclusive Falsification (CF) for Successful and Unsuccessful Subjects

		Feedback	
Outcome	Test Type	"Yes"	"No"
Successful	+H	49% AV	27% CF
	−H	12% CF	12% AV
Unsuccessful	+H	47% AV	41% CF
	−H	5% CF	7% AV

The qualitative difference between the two forms of incorrectness suggests that it might be worthwhile to look at the dependent measures with respect to a more liberal measure of success. Consequently, we scored narrow-rule subjects as near-successful if they stated, at a minimum, the core condition (see Figure 2). The following analysis summarizes the most important dependent measures with respect to near success. Collapsing the results over rule and domain revealed that the near-success rate for no-error subjects was 71%, while for error subjects it was 28% [$\chi^2(1, N = 120) = 19.22, p < .0001$]. A chi-square analysis showed a significant difference between conditions for rule discovery [$\chi^2(7, N = 120) = 35.64, p < .0001$]. Inspection of the post hoc cell contributions revealed that within the numerical domain, more narrow-rule no-error subjects and fewer broad-rule error subjects were successful than expected by chance alone ($p < .05$ for both cases). Within the animal domain, both broad- and narrow-rule error subjects were less successful than expected ($p < .05$ for both comparisons).

A series of analyses of variance revealed no difference between near-successful and unsuccessful subjects with respect to proportion of new hypotheses generated following falsification, proportion of +Htests generated, proportion of verified +Htests and −Htests, proportion of trials replicated, or proportion of error trials replicated. Moreover, the near-success criterion did little to change the pattern of conclusive falsification found with the strict-success criterion.

Near-success subjects correctly identified 79% of their false positives and 88% of their false negative errors. Unsuccessful subjects correctly identified 46% of their false positives and 52% of their false negatives. The difference between near-successful and unsuccessful subjects was significant for the identification of false positives [$F(1,52) = 7.66, p < .01$] and false negatives [$F(1,52) = 11.84, p < .01$].

Discussion

The results of Experiment 1 show that in the classic numerical domain, successful rule discovery on the Wason task is affected by system error and rule type. However, in the animal domain, there was an effect only for system error.

The inclusion of the animal domain was motivated by a desire to see how system error interacted with the unique aspects of qualitatively different content domains. Theoretically, any number of possible rules could be constructed to describe either domain. However, while subjects generated different hypotheses for the two domains, they also differed in the number of categories from which they drew their hypotheses. The subjects' record sheets reflect these differences: within the numerical domain, subjects focused exclusively on arithmetical relationships (e.g., even numbers, even numbers increasing by two, etc.); in contrast, animal domain hypotheses were

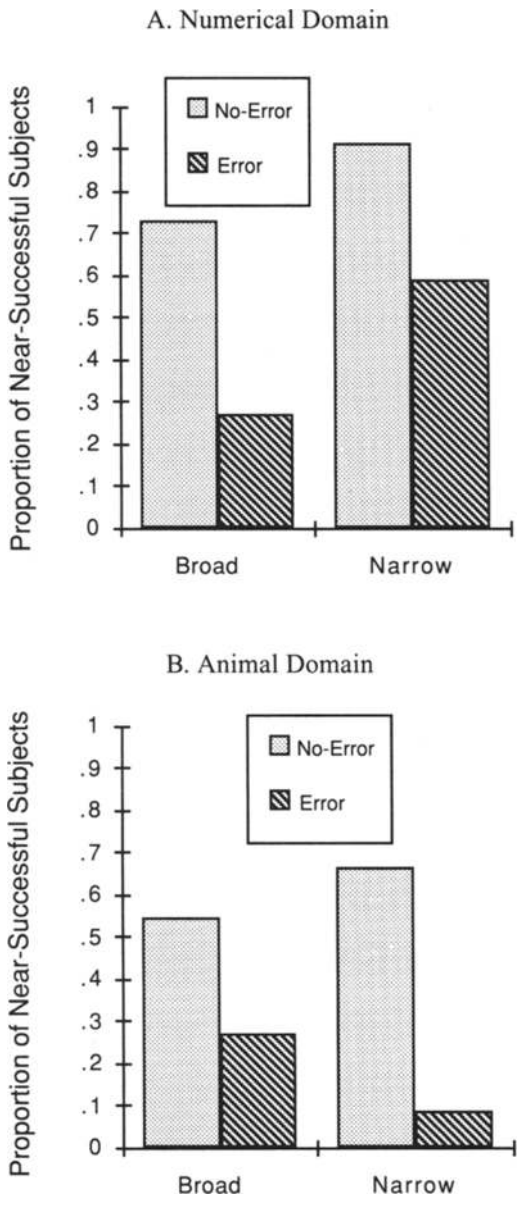

Figure 2. Proportion of subjects discovering (near success) the rule for each rule type and error condition. (A) Numerical domain (broad rule, numbers in ascending order; narrow rule, sequential, even numbers between 2 and 100, inclusive). (B) Animal domain (broad rule, living things; narrow rule, mammals in increasing order of size).

drawn from a wide range of qualitatively different categories (e.g., size, appearance, taxonomy, etc.). Thus, the two domains differ in the number, and types, of plausible hypotheses they support.

These results suggest that the effects of rule on task success are mediated by the specific properties of the domain: The inclusion of the narrow-rule condition lowered the proportion of successful subjects only in the numerical domain. A detailed examination of subjects' proposed rules revealed that in the numerical domain, narrow-rule subjects often failed to include the boundary conditions as part of their final rules. In contrast, within the animal domain, many subjects proposed a version of the narrow rule in both the narrow- and broad-rule conditions. This interaction of rule type and domain raises questions about analyzing the dependent measures with respect to task success. This issue will be addressed below.

While the inclusion of system error does affect the amount of time people spend in trying to find the correct rule, there is little effect on people's test strategies. All subjects relied primarily on a +Htest strategy. In contrast, however, to Klayman and Ha's (1987) argument for the efficacy of a +Htest strategy when there is a possibility of error, we found that the successful subjects in both error and no-error conditions were those who generated fewer +Htests. But, in order to fully understand the relationship between a +Htest strategy and task success, the type of feedback subjects received must also be considered.

Klayman and Ha (1987) argued that when the hypothesis space is large, and errors are possible, high frequencies of +Htest verification provide support for the current hypothesis and indicate which region of the hypothesis space to further explore. In contrast, Gorman (1986, 1989) suggested that higher proportions of falsification are associated with success. These positions are not mutually exclusive, and indeed, Experiment 1 supported both: higher proportions of +Htest verification and higher proportions of −Htest falsification were associated with task success.

One of the major difficulties facing error subjects is that of trying to isolate system error trials. One strategy is to replicate suspect trials. As our results show, people varied widely in their decision to replicate. However, when people did replicate, they replicated error trials more often than would be expected by chance.

As discussed above, many subjects were able to identify the core portion of the narrow rule, but failed to discover its boundary conditions. This suggests that the conventionally used strict-success criterion may underestimate the progress people make toward rule discovery. Subsequent analyses using near success revealed two main differences from the results found using strict success. The original analyses showed that successful subjects generated fewer +Htests, and received less +Htest falsification; the near-success analyses revealed no differences on these measures. Thus, over-reliance on a +Htest strategy appears to hinder discovery, not of the core rule, but of the boundary conditions.

The use of the near-success criterion does raise some issues about determining success. First, the liberal criterion only affects the narrow-rule conditions. Since the broad rules only have a single proposition, there is no possibility for partial success. Second, the criterion has a differential affect on the two domains. Comparison of Figures 1 and 2 shows that the near-success criterion increased success for both narrow-rule conditions in the numerical domain, but only for the no-error condition in the animal domain. In the numerical domain, virtually

all narrow-rule subjects, regardless of error condition, generated the core portion of the rule. However, in the animal domain system, error severely affected people's ability to form the core portion of the rule; error subjects generated idiosyncratic rules that bore no resemblance to the true rule. The impact of system error on the types of hypotheses generated in different domains is an area that needs further investigation.

A major motivation for Experiment 1 was to investigate Kern's (1982) and Gorman's (1986, 1989) contention that people are biased to consider only falsification trials as possible errors. In order to investigate this claim, we needed to unconfound error trials and false negatives by making false positives about as likely as false negatives. The inclusion of the narrow rule led to a more balanced distribution of verification and falsification trials and to the presence of both false positives and false negatives. While there was a trend for people to prefer falsification trials as errors, this result needs to be considered with respect to rule type: broad- but not narrow-rule subjects labeled proportionally more falsification than verification trials as errors.

A more liberal interpretation of an immunization bias is that, given the possibility of system error, people are reluctant to give up a hypothesis following falsification. Inclusion of a narrow rule allowed us to investigate this possibility. Our results show that virtually all subjects changed their current hypothesis following disconfirmation at least once, although error subjects were less likely to do so than no-error subjects. Thus, our results do not support a strict interpretation of an immunization bias; they do, however, provide some support for a more liberal interpretation.

Obviously, there is a difference between labeling and correctly identifying a trial as an error. Success was associated with the identification of higher proportions of system error. Yet, while successful subjects correctly identified all of their false negatives, they did not identify all of their false positives even though they had sufficient information to do so. Why, then, did successful subjects not identify all of their false positives?

One explanation is that since subjects have already concluded that they know the rule, there is little incentive for them to spend time looking for errors that they might have missed. Consequently, they fail to identify some of the false positives that they have received. However, this explanation fails to account for the fact that successful subjects did identify all of their false negatives, some of which were identified at the conclusion of experimentation. This may reflect people's preference for reasoning from exemplars rather than nonexemplars of a concept (see, e.g., Bruner, Goodnow, & Austin, 1956). Identifying false negatives as errors allows one to recast nonexemplars as exemplars. This recasting may facilitate the induction of a viable hypothesis by increasing the number of positive exemplars.

Experiment 1 raised a number of issues for further exploration. Our procedure required subjects to have a hypothesis for each test. However, during the course of the study some subjects commented on their difficulty in generating a hypothesis to test. This may have had an unexpected effect: subjects may have occasionally listed hypotheses that they were not explicitly testing. As Klahr and Dunbar (1988) have shown, some subjects prefer to use experimentation as a means of generating, rather than evaluating, a hypothesis during experimentation. That is, they conduct experiments without explicit hypotheses, using the resulting feedback to eventually generate a hypothesis. Since we required explicit hypotheses for every trial, it is impossible to tell whether subjects were using such a strategy. One alternative to this procedure is to allow subjects to run experiments without explicit hypotheses.

The separate roles of hypothesis generation and evaluation raise a second point. Tweney et al. (1980) conducted a modified version of the Wason task which illustrates how competing hypotheses might help task success. In this study, subjects were told that the experimenter would classify each test as being one of two mutually exclusive rules. That is, all tests were positive exemplars of one rule or the other. The manipulation boosted the success rate to 85%; it was approximately 30% in the traditional Wason task.

However, this research has two drawbacks. First, competing hypotheses in real-world tasks are unlikely to be either mutually exclusive or exhaustive. Second, the manipulation does not really investigate how multiple hypotheses affect task performance; rather, it shows how receiving only confirmatory feedback affects the discovery of two mutually exclusive rules.

A number of researchers (e.g., Freedman, 1992b; Klahr & Dunbar, 1988; Tweney, Doherty, & Mynatt, 1981) have investigated the effects of generating alternative hypotheses prior to experimentation. This work suggests that people can generate a correct hypothesis without conducting any experiments; moreover, the manipulation improves the efficiency with which the correct hypothesis is subsequently discovered. However, the effects of system error on such a strategy have not been investigated in any of these studies.

In order to explore these issues, we designed a second experiment in order to investigate the following questions: (1) Does initially listing multiple hypotheses increase the proportion of error subjects who discover the rule, and does it improve the efficiency of their experimentation? (2) Do subjects choose to conduct experiments without stated hypotheses, and if so, is this an effective strategy for rule discovery, given the possibility of system error?

EXPERIMENT 2

Method

Subjects

The subjects were 59 college undergraduates. They were run separately, and they received course credit for participating in the study.

Design

A prior hypotheses (presence/absence) × error (presence/absence) between-subjects design was used. In order to provide a balance of "yes" and "no" feedback, all subjects were required to

discover the narrow rule, "Sequential, even, numbers between 2 and 100, inclusive." The seed trial was [2–4–6].

Procedure

The procedure was identical to that of Experiment 1, with the following differences. Prior to experimentation, subjects in the prior-hypotheses conditions were shown the set [2–4–6] and were instructed to list as many rules as they could that described this set of numbers. The subjects' hypotheses were available to them throughout the remainder of the study.

All subjects were told that some people found it useful to sometimes conduct experiments without hypotheses. Thus, if they chose to, subjects could conduct *nil* hypothesis trials.

Results

Prior Hypotheses

In order to ensure that the error and no-error subjects did not differ in the number and quality of their prior hypotheses, we compared them on both of these measures. While no-error subjects generated slightly fewer prior hypotheses than did error subjects (3.8 vs. 5.3), the difference was nonsignificant.

The quality of the prior hypotheses was judged by 10 psychology graduate students. Each proposed hypothesis was rated against the correct rule on a scale from 0 (*no match*) to 1 (*perfect match*). A Kendall coefficient of concordance was calculated in order to measure the degree of agreement between the judges. Although the judges did not completely agree in their rankings, they did agree more than would be expected by chance ($W = .46, p < .001$).

No subject generated the correct rule prior to experimentation. In order to estimate subjects' best guess as to the correct rule, we compared subjects' highest rated hypothesis. Mean ratings for these "best prior" hypotheses ranged from .33 to .78, with overall mean ratings of .51 for error subjects and .52 for no-error subjects.

Rule Discovery

In the no-error condition, 53% and 29% of the subjects discovered the rule in the no-prior and prior-hypotheses conditions, respectively, though the difference was nonsignificant. In the error condition, only 7% of the subjects in either hypothesis condition were successful. A chi-square analysis revealed an overall difference between conditions [$\chi^2(3, N = 59) = 12.27, p < .01$]. Inspection of the post hoc cell contributions revealed that more no-error no-prior subjects were successful than would be expected by chance ($p < .05$). Collapsing hypotheses revealed an overall effect for the presence of system error [$\chi^2(1, N = 59) = 9.82, p < .01$]. Analysis of the post hoc cell contributions showed that fewer error subjects were successful than would be expected by chance ($p < .05$).

Number of Trials

One possible effect of generating prior hypotheses may be to increase subjects' search efficiency. In particular, given the confusion that system error evokes, a source of prior hypotheses may be especially beneficial for error subjects. If so, there should be a decrease in the number of trials that these subjects generate. However, as Table 6 shows, there was no effect of prior hypotheses on the number of trials conducted. As in Experiment 1, error subjects generated approximately twice as many trials as did no-error subjects [17.2 vs. 9.0; $F(1,55) = 22.4, p < .01$].

Nil Hypothesis Trials

During Experiment 1, error subjects' comments suggested that they had difficulty generating new hypotheses to test. Since prior-hypotheses subjects have a set of hypotheses to fall back on, we expected that they would produce fewer nil trials than would no-prior-hypotheses subjects. In the no-prior-hypotheses condition, 10% of no-error subjects' trials and 3% of the error subjects' trials were conducted without hypotheses; however, all of the no-error nil trials were by a single subject. None of the prior-hypotheses subjects conducted a nil hypothesis trial. Thus, contrary to our expectations, subjects rarely conducted experiments without an explicit hypothesis.

+Htest Strategy

Regardless of when hypotheses are generated, they must still be evaluated by generating experimental trials. As in Experiment 1, approximately 80% of subjects' trials involved +Htests.[4] There was no effect of error or of prior hypotheses on the proportion of +Htests proposed.

Experiment Feedback

As in Experiment 1, we analyzed feedback to +Htests and −Htests separately.

+Htest feedback. Virtually all subjects received verification on approximately 50% of their +Htests. Across conditions, the proportion of verified +Htests ranged from 39% to 50%; there was no effect of error or hypotheses.

−Htest feedback. Approximately 75% of the error subjects received only −Htest verification, as did 50% of the no-error subjects. Both error and no-error subjects received verification on roughly 75% of their −Htests.

Hypothesis Change

As discussed in Experiment 1, if people have an immunization bias, error subjects should be less likely to change their hypothesis following falsification than should no-error subjects. No-error subjects changed their current hypothesis following 66% of their falsification trials, whereas 73% of the error subjects did so; the difference was nonsignificant.

Table 6
Mean Number (and Standard Deviation) of Trials (Experiment 2)

	No Prior Hypotheses		Prior Hypotheses		Combined	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*
No errors	10.0	4.0	7.9	3.5	9.0	3.9
Errors	17.7	10.0	16.6	6.8	17.2	8.4
M	13.9	8.5	12.4	7.0		

Response to Error Trials

The prior-hypotheses condition was designed to determine the extent to which the effects of system error could be mitigated. However, we have already shown that this manipulation did not improve task success; this suggests that a pool of hypotheses does not aid in the identification of error trials.

There was no effect of prior hypotheses on the proportion of false positives or false negatives identified (see Table 7). Collapsing over conditions, 34% of the trials that subjects marked as errors were verification trials; 66% were falsification trials [paired $t(25) = 2.15$, $p < .05$].

Virtually all of the error subjects marked at least one falsification trial as being an error; however, only about 50% of the subjects, in either error condition, labeled at least one verification trial as an error.

The proportions of correct error identification in Experiment 2 were similar to those for the corresponding condition in Experiment 1: The subjects in both hypotheses conditions identified approximately 50% of their false positives; similarly, the subjects identified approximately 62% of their false negatives.

Experiment Replication

As in Experiment 1, error subjects varied widely in their use of a replication strategy; 52% of the subjects never replicated any of their test triples. The remaining subjects replicated between one and nine trials. On the average, error subjects replicated 13% of their trials. There was no effect of the hypothesis manipulation on the proportion of trials replicated.

Power of replication. As discussed in Experiment 1, replication is most useful if subjects choose error trials to replicate: Overall, 48% of subjects' replications involved false feedback trials. Testing the proportion of error trial replications against the actual error rate of 20% showed that error trials were more likely to be replicated than would be expected by chance [$t(13) = 2.84$, $p < .05$].

Task Success

Strict success. As in the first experiment, successful and unsuccessful subjects did not differ with respect to the proportion of new hypotheses generated following falsification (approximately 65% in each case).

Success in Experiment 1 was associated with lower proportions of +Htests. This result was replicated in Experiment 2. Sixty-nine percent of successful subjects' tests were +Htests; 86% of unsuccessful subjects' tests were +Htests [$F(1,57) = 8.17$, $p < .01$].

Table 7
Mean Proportion (and Standard Deviation) of Errors Identified by Condition (Experiment 2)

	No Prior Hypotheses			Prior Hypotheses			Combined	
Errors	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
False positives	.59	.39	14	.46	.50	14	.53	.44
False negatives	.71	.49	7	.55	.48	11	.61	.48

We also assessed the association between success and proportion of verified +Htests in no-error and error conditions. Successful and unsuccessful subjects received verification on approximately 50% of their +Htests. Although there was a difference on this measure in Experiment 1, both results support our argument that successful rule discovery is not associated with a higher proportion of verified +Htests. As in Experiment 1, successful and unsuccessful subjects did not differ on the proportion of verified −Htests (approximately 74% for both).

As in Experiment 1, conclusive falsification played an important role in task success. Successful and unsuccessful subjects received about the same proportions of conclusive falsification (42% and 49%, respectively). However, the two groups differed in the source of that feedback: for successful subjects, 19% were in the form of falsified −Htests; for unsuccessfuls, only 6% of their conclusive falsification was in this form.

Since only two error subjects successfully discovered the rule, analyses of error identification and experiment replication with respect to strict success were not conducted.

Near success. We analyzed the dependent measures with respect to the near-success criterion described in Experiment 1. In the no-prior condition, 93% of the no-error subjects and 40% of the error subjects were classified as near successful; in the prior-hypotheses condition, 86% of the no-error and 60% of the error subjects met the near-success criterion. A chi-square analysis showed a significant difference between conditions for rule discovery [$\chi^2(3, N = 59) = 12.55$, $p < .01$]. Inspection of the post hoc cell contributions revealed that in the no-prior condition, no-error subjects were more successful, whereas error subjects were less successful, than would be expected by chance alone ($p < .05$ for both cases).

Collapsing the hypotheses conditions revealed a main effect for error [$\chi^2(1, N = 59) = 10.94$, $p < .001$]. Examination of the post hoc cell contributions revealed that more error subjects were successful than would be expected by chance ($p < .05$).

A series of analyses of variance revealed no difference between near-successful and unsuccessful subjects with respect to proportion of nil hypothesis trials, proportion of hypothesis changes following falsification, proportion of +Htests generated, proportion of verified +Htests and −Htests, proportion of false positive errors identified, and proportion of trials, error and correct, replicated. Use of the near-success criterion had little effect on the pattern of conclusive falsification found using the strict-success criterion.

Analysis did reveal that near-successful subjects correctly identified 88% of their false negative errors, whereas 40% of the unsuccessful subjects did so [$F(1,16) = 5.67$, $p < .05$].

Discussion

The results were generally consistent with those of Experiment 1, with one important difference. The subjects in Experiment 1 identified an approximately equal pro-

portion of verifications and falsifications as errors. This led us to conclude that previous attributions of an immunization bias in subjects' responses to system error derive from a methodological artifact that disappears when the task promotes a more balanced proportion of both types of feedback. In Experiment 2, however, subjects' behavior was consistent with an immunization bias: they identified proportionally fewer verification trials than falsification trials as errors. The difference in the two results suggests that people vary widely in their error attribution strategies. That is, error attribution appears to be much more complicated than is suggested by a simple immunization bias. We will discuss this issue below.

There were two principal motivations for Experiment 2: (1) to investigate whether generating hypotheses prior to experimentation improved success in the error condition; and (2) to see whether subjects would choose to use a nil-hypothesis strategy, and if so, whether such a strategy would be effective.

We had expected subjects to conduct nil-hypothesis trials in order to establish a base of experimental results that could then be used to induce a hypothesis. However, the results suggest—contrary to our expectations—that people prefer to test, rather than induce, hypotheses, even when they lack specific information to initially guide their search. That is, people prefer to use their content knowledge to guide experimentation right from the start, even if they have no idea which aspects of that knowledge are applicable.

Freedman (1992a) suggested that considering multiple hypotheses mitigated the effects of system error. In contrast, our results showed no effect of prior hypotheses on task success. However, there are important differences between the two studies. Freedman instructed subjects to test either one or two hypotheses at a time. This produced a situation similar to that of Tweney et al. (1980). Subjects could construct two exclusive hypotheses and a test sufficient to eliminate one or the other of the two. In contrast, we did not instruct subjects in how they should go about testing their hypotheses.

Second, Freedman (1992a) did not actually introduce system errors; he only suggested that such errors might occur. As Gorman (1989) showed, there are differences in how people respond to actual and implied system error manipulations. These differences make it difficult to compare our results with those of Freedman.

Our procedure is closer to that of Klahr and Dunbar (1988), with the addition of system error. In contrast to that study, in the present work we did not find that prior generation of hypotheses facilitated task success, improved the efficiency of the discovery process, or mitigated the effects of system error. Thus, simply listing hypotheses before experimentation does not necessarily mitigate the effects of system error. However, this conclusion must be considered with respect to the size of the hypothesis space that subjects explore.

A priori hypotheses will be useful to the degree that they encompass the area of the hypothesis space within which the correct hypothesis lies. For example, in Klahr and Dunbar's (1988) task, most people correctly assumed that the name of the "mystery" computer key—RPT—was representative of the key's function. That is, although the key could have an infinite number of arbitrary functions, people used its name to constrain the range of hypotheses they proposed. In contrast, the seed [2–4–6] was purposely designed to suggest a wide range of plausible hypotheses (Wason, 1960). Consequently, people have little a priori information on how to constrain their set of possible hypotheses. As our judges' ratings reflect, subjects' a priori hypotheses varied considerably in their resemblance to the correct rule.

GENERAL DISCUSSION

This research was motivated by four main issues: (1) to investigate Kern's (1982) and Gorman's (1986, 1989) contention that people are biased to consider only falsification trials as possible errors; (2) to see whether, given the possibility of system error, a +Htest strategy is a useful search strategy, as Klayman and Ha (1987) have claimed; (3) to investigate whether or not generating hypotheses prior to experimentation mitigated the effects of system error; and (4) the interaction of system error and domain knowledge.

By using both broad and narrow rules, we were able to contrast situations with unbalanced and balanced proportions of verification and falsification trials. Our results show that, in contrast to the claim that people are biased to label falsification trials as errors in order to preserve their current hypothesis, people are equally able to detect false positives and false negatives.

Our results also show that, with respect to the efficacy of +Htests, successful subjects in both error and no-error conditions conducted proportionally *fewer* +Htests than did unsuccessful subjects. This suggests that although a +H strategy may be useful in establishing a viable hypothesis for further exploration, over-reliance on +Htests negatively affects task success (Tweney et al., 1981).

Previous research suggests that generation of prior hypotheses can be a useful discovery strategy (Klahr & Dunbar, 1988). However, the current research shows that without some constraints on their search, people are unlikely to generate hypotheses that cover the appropriate region of the hypothesis space.

When we used a strict criterion for success, we found similar search strategies and success rates in the two domains. However, the near-success analysis highlighted the effect of system error on domain: subjects in the narrow-rule animal domain were severely affected by the inclusion of system error, unlike their counterparts in the numerical domain.

The near-success analysis raises questions about the determination of success in studies of scientific reasoning, and in real-world scientific endeavors. In such contexts, success is not an all-or-none affair. Moreover, the form of the rule, the domain, and the type of feedback all affect assessments of task success. That is, although using a strict-success criterion does not always adequately rep-

resent people's ability, it is not clear how rule, domain, and feedback interact to affect the discovery process. Future work needs to address this issue.

Our analysis of successful and unsuccessful subjects' error identification patterns did produce one unexpected result. In both studies, successful subjects identified all of their false negatives, but not all of their false positives. Since knowing the rule provides sufficient information for identifying all errors, why then did successful subjects not identify all of their false positives? One explanation would be that they simply did not care to go back at the end of the study and isolate such errors. However, all subjects were encouraged to, and did, check over their record sheets at the conclusion of the study. Thus, it is unlikely that subjects cared about identifying their false negatives, but not their false positives.

One difficulty in understanding the link between error identification and the discovery process is to know how the two interact: establishing a viable hypothesis is necessary for identifying error trials; but identifying error trials is necessary for deciding on a viable hypothesis. It is unsurprising that the identification of error trials is critical to task success. However, it is less clear what strategies people use to identify possible errors. One strategy is to replicate suspect trials. Our results show that people varied widely in their use of this strategy. However, the subjects who did replicate tended to replicate error trials, rather than no-error trials, more often than would be expected by chance. What is not clear from the current work is *how* they made this decision.

One possible answer to this question is suggested by recent work on the treatment of anomalous data. Chinn and Brewer (1993) found that the strongest influence on how such data were treated was a person's theoretical commitment. Strong theoretical commitments lead people to discount data that do not fit with their position; that is, theory drives the determination of data validity.

However, Chinn and Brewer's (1993) methodology is considerably different than ours. They provided subjects with a theoretical position in a domain where most lay people have little knowledge (i.e., the mass extinction of dinosaurs at the end of the Cretaceous period). Subjects were then asked to assess data that did or did not fit with their assigned theoretical stance. Thus, subjects were immediately aware whether or not the presented data was consistent with their assigned theory. In contrast, we presented subjects with a task in which they had considerable domain knowledge, but little insight into the specific structure of the rule they were to discover. This produced a situation in which subjects had first to generate a hypothesis and then to evaluate the feedback with respect to that hypothesis.

So, when *do* people trust the data? It appears that the default is to trust the data when you do not distrust it. That is, if people are not sure whether or not feedback is erroneous, they initially accept it as veridical and use this information to develop their initial hypotheses. Increasing confidence in one's hypothesis leads to the flagging of suspect trials. Flagged trials are often replicated, because, as we teach our students from their very first research methods course, replication can be a powerful tool for error detection and correction.

REFERENCES

BREHMER, B. (1979). Preliminaries to a psychology of inference. *Scandinavian Journal of Psychology*, **20**, 193-210.

BREHMER, B. (1980). In one word: Not from experience. *Acta Psychologica*, **45**, 223-241.

BREHMER, B. (1987). Note on subjects' hypotheses in multiple-cue probability learning. *Organizational Behavior & Human Decision Processes*, **40**, 323-329.

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. (1956). *A study of thinking*. New York: New York Science Editions.

CASTELLAN, N. J. (1977). Decision making with multiple probabilistic cues. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 117-147). Hillsdale, NJ: Erlbaum.

CHINN, C. A., & BREWER, W. F. (1993). Factors that influence how people respond to anomalous data. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 318-323). Hillsdale, NJ: Erlbaum.

DOHERTY, M. E., & TWENEY, R. D. (1988). *The role of data and feedback error in inference and prediction*. (Final report for ARI Contract MDA903-85-K-0193). Bowling Green, OH: Bowling Green State University.

EINHORN, H. J., & HOGARTH, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, **32**, 22-53.

FARRIS, H. H. (1992). *Rule discovery heuristics: Goal-switching between counterfactual and positive test strategies in an adaptive system of heuristic search*. Unpublished doctoral dissertation, University of California, Santa Barbara.

FREEDMAN, E. G. (1992a, November). *The effects of possible error and multiple hypotheses on scientific induction*. Paper presented at the meeting of the Psychonomic Society, St. Louis.

FREEDMAN, E. G. (1992b). Scientific induction: Multiple hypotheses and individual and group processes. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 183-188). Hillsdale, NJ: Erlbaum.

GORMAN, M. E. (1986). How the possibility of error affects falsification on a task that models scientific problem solving. *British Journal of Psychology*, **77**, 85-96.

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

GORMAN, M. E. (1992). *Simulating science: Heuristics, mental models, and technoscientific thinking*. Bloomington: Indiana University Press.

KERN, L. H. (1982). *The effect of data error in inducing confirmatory inference strategies in scientific hypothesis testing*. Unpublished doctoral dissertation, Ohio State University.

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

KLAYMAN, J. (1984). Learning from feedback in probabilistic environments. *Acta Psychologica*, **56**, 81-92.

KLAYMAN, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 317-330.

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, & Cognition*, **15**, 596-604.

MOERTEL, C. G., FLEMING, T. R., MACDONALD, J. S., HALLER, D. G., LAURIE, J. A., & TANGEN, C. (1993). An evaluation of the carcinoembryonic antigen (CEA) test for monitoring patients with resected colon cancer. *Journal of the American Medical Association*, **270**, 943-948.

O'CONNOR, R. M., JR., DOHERTY, M. E., & TWENEY, R. D. (1989). The effects of system failure error on predictions. *Organizational Behavior & Human Decision Processes*, **44**, 1-11.

SLOVIC, P., FISCHHOFF, B., & LICHTENSTEIN, S. (1971). Behavioral decision theory. *Annual Review of Psychology*, **28**, 1-39.

TWENEY, R. D., DOHERTY, M. E., & MYNATT, C. R. (1981). Hypothesis testing: The role of confirmation. In R. D. Tweney, M. E. Doherty, & C. R. Mynatt (Eds.), *On scientific thinking* (pp. 115-128). New York: Columbia University Press.

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.

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

Whether positive or negative, result of prostate cancer test can create a maze of questions. (1993, June 23). *The New York Times*, p. C12.

YORK, K., DOHERTY, M., & KAMOURI, J. (1987). The influence of cue unreliability on judgment in a multiple cue probability learning task. *Organizational Behavior & Human Decision Processes*, **39**, 303-317.

NOTES

1. See Gorman (1992) for a comprehensive review of his (and other's) work on the Wason task, and related scientific discovery tasks.

2. For example, following surgery, colon cancer patients are monitored for increased levels of a biological marker for colorectal cancer. A 6-year study has shown that the test used produced false negatives for 41% of the cases, and false positives for 16% of the cases (Moertel et al., 1993). Or to cite another example, a recently developed and widely used test for early detection of prostate cancer has a 20% false negative rate and a 25% false positive rate ("Whether Positive or Negative," 1993).

3. The error subjects' feedback pattern is partially determined by which trials are marked as errors. That is, the *perceived* pattern of "yes" and "no" responses depends in part on which trials are marked as possible errors. Analyses based on subjects' perceived data patterns will be noted as they arise.

4. Because +Htest classification is defined with respect to subjects' stated hypotheses, nil trials are excluded from this analysis.

APPENDIX
Experiment Instructions

You will be given a set of three observations that conform to a simple rule. The rule is concerned with a relation between sets of three observations. Your task is to discover this rule by generating a possible rule (a hypothesis) and testing it with your own set of three observations. After you have written down your set of observations, I will indicate whether or not your test conforms to the rule by placing a check under either the "Conform" or "Does Not Conform" column on your paper.

Consider this procedure as similar to performing a number of mini-experiments: You propose a hypothesis, test it with an experiment, and evaluate the outcome of the experiment (i.e., feedback from the experimenter). Treat the feedback from the experimenter as evidence, and use it to evaluate your hypothesis.

You may propose as many rules and tests as you wish. You are free to refer to previous trials at any time during the study. If you wish, you may retain a hypothesis, and/or a test, from a previous trial. Remember, your aim is to discover the rule that describes the relationship between the objects in the set. Continue until you are sure that you know what the rule is. ONLY THEN, AND NOT BEFORE, ARE YOU TO WRITE THIS RULE DOWN ON THE BACK OF YOUR RECORD SHEET. Do you have any questions at this time?

Instructions for the error conditions included the following: To make this more like real science, there may be some "noise" or random error in the feedback you receive, i.e., if the triad actually DOES match the rule, it may be labeled as Does Not Conform, and vice versa. On anywhere from 0%–25% of the trials, the feedback you receive will be incorrect. For example, your 7th trial may be classified incorrectly, as may your 17th trial. The amount of error can never exceed an average of one trial in five, and there may be NO error at all.

To indicate where you think random error has occurred, please mark the trials you think have been classified incorrectly with an X. If you change your mind, place a check mark next to the X. This will indicate that you no longer believe you received incorrect feedback on that trial.

(Manuscript received March 29, 1994;
revision accepted for publication August 10, 1995.)