

Do Conditional Hypotheses Target Rare Events?

Craig R. M. McKenzie, Victor S. Ferreira, Laurie A. Mikkelsen,
Kristine J. McDermott, and Ryan P. Skrable

University of California, San Diego

When testing hypotheses, rare or unexpected observations are normatively more informative than common observations, and recent studies have shown that participants' behavior reflects this principle. Research has also shown that, when asked to test conditional hypotheses ("If X , then Y ") that are abstract or unfamiliar, participants overwhelmingly consider a supporting observation mentioned in the hypothesis ($X \& Y$) to be more informative than a supporting observation not mentioned ($\sim X \& \sim Y$). These two empirical findings would mesh well if conditional hypotheses tend to be phrased in terms of rare, rather than common, events. Six experiments are reported indicating that people do have a tendency—often a very strong one—to phrase conditional hypotheses in terms of rare events. Thus, observations mentioned in conditional hypotheses might generally be considered highly informative because they usually *are* highly informative. © 2001

Academic Press

In the context of Bayesian hypothesis testing, data are informative to the extent that they are unexpected, or rare (e.g., Horwich, 1982; Howson & Urbach, 1989). Consider testing a forecaster's claim of being able to predict the weather in San Diego, where rainy days are rare. Which outcome would leave you more convinced of the forecaster's ability, a correct prediction of rain or a correct prediction of no rain? The correct prediction of rain should impress you more.¹

This research was supported by National Science Foundation Grants SBR-9515030 and SES-0079615. The authors thank Mike Ziolkowski and two anonymous reviewers for helpful comments on earlier drafts. Some of the results were presented at the 2000 Behavioral Decision Research in Management Conference, Tucson, AZ.

Address correspondence and reprint requests to Craig McKenzie, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109. E-mail: cmckenzie@ucsd.edu.

¹ To illustrate the normative principle, consider testing the forecaster's claim that he or she can predict the weather: "If I predict rain, then it will rain," which we take to mean that the forecaster can correctly predict the days it will and will not rain. Does a correct prediction of rain or a correct prediction of no rain constitute stronger evidence for his or her claim? For the sake of concreteness, assume that the base rate of rain is 5%, or $p(\text{rain}) = .05$ and $p(\text{no rain}) = .95$. Assume further

This normative principle is routinely exploited by researchers, who strive to correctly predict rare or unusual events because, intuitively, such results constitute strong evidence for the hypothesis or theory being tested.

The notion that people are sensitive to the rarity of data has important implications for how (and how well) they intuitively test hypotheses. Hypothesis-testing tasks often involve testing conditional hypotheses, “If X_1 , then Y_1 ,” where each variable, X and Y , has two levels (X_1 and X_2 , Y_1 and Y_2). Reasoning about conditionals is fundamental to human thinking (e.g., it is necessary for understanding causality and free will; see Sanford, 1989) and, accordingly, has been an active topic of research for the past four decades (for reviews, see Evans, Newstead, & Byrne, 1993; Oaksford & Chater, 1994; Wason & Johnson-Laird, 1972). Recently, McKenzie and Mikkelsen (2000) had participants test hypotheses of the form “If X_1 , then Y_1 ” and asked them whether an X_1 & Y_1 observation or an X_2 & Y_2 observation—both of which support the hypothesis—provided stronger support. Decades of hypothesis-testing research have shown that participants overwhelmingly prefer confirming observations named in the hypothesis, or X_1 & Y_1 observations (e.g., Evans, 1989; Fischhoff & Beyth-Marom, 1983; Johnson-Laird & Tagart, 1969; Klayman & Ha, 1987; McKenzie, 1994). McKenzie and Mikkelsen also found this—but only when the hypothesis regarded unfamiliar variables and there was no information regarding the rarity of the observations. When participants were told that X_2 and Y_2 were

that the forecaster predicts the two events as often as they occur, or that $p(\text{predict rain}) = .05$ and $p(\text{predict no rain}) = .95$. (We relax these assumptions below.) From a Bayesian viewpoint, data (D) support a hypothesis (H1) over its alternative (H2) to the extent that the likelihood ratio, $p(D|H1)/p(D|H2)$, exceeds 1. Let H1 be that the forecaster is perfect: Every time he or she predicts rain, it does rain, and every time he or she predicts no rain, it does not. Let H2 be that the forecaster achieves only chance-level performance. Let D correspond to the conjunction of a prediction and an outcome (e.g., correctly predicting rain). If H1 is true, then the forecaster will correctly predict rain on 5% of the days (every day it rains), or $p(\text{correctly predict rain} | \text{perfect performance}) = .05$. Similarly, $p(\text{correctly predict no rain} | \text{perfect performance}) = .95$. If H2 is true (i.e., predictions and outcomes are independent), the probability of correctly predicting rain is $.05^2$, or $.0025$. Similarly, the probability of correctly predicting no rain given chance performance is $.95^2$, or $.9025$. Now we can compare the likelihood ratios for the correct predictions of rain vs. no rain:

$$\frac{p(\text{correctly predict rain} | \text{perfect performance})}{p(\text{correctly predict rain} | \text{chance performance})} = \frac{0.0500}{0.0025} = 20,$$

whereas

$$\frac{p(\text{correctly predict no rain} | \text{perfect performance})}{p(\text{correctly predict no rain} | \text{chance performance})} = \frac{0.9500}{0.9025} = 1.053.$$

