*Psychonomic Bulletin & Review*
*2000, 7 (2), 360-366*

# The psychological side of Hempel's paradox of confirmation

CRAIG R. M. McKENZIE and LAURIE A. MIKKELSEN
*University of California, San Diego, California*

People often test hypotheses about two variables ($X$ and $Y$), each with two levels (e.g., $X1$ and $X2$). When testing "If $X1$, then $Y1$," observing the conjunction of $X1$ and $Y1$ is overwhelmingly perceived as more supportive than observing the conjunction of $X2$ and $Y2$, although both observations support the hypothesis. Normatively, the $X2\&Y2$ observation provides stronger support than the $X1\&Y1$ observation if the former is rarer. Because participants in laboratory settings typically test hypotheses they are unfamiliar with, previous research has not examined whether participants are sensitive to the rarity of observations. The experiment reported here showed that participants were sensitive to rarity, even judging a rare $X2\&Y2$ observation more supportive than a common $X1\&Y1$ observation under certain conditions. Furthermore, participants' default strategy of judging $X1\&Y1$ observations more informative might be generally adaptive because hypotheses usually regard rare events.

A fundamental issue in the study of human inference is what, psychologically speaking, constitutes confirmatory evidence for a hypothesis. In 1945, philosopher Carl Hempel noted the following paradox regarding confirmatory evidence. Assume that the hypothesis of interest is "All ravens are black." This statement can be rewritten as "If something is a raven, then it is black," or Raven → Black. Clearly, observing a black raven would count as confirming evidence. Similarly, if the hypothesis were "If something is not black, then it is not a raven," or ~Black → ~Raven, observing a nonblack nonraven (e.g., a white shoe or a yellow pencil) would clearly be confirming evidence. Because these two hypotheses are logically equivalent (one is the contrapositive of the other), any evidence that confirms one must confirm the other. It follows, then, that observing a nonblack nonraven confirms Raven → Black. Thus, one could apparently confirm the hypothesis about the color of ravens by sitting in one's office and never even observing a raven. Most people find this highly counterintuitive; hence, the paradox.

Other philosophers have pointed out that the paradox can be resolved if one conceives of confirmation as a matter of degree rather than all-or-none. Although black ravens and nonblack nonravens both confirm Raven → Black, they do not do so equally strongly. From a Bayesian perspective, confirming evidence supports a hypothesis to the extent that it is rare, or surprising. Because nonblack things and nonravens are both common, observing a nonblack nonraven would not be unusual and would therefore confirm the hypothesis only negligibly. In contrast, because few things are black and few things are ravens, observing a black raven would be surprising and would constitute stronger confirmation (Alexander, 1958; Good, 1960; Horwich, 1982; Hosiasson-Lindenbaum, 1940; Howson & Urbach, 1989, pp. 88–91; Mackie, 1963).[1] The paradox appears to stem from our inability to distinguish intuitively between nonconfirmatory and minutely confirmatory evidence: The nonblack nonraven appears completely uninformative but, in fact, provides very weak confirmation.

The question examined in this article is whether people are sensitive to how "surprising" data are. If so, then they should consider, for example, observing a nonblack nonraven to be of virtually no value when testing Raven → Black. This, of course, is consistent with both intuition and normative analyses, so showing this would not be particularly interesting. However, of considerable interest is whether people consider, for example, observing a black raven (the rare confirming observation) to be informative when testing ~Black → ~Raven.

Because we did not examine negations in the experiment reported here, consider the issue in negation-free terms. Assume two variables, $X$ and $Y$, each with two mutually exclusive and exhaustive levels ($X1$ and $X2$, $Y1$ and $Y2$). Assume further that $X1$ is more common than $X2$, and $Y1$ is more common than $Y2$. When testing "If $X1$, then $Y1$," will participants consider the rare $X2\&Y2$ observation to be more informative than the common $X1\&Y1$ observation?

The extant literature on the psychology of hypothesis testing clearly predicts "no." The consensus is that people test hypotheses of the form $X1 \rightarrow Y1$ by checking to see whether $Y1$ occurs (or is true) when $X1$ occurs. What happens when $X2$ occurs is considered less important, if not irrelevant. Thus, of the two confirming instances, the $X1\&Y1$ observation would be predicted to be judged

This research was supported by National Science Foundation Grant SBR-9515030. The authors thank Mark Machina for introducing us to Hempel's paradox and thank Nick Chater, Josh Klayman, Raymond Nickerson, and Mike Ziolkowski for many helpful comments. Correspondence should be addressed to C. R. M. McKenzie, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109 (e-mail: cmckenzie@ucsd.edu).

 

more informative (Evans, 1989; Fischhoff & Beyth-Marom, 1983; Johnson-Laird & Tagart, 1969; Klayman & Ha, 1987; McKenzie, 1994, 1998; Wason, 1960, 1966). Put differently, participants judge the supporting observation mentioned in the hypothesis to be more informative than the unmentioned supporting observation.

However, virtually all psychological studies use hypotheses about which participants know essentially nothing. This is a reasonable strategy in that it puts participants on equal footing when testing hypotheses in the laboratory. Among other things, it eliminates (or at least reduces) variance in prior beliefs regarding the truth of the hypotheses to be tested. But in order to examine whether people are sensitive to the rarity of data, participants must know the relative commonality of *X*1 versus *X*2 and of *Y*1 versus *Y*2. When presented with unfamiliar or abstract hypotheses, participants might resort to the general heuristic of examining instances named in the hypothesis (Klayman & Ha, 1987), but they might utilize information regarding rarity when it is available. This could occur either by presenting participants with familiar hypotheses, in which case general knowledge about rarity might be exploited, or by directly informing participants about rarity.

In the experiment reported here, participants were given a hypothesis to test and chose whether observing *X*1&*Y*1 or *X*2&*Y*2 provided stronger support. One group tested one of two *abstract* hypotheses with which the participants were unfamiliar. The hypothesis tested was either "If a person has genotype *A*, then he/she has personality type *X*" or "If a person has personality type *Y*, then he/she has genotype *B*." Participants were told that everyone had personality type *X* or *Y*, and everyone had genotype *A* or *B*. They then chose which of two observations was more supportive of the hypothesis: a person with genotype *A* and personality type *X*, or a person with genotype *B* and personality type *Y*. Both observations support both hypotheses, but, on the basis of previous findings, we expected participants to choose as more supportive the observation named in the hypothesis. That is, we expected those testing Genotype *A* → Personality Type *X* to select as more supportive the genotype *A*/personality type *X* person and those testing Personality Type *Y* → Genotype *B* to select the genotype *B*/personality type *Y* person. A second group tested either abstract hypothesis but was told that a small minority of people have genotype *B* and a small minority have personality type *Y*. At issue was whether participants would be more likely to select the genotype *B*/personality type *Y* person as more supportive when provided with explicit information about rarity.

A third group tested one of two *concrete* hypotheses, either "If a person is HIV+, then he/she is psychotic" or "If a person is mentally healthy, then he/she is HIV−." They then selected which of two observations, an HIV+/psychotic person and a mentally healthy/HIV− person, provided stronger support. Both are confirming observations, but the former is rarer than the latter—and participants presumably know this. If participants are sensitive to the rarity of data, then the rare observation should be considered more supportive across hypotheses. A final group performed the same task with one of the concrete hypotheses but was also "reminded" that a small minority of people are HIV+ and a small minority are psychotic.

Thus, the abstract group had neither real-world knowledge nor statistical information regarding the rarity of the observations; the abstract + statistics group had only statistical information; the concrete group had only real-world knowledge; and the concrete + statistics group had both real-world knowledge and corroborating statistical information.

## METHOD

The participants were 282 University of California, San Diego, students who received partial course credit for introductory psychology courses. They were given a three-page booklet, the first page of which provided general instructions. The second page consisted of text to read and one question to answer. The participants in the abstract and abstract + statistics groups were told they were researchers investigating a possible relation between genetics and personality (or personality and genetics). For their purposes, there were four categories of people: (1) those who have genotype *A* and personality type *X*, (2) those who have genotype *A* and personality type *Y*, (3) those who have genotype *B* and personality type *X*, and (4) those who have genotype *B* and personality type *Y*. (Order of these categories was reversed for half of the participants.) Half of the participants were told that the hypothesis to be tested was "If a person has genotype *A*, then he/she has personality type *X*," and the other half were told to test "If a person has personality type *Y*, then he/she has genotype *B*." They were then informed that, of the first two people observed, one had personality type *X* and genotype *A* and one had personality type *Y* and genotype *B*, and they were then asked to choose which observation provided stronger support of the hypothesis they were testing. (Order of the observations was reversed for half of the participants.) Before responding, abstract + statistics participants read information that they "might want to take into consideration": (1) A large majority of the population has personality type *X* and a small minority has personality type *Y*. (2) A large majority of the population has genotype *A* and a small minority has genotype *B*.

Because a small minority has either personality type *Y* or genotype *B*, the personality type *Y*/genotype *B* person will be referred to as the "rare observation," and the personality type *X*/genotype *A* person as the "common observation." Similarly, the personality type *Y*/genotype *B* hypothesis and the genotype *A*/personality type *X* hypothesis will be referred to as the "rare hypothesis" and the "common hypothesis," respectively, but this should not be taken to reflect the prior probability of the hypotheses, which are equivalent in the sense that both are supported by the same observations. Furthermore, though the abstract group has no basis for inferring rarity, we will still refer to "rare" and "common" observations and hypotheses when discussing this group to maintain consistency and provide benchmarks for comparison.

The participants in the concrete and concrete + statistics groups were told they were researchers examining the possible relation between mental health and AIDS. For their purposes, there were four categories of people: (1) those who were HIV+ and psychotic, (2) those who were HIV+ and mentally healthy, (3) those who were

HIV− and psychotic, and (4) those who were HIV− and mentally healthy. Half of these participants were told that the hypothesis to be tested was "If a person is HIV+, then he/she is psychotic," and half were told to test the hypothesis "If a person is mentally healthy, then he/she is HIV−." They were then informed that, of the first two people observed, one was HIV+ and psychotic and the other was mentally healthy and HIV−, and they were asked to select which of the two observations provided stronger support for the hypothesis. Before responding, the concrete + statistics group read: "(1) A large majority of the population is HIV− and a small minority is HIV+. (2) A large majority of the population is mentally healthy and a small minority is psychotic." For the two concrete groups, the HIV+/psychotic individual is rare, and the mentally healthy/HIV− individual is common. Similarly, the HIV+/psychotic hypothesis and the mentally healthy/HIV− hypothesis are the "rare hypothesis" and "common hypothesis," respectively.

On the final page, the participants justified their choice of which observation provided stronger support of the hypothesis. They were instructed not to change the choice they had already made.

To summarize, the experiment was a completely between-participants 2 (context: abstract vs. concrete) × 2 (statistical information: yes vs. no) × 2 (hypothesis tested: common vs. rare) design. There were between 34 and 36 participants in each condition. They chose either the common or rare observation as more supportive of the hypothesis tested and explained why.

## RESULTS

Figure 1 shows the percentage of participants who selected the rare observation as providing stronger support as a function of group (abstract, abstract + statistics, concrete, and concrete + statistics) and hypothesis tested (common and rare). The abstract participants (far left) serve as a control group, showing how often the observation (not) named in the hypothesis is chosen when no information is available regarding the rarity of the observations. The low value for the common hypothesis reflects the fact that few participants in the abstract group selected the rare observation as more supportive when testing the common hypothesis, and the high value for the rare hypothesis shows that many chose the rare observation when testing the rare hypothesis. These results replicate those of many previous studies and illustrate the tendency to perceive as more informative the observation named in the hypothesis.

For the rest of the groups, who were told the relevant statistical information and/or were assumed to have already known it, the normative response was to choose the rare observation as more supportive, regardless of the hypothesis tested. Thus, to the extent that both columns in Figure 1 are large for a given group, the better the group's performance. It is easy to see that, across the four groups, there was little variation in selecting the rare observation when the rare hypothesis was tested (hatched columns). All groups tended to select the rare observation when testing the rare hypothesis, though the two concrete groups did so slightly more often than the two abstract groups.

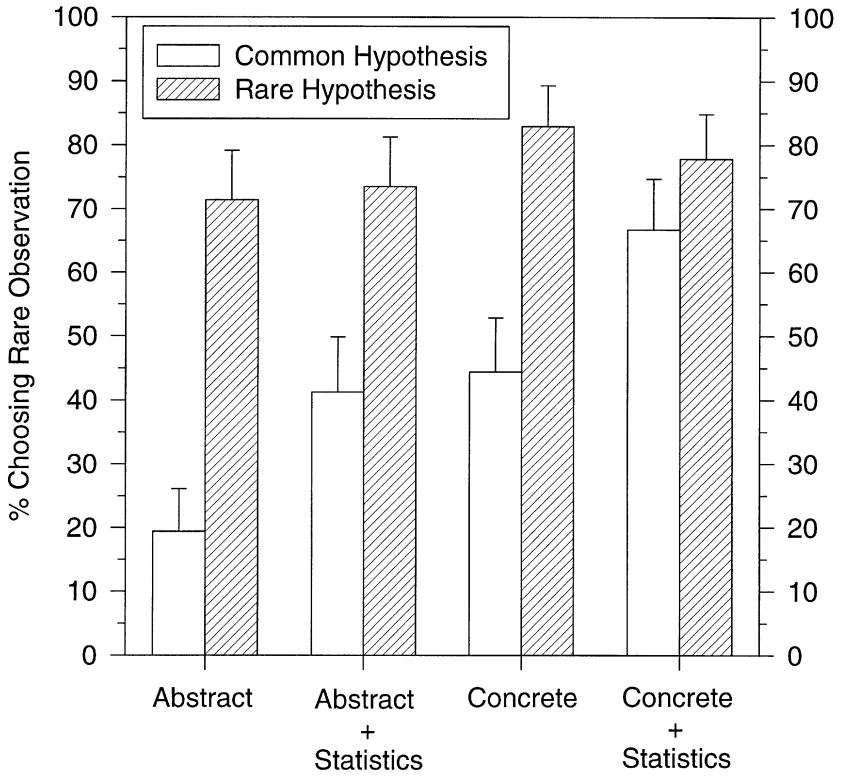


**Figure 1. Percentage of participants selecting the rare (rather than common) observation as a function of group (abstract, abstract + statistics, concrete, and concrete + statistics) and hypothesis tested (common and rare). Standard error bars are shown.**

Choosing the rare observation when testing the rare hypothesis is consistent with both preferring the named observation and preferring the normatively more supportive observation (though see the analysis of the participants' reasons for their choices below). More informative was the participants' behavior when testing the common hypothesis. Figure 1 reveals that the percentage of rare observation choices when testing the common hypothesis varied considerably across the groups. Relative to the abstract group, the abstract + statistics group was more likely to select the rare observation, showing that the statistical information led to more normative responses. The concrete group also selected the rare observation more often than the abstract group, which indicates that the participants' real-world knowledge was helpful. Of interest is that the concrete group performed as well as or better than the abstract + statistics group, demonstrating that implicit real-world knowledge affected behavior at least as much as the explicit statistical information. The concrete + statistics group was by far the most likely to select the rare observation when testing the common hypothesis. Indeed, this group chose the rare observation twice as often as the common observation when testing the common hypothesis. As can be seen, the concrete + statistics group was about equally likely to select the rare observation regardless of hypothesis tested.

A 2 (context: abstract vs. concrete) × 2 (statistical information: yes vs. no) × 2 (hypothesis tested: common vs. rare) × 2 (response: common vs. rare observation) log-linear analysis corroborated the patterns noted above. The participants were more likely to select the rare observation when testing the rare hypothesis than when testing the common hypothesis ($p < .001$); they were also more likely to select the rare observation when presented with the concrete scenario than when presented with the abstract scenario ($p = .002$). In addition, they were marginally significantly more likely to choose the rare observation when presented with statistical information ($p = .056$). Furthermore, statistical information affected the participants testing the common hypothesis more than those testing the rare hypothesis ($p = .043$). Separate analyses confirmed that the proportion of participants choosing the rare observation when testing the rare versus common hypothesis was different for each of the abstract, abstract + statistics, and concrete groups ($p$s < .02, Fisher's exact tests), but not the concrete + statistics group ($p = .43$). In other words, only the concrete + statistics participants were not significantly influenced by the hypothesis tested when choosing an observation.

The participants' reasons for their choices were also analyzed. Reasons were categorized as "correct" if the participant both (1) chose the rare observation and (2) mentioned the relative commonality of one level of either variable (e.g., being HIV+) or one of the observations (e.g., being HIV+ and psychotic). Reasons were categorized as "naming" if the participant mentioned that the observation named in the hypothesis was more relevant. Any reason not falling under these two categories was categorized as "other." If more than one reason was given, any mention of rarity combined with choosing the rare observation was categorized as "correct," and reasons were categorized as "other" only if neither correct nor naming reason(s) were stated. Thus, each participant's reasons fell into exactly one of the three categories.

The results are shown in Table 1 by condition. Not surprisingly, participants in the abstract group never gave a "correct" reason for their choices because they had no information regarding rarity. More interesting is this group's percentage of "naming" reasons, which serves as a useful benchmark: About three fourths of these participants testing either hypothesis justified their choice by stating that the observation named in the hypothesis was more relevant. (The most common "other" reason for both the abstract group and the concrete group was stating that the unnamed observation must be observed to confirm the hypothesis.) The abstract + statistics group was more likely to provide "correct" reasons when testing either hypothesis, and more than half did so when testing the rare hypothesis. Thus, though the abstract and abstract + statistics groups chose the rare observation about equally often when testing the rare hypothesis (see Figure 1), they apparently did so for different reasons. (The most frequent "other" reason for both the abstract + statistics group and the concrete + statistics group was to mention rarity but select the common observation.)

Table 1 also shows that the concrete group stated correct justifications more often when testing the rare hypothesis than when testing the common hypothesis, but these responses constituted only a minority in either

**Table 1**
**Categorizations of Participants Based on Their Reasons for Selecting the Common Observation or the Rare Observation as Providing Stronger Support for the Hypothesis**

| | Abstract Scenario | | | | Concrete Scenario | | | |
|---|---|---|---|---|---|---|---|---|
| | No Statistics | | Statistics | | No Statistics | | Statistics | |
| Reason | Common H | Rare H | Common H | Rare H | Common H | Rare H | Common H | Rare H |
| Correct | 0.0 | 0.0 | 38.2 | 55.9 | 16.7 | 28.6 | 55.6 | 58.3 |
| Naming | 77.8 | 74.3 | 17.6 | 17.6 | 41.7 | 51.4 | 25.0 | 19.4 |
| Other | 22.2 | 25.7 | 44.1 | 26.5 | 41.7 | 20.0 | 19.4 | 22.2 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Note—"Common H" and "Rare H" refer to testing the common hypothesis and the rare hypothesis, respectively.

case. Given that this group chose the rare observation at least as often as the abstract + statistics group (see Figure 1), it is interesting that they were less likely to provide a correct justification. We compared the concrete and abstract + statistics groups in terms of the number of correct versus naming responses (ignoring other responses). The concrete group was less likely to provide a correct reason when testing both the common hypothesis and the rare hypothesis ($ps < .03$, Fisher's exact tests). Finally, Table 1 shows that more than half of the participants testing either hypothesis in the concrete + statistics group reported the correct reason for their choices. Thus, regardless of the hypothesis tested, a majority of this group chose the rare observation and reported the correct reason for doing so.

## DISCUSSION

The results of this experiment corroborate earlier findings that observations mentioned in the hypothesis are generally considered more valuable. The results also extend previous findings by demonstrating that participants' knowledge regarding the rarity of observations can reduce, perhaps even overcome, this tendency. The rare observation was more often correctly chosen as more supportive when (1) the hypotheses were concrete rather than abstract, which allowed the participants to exploit their real-world knowledge about rarity, and (2) statistical information about rarity was provided. The combination of a concrete hypothesis and statistical "reminder" led most participants to select the rare observation even when testing the common hypothesis. The participants in this latter condition were about equally likely to select the rare observation regardless of the hypothesis tested.

It was also found that participants' knowledge—both implicit real-world knowledge and explicit statistical information provided to them—about rarity led to more correct justifications of choices. Results for the concrete group were most interesting: This group selected the rare observation at least as often as the abstract + statistics group when testing either hypothesis, but they were less likely to provide a correct reason why. It seems reasonable to assume that the concrete group's choices were affected by real-world knowledge regarding rarity. Thus, the participants in this group could apparently express their real-world knowledge through their choices, but not through verbalization (Nisbett & Wilson, 1977). This is especially intriguing because the concrete scenario represents the most realistic situation. That is, people typically make inferences about variables they are familiar with, but they are not usually handed the relevant statistical information. The mechanism(s) governing hypothesis-testing behavior under these circumstances might operate at a level that goes unnoticed by most people.

The fact that rarity affected judgments illustrates that using familiar hypotheses enables participants to exploit an aspect of the hypothesis-testing situation that is unavailable when abstract or unfamiliar hypotheses are used—which is the norm in the laboratory. Klayman and Ha (1987) speculated that participants might prefer cases named in the hypothesis when presented with unfamiliar tasks but use more sophisticated strategies in more familiar territory. The present study provides empirical support for this notion. Generalizing findings based on unfamiliar hypothesis-testing tasks might mislead us as to people's ability in their natural environment.

Two concerns about our experiment could be raised. First, the hypotheses we asked the participants to test were different from the Raven → Black example in that the raven hypothesis is probably more likely to be viewed as deterministic. That is, observing a nonblack raven might be more likely to be seen as conclusively falsifying the raven hypothesis than, for example, observing a mentally healthy/HIV+ person would be seen as conclusively falsifying the Mentally Healthy → HIV− hypothesis. However, as we mentioned in note 1, the ordering of $X1\&Y1$ and $X2\&Y2$ observations in terms of their informativeness remains the same regardless of whether the hypothesis tested is deterministic or probabilistic.

Second, when analyzing and interpreting our data, we treated "context" as a variable, although there are undoubtedly many factors that differ between the abstract and concrete scenarios. For example, the statistical information presented to the participants might have been interpreted differently in the two scenarios. Nonetheless, our conclusions are similar whether we treat the two context conditions as two levels of a variable or as separate experiments. Relatedly, all participants were told that the four categories of observations were mutually exclusive and exhaustive, but this information might have been reinforced by the participants' real-world knowledge in the concrete, but not the abstract, conditions. For instance, it might have been easy to understand that a person must be either HIV− or HIV+, and not both, but perhaps this was less clear for the fictitious personality type *X* versus *Y* categories. There was, however, no evidence of any such doubt or confusion in the participants' reported reasons for their choices.

Our findings indicate that participants are sensitive to the rarity of data, but there is still the question of why, when presented with unfamiliar hypotheses, they tend to judge the confirming instance named in the hypothesis more informative than the confirming instance not named. Why is this the default strategy? We speculate that it is because hypotheses tend to be formed about rare events (see also Klayman & Ha, 1987). It is rare, not common, events that we seek to explain or predict (Einhorn & Hogarth, 1986; Mackie, 1974). Hypotheses are generated about the factors leading to being HIV+, not HIV−, about what causes plane crashes, not normal flights, about what leads to success, not mediocrity, and so on. Rare events are salient and "demand" explanation. If hypotheses tend to be framed in terms of rare events, a confirming observation named in the hypothesis will be more informative than a confirming observation not named. In the absence of information indicating other-

wise, behaving as though the named observation is rare might be generally adaptive.

Recent analyses of Wason's (1966, 1968) selection (or four-card) task have led to similar conclusions. In this task, participants are given a rule to test of the form "If *P*, then *Q*" and are presented with four cards: *P*, ~*P*, *Q*, and ~*Q*. On the back of the first two cards is either *Q* or ~*Q*, and on the back of the last two is either *P* or ~*P*. For example, participants might test the rule "If there is a vowel on one side of the card, then there is an even number on the other side," with the four cards showing A, K, 2, and 7. Each card has a letter on one side and a number on the other. Which cards are necessary to turn over in order to see if the rule is true or false? According to one interpretation of the rule, propositional logic dictates that the *P* and ~*Q* cards should be turned over (A and 7 in the example). However, a common finding is that participants want to turn over the cards mentioned in the rule: the *P* and *Q* cards (e.g., Wason, 1966, 1968). Oaksford and Chater (1994; see also Nickerson, 1996) have argued that, from a Bayesian perspective, the *P* and *Q* cards are the most informative—if one assumes that *P* and *Q* are rare relative to ~*P* and ~*Q*. As predicted by this account, participants presented with a reduced array selection task, where only the *Q* and ~*Q* cards are present, were more likely to select the ~*Q* card as the commonality of *Q* was increased (Oaksford, Chater, Grainger, & Larkin, 1997). Oaksford and Chater's (1994) account of the selection task meshes well with our account of hypothesis testing: When testing "If *P*, then *Q*," participants generally behave as though *P* and *Q* are rare, and information to the contrary changes behavior in the appropriate direction.[2]

Another related and highly studied task is covariation assessment. In a typical task, there are two variables, either present or absent, resulting in four types of observation. Participants are asked to assess the relation between the two variables given the frequencies of the four types of observation. Both the joint presence and the joint absence of the variables are confirming evidence for a positive relation between the variables, but the former observation has a much larger influence on judgments (e.g., Kao & Wasserman, 1993; Lipe, 1990; McKenzie, 1994; Wasserman, Dorner, & Kao, 1990). Thus, as with hypothesis testing, the confirming observation named in the relation (or hypothesis) to be tested has a bigger impact on behavior than the confirming observation not named. However, also as with hypothesis testing, covariation tasks tend to use variables that are unfamiliar or abstract, conditions that the present study shows lead to a preference for the named observation. Furthermore, the apparent default preference for the named conjunction might be due to the fact that a variable's presence is generally rarer than its absence. For example, when assessing the relation between a symptom and a disease, most people will not have the symptom and most people will not have the disease, making the probability of having both much lower than the probability of having neither. Participants might consider the joint presence of variables to be more informative than their joint absence because, outside the laboratory, the former observation is typically rarer—and, therefore, normatively more informative, at least from a Bayesian perspective (see also Anderson, 1990, pp. 149–160).

In sum, participants' tendency to deem as more informative confirming observations named rather than not named in the hypothesis can be reduced, maybe even eliminated, when participants know that the unnamed observation is rare. The norm in the laboratory, however, is to use hypotheses that participants are unfamiliar with, which precludes their ability to exploit such knowledge. Furthermore, that participants are sensitive to the rarity of data is consistent with our speculation that, outside the laboratory, named confirming observations are typically rarer than unnamed confirming observations, making the former generally more informative. Thus, participants' default hypothesis-testing strategy in the laboratory appears generally adaptive as well as *adaptable* (Klayman & Brown, 1993) in that it changes appropriately when normatively relevant information is available.

## REFERENCES


Alexander, H. G. (1958). The paradoxes of confirmation. *British Journal for the Philosophy of Science*, **9**, 227-233.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, **99**, 3-19.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis testing from a Bayesian perspective. *Psychological Review*, **90**, 239-260.

Good, I. J. (1960). The paradox of confirmation. *British Journal for the Philosophy of Science*, **11**, 145-149.

Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind*, **54**, 1-26, 97-121.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

Hosiasson-Lindenbaum, J. (1940). On confirmation. *Journal of Symbolic Logic*, **5**, 133-148.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.

Johnson-Laird, P., & Tagart, J. (1969). How implication is understood. *American Journal of Psychology*, **82**, 367-373.

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 1363-1386.

Klayman, J., & Brown, K. (1993). Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition*, **49**, 97-122.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, **94**, 211-228.

Lipe, M. G. (1990). A lens-model analysis of covariation research. *Journal of Behavioral Decision Making*, **3**, 47-59.

Mackie, J. L. (1963). The paradox of confirmation. *British Journal for the Philosophy of Science*, **13**, 265-277.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford: Oxford University Press, Clarendon Press.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, **26**, 209-239.

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 771-792.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking & Reasoning*, **2**, 1-31.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231-259.

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

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 441-458.

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. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135-161). Harmondsworth, U.K.: Penguin.

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

Wasserman, E. A., Dorner, W. W., & Kao, S.-F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 509-521.

## NOTES

1. From a Bayesian viewpoint, data, $D$, support a hypothesis, $H1$, over its alternative, $H2$, to the extent that the likelihood ratio, $p(D|H1)/p(D|H2)$, is greater than 1. Let $H1$ be Raven → Black, and let $H2$ be that the properties of being a raven and of being black are independent. Assume that the unconditional probability that an object (randomly sampled) is black is .01 and that the unconditional probability that an object is a raven is also .01. If being black and being a raven are independent ($H2$), then the probability of an object being both black and a raven is $.01^2$, or .0001, which is $p(\text{black raven}\,|\,H2)$. Similarly, given the independence hypothesis, the probability of observing a nonblack nonraven, $p(\text{nonblack nonraven}\,|\,H2)$, is $.99^2$, or .9801. If the Raven → Black hypothesis ($H1$) is true, then the probability of observing a black raven, $p(\text{black raven}\,|\,H1)$, is .01, because all ravens will be black. Again, because there are no nonblack ravens, all nonblack things must be nonravens, so $p(\text{nonblack nonraven}\,|\,H1) = .99$. Now we can compare likelihood ratios for the black raven observation,

$$p(\text{black raven}\,|\,H1)/p(\text{black raven}\,|\,H2) = .01/.0001 = 100,$$

and the nonblack nonraven observation,

$$p(\text{nonblack nonraven}\,|\,H1)/p(\text{nonblack nonraven}\,|\,H2) = .99/.9801 = 1.01.$$

In this example, the likelihood ratio for the black raven is 99 times greater than that for the nonblack nonraven. Note also that the latter likelihood ratio barely exceeds 1, showing that the nonblack nonraven is virtually uninformative.

When testing "If $X1$, then $Y1$," an $X1\&Y1$ observation is more informative than an $X2\&Y2$ observation whenever $p(X1) < 1 - p(Y1)$. This is true even when the hypothesized relation between $X$ and $Y$ is probabilistic, rather than deterministic, as in the above example. To model the probabilistic case, one can use the phi coefficient to represent an imperfect relationship between $X$ and $Y$ (e.g., $H1$ could be $\phi = .5$). Furthermore, the alternative hypothesis need not be independence (e.g., $H2$ could be $\phi = .1$). The result appears to be very general.

2. Our analysis does, however, differ from Oaksford and Chater's (1994) in an interesting way. Consider Raven → Black as a rule to be tested in a selection task. The four cards would be raven ($R$), not a raven ($\sim R$), black ($B$), and not black ($\sim B$). Analyses of the selection task—both logical and Bayesian—usually claim that turning over the $\sim R$ card is uninformative (e.g., Oaksford & Chater, 1994; Wason, 1966, 1968). However, our analysis indicates otherwise: Turning over the $\sim R$ card could reveal $\sim R\&\sim B$, which we showed is informative in note 1 above. Indeed, observing $\sim R\&B$ is also informative, so we claim that turning over the $\sim R$ card has informational value, at least when the marginal probabilities, $p(R)$ and $p(B)$, are not different under the competing hypotheses (see also Nickerson, 1996).

Our claims in this note do not depend on $R$ and $B$ being rare, but, for the sake of consistency and concreteness, assume as in note 1 that $p(R) = .01$, $p(\sim R) = .99$, $p(B) = .01$, and $p(\sim B) = .99$. If the rule is true ($H1$), then there should be no instances of $R\&\sim B$. Let $H2$ be that the properties $R$ and $B$ are independent. Because the marginals are fixed and $p(R) = p(B)$, $H1$ entails not only that there are no instances of $R\&\sim B$ but that there are no instances of $\sim R\&B$. If all ravens are black, and the number of ravens equals the number of black things, then nothing other than ravens can be black. Observing an instance of $\sim R\&B$ is informationally equivalent to observing an instance of $R\&\sim B$. When $p(R) = p(B)$, turning over the $\sim R$ card is informative: Finding $B$ on the back of the card falsifies the rule, whereas finding $\sim B$ provides evidence for the rule (note 1).

Now consider the case in which $p(R) < p(B)$. That is, even if all ravens are black, there will still be instances of black nonravens. Nonetheless, these $\sim R\&B$ instances are informative, though less so than $R\&\sim B$ instances. If $p(R) = .01$, and $p(B) = .05$, then the likelihood ratio for the $\sim R\&B$ observation, $p(\sim R\&B\,|\,H1)/p(\sim R\&B\,|\,H2)$, equals $.04/.0495$. The ratio is less than 1, showing that a $\sim R\&B$ observation is evidence against $H1$. The likelihood ratio for the $\sim R\&\sim B$ observation is $.95/.9405$, which is greater than 1, and shows that it is evidence for $H1$. Thus, turning over the $\sim R$ card when $p(R) < p(B)$ is also informative: $B$ on the other side provides evidence against the rule, whereas $\sim B$ on the other side provides evidence for the rule.

Finally, consider the situation in which $p(R) > p(B)$. Because there are more ravens than black things, there must be instances of nonblack ravens. One need not turn over any cards to know the rule is false.

In short, our analysis indicates that, under circumstances in which at least one card should be turned over, turning over the $\sim R$ card is informative if the marginal probabilities of the properties being tested are the same under the competing hypotheses. In terms of our example, this means that whether or not the rule is true does not affect the proportion of things that are ravens and the proportion of things that are black. Oaksford and Chater (1994), on the other hand, fix $p(R)$ but let $p(B)$ depend on the truth of the rule in a way that, *by design*, renders the $\sim R$ card uninformative (p. 610). The reasonableness of the assumptions regarding if and how the marginal probabilities vary under the competing hypotheses might depend on the specific context.

(Manuscript received August 11, 1998;
revision accepted for publication June 28, 1999.)