The first likelihood ratio is 19 times greater than the second one, showing that correctly predicting rain is much more supportive of the forecaster’s claim than correctly predicting no rain. Furthermore, the second likelihood ratio barely exceeds 1, indicating that correctly predicting no rain, a common event, is virtually uninformative. Generally, given two dichotomous variables, X and Y , when testing “If X_1 , then Y_1 ,” an X_1 & Y_1 observation will be more informative than an X_2 & Y_2 observation whenever $p(X_1) < 1 - p(Y_1)$ (Horwich, 1982; Mackie, 1963; McKenzie & Mikkelsen, 2000).

rare relative to X_1 and Y_1 , or when this could be inferred because familiar, concrete variables were used, participants were more likely to select the X_2 & Y_2 observation as more supportive. The combination of familiar variables and a “reminder” that X_2 and Y_2 were rare led participants to correctly select the X_2 & Y_2 observation more often than the X_1 & Y_1 observation, even though they were testing “If X_1 , then Y_1 .”

Another example of participants’ sensitivity to rarity when testing conditional hypotheses comes from Wason’s selection (or four-card) task. In this task, participants test the rule “If X_1 , then Y_1 ” and are presented with four cards: X_1 , X_2 , Y_1 , and Y_2 . On the back of the first two cards is either Y_1 or Y_2 , and on the back of the last two is either X_1 or X_2 . 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 X_1 and Y_2 cards should be turned over (A and 7 in the example). A common finding, though, is that participants want to turn over the cards mentioned in the rule: the X_1 and Y_1 cards (e.g., Wason, 1966, 1968). This has been widely regarded as a classic example of irrational behavior. However, Oaksford and Chater (1994, 1996; see also Nickerson, 1996) have argued convincingly that, from a Bayesian perspective, the X_1 and Y_1 cards are the most informative for testing the rule—if one assumes that X_1 and Y_1 are rare relative to X_2 and Y_2 . As predicted by this account, participants presented with a reduced array selection task, where only the Y_1 and Y_2 cards are present, were more likely to select the Y_2 card as it became rarer (Oaksford, Chater, Grainger, & Larkin, 1997).

Given these results, people’s hypothesis-testing behavior can be characterized as follows: When testing “If X_1 , then Y_1 ,” the default strategy is to deem an X_1 & Y_1 observation more supportive than an X_2 & Y_2 observation, but the strategy changes appropriately when information (implicit or explicit) regarding the rarity of data is available.

However, an important question remains unanswered: *Why* is the default strategy to judge confirming observations mentioned in the conditional hypothesis most supportive? One possibility is that simply mentioning the X_1 & Y_1 observation makes it more salient. However, McKenzie and Mikkelsen (2000; see also Oaksford & Chater, 1994) speculated that mentioned observations are judged more supportive because conditional hypotheses tend to be phrased in terms of rare, rather than common, events. If so, then a confirming observation mentioned in the hypothesis would be *normatively* more supportive than an unmentioned confirming observation. In the absence of information indicating otherwise, behaving as though the mentioned observation is rare, and therefore highly informative, might make good normative sense.

To be more concrete, consider two people applying to a selective college, where admission is rare. One applicant currently attends a private preparatory school (relatively rare) and is admitted, and the other attends a public school (relatively common) and is rejected. If asked to explain why the two applicants

experienced different outcomes, would people phrase their explanations, or tentative hypotheses, in terms of "If applicants attend private schools, then they will be admitted" or in terms of "If applicants attend public schools, then they will be rejected"? The former phrasing targets rare events and the latter common events. Targeting rare events makes a mentioned observation (an admitted applicant who attended private school) more supportive of the hypothesis than an unmentioned observation (a rejected applicant who attended public school). In contrast, targeting common events makes a mentioned observation *less* supportive of the hypothesis than an unmentioned observation.

Interestingly, research has shown that people's counterfactual statements, which are often conditionals ("If only . . . , then . . ."), mention common, not rare, events (for a review, see Roese, 1997). Consider, for example, Kahneman and Tversky's (1982) scenario, in which a Mr. Jones leaves his office at the usual time, but takes an unusual route home and is killed in a car accident. When asked to complete "If only . . ." stems, participants often mentioned the route in the antecedent (e.g., "If only Mr. Jones had taken his usual route home, then he would still be alive"). When the scenario was changed such that Mr. Jones left his office unusually early, but took the usual route home, participants were more likely to mention the time in the antecedent ("If only he had left at his usual time . . ."). Note that although it is the unusual preceding event that is targeted (i.e., the unusual route in the first case and the unusual departure time in the second), it is changed back into its common form (e.g., the *usual* route) in the counterfactual statement. Counterfactuals involve "the conversion of exceptional antecedents back into *normal* antecedents" (Roese, 1997, p. 137; emphasis added). Thus, what gets explicitly mentioned in "If only . . ." counterfactual statements is what is common, not rare.

However, how people phrase counterfactuals, which regard what might have occurred (but did not), may not tell us much about how people phrase conditional hypotheses, which are tentative statements about what is, or will be, true. Indeed, studies examining how people explain events, an area of research more closely related to the current topic, lend some credence to the idea that conditional hypotheses are phrased in terms of rare events. For example, attribution theorists have demonstrated that rare (or unexpected) events trigger a search for explanation (see Weiner, 1985, for a review). They have also shown that, when asked to explain a particular outcome, people point to rare (or abnormal) preceding events (Hilton & Slugoski, 1986; see also Einhorn & Hogarth, 1986; Klayman and Ha, 1987; Mackie, 1974). Though relevant, these findings are nonetheless inconclusive with respect to the phrasing of conditional hypotheses. For example, in the college application scenario described earlier, the student being admitted might trigger a search for explanation, but that does not necessitate phrasing the relation in terms of "If private school, then admitted" rather than "If public school, then rejected." Similarly, if asked to explain why the first applicant in our above example was admitted, people might point to the private schooling, but note that our example did not ask why a particular outcome occurred, but why the applicants experienced different outcomes.

Miller, Taylor, and Buck (1991) conducted studies even more closely related to the current topic. Testing ideas inspired by norm theory (Kahneman & Miller, 1986), Miller *et al.* asked participants to explain why men and women visited their doctors differentially often. The authors manipulated whether the reference group was elementary school teachers or college professors. Women are more common in the first group, while men are more common in the second (and the majority gender is viewed as prototypical of each group as well). Participants tended to mention men when asked to explain why male and female elementary school teachers behaved differently (“men tend to . . .”), and they tended to mention women when asked to explain why male and female college professors behaved differently (“women are more likely to . . .”). That is, participants’ spontaneous explanations targeted attributes of the nonprototypical, or rare, category members.

Miller *et al.*’s (1991) results make more plausible the idea that conditional hypotheses target rare events, given that such hypotheses are often uttered in the context of explaining, or at least describing, how variables are related. However, establishing a common link between how conditional hypotheses are phrased and how they are tested has important theoretical implications for lay hypothesis testing, and, accordingly, it is crucial to examine the phrasing of conditional hypotheses directly. Hence, this article specifically examines whether conditional hypotheses are phrased in terms of rare events.

EXPERIMENT 1

In our first experiment, participants were presented with two scenarios in which various people, differing in several respects, experienced different outcomes. In each scenario, there were only two outcomes and two preceding events. One outcome and one preceding event was rare and one was common. Based on the information in each scenario, participants reported the most plausible explanation as to why the participants experienced different outcomes in terms of a conditional hypothesis. At issue was whether the conditional hypotheses targeted the rare or common events.

Method

Participants were 33 students in psychology courses at University of California, San Diego (UCSD) and received partial course credit for participation. They were given a questionnaire in a laboratory setting. There were two groups of participants, differing in terms of which outcome/preceding event was rare in the two hypothetical scenarios. In the first scenario for one group, 5 high school seniors had applied to a prestigious college that receives many applications. Participants were told that few applicants are accepted; most are rejected. Consistent with this, only 1 of the 5 applicants had been accepted and 4 had been rejected. Next to each applicant was a rating from the college in 5 catego-

ries. In each category, 4 applicants received “average” and 1 “high,” as shown below:

	GPA	SAT Scores	Letters of Recomm.	Interview	Extra-curricular	APPLICATION OUTCOME
Alice	Average	Average	High	Average	Average	REJECTED
Bill	Average	High	Average	Average	Average	ACCEPTED
Cindy	Average	Average	Average	Average	High	REJECTED
Dennis	Average	Average	Average	High	Average	REJECTED
Emily	High	Average	Average	Average	Average	REJECTED

Participants were then asked:

Based on the information above, *what explanation of the different outcomes seems most plausible to you?*

IF APPLICANTS _____

THEN _____

The ratings in only one category (SAT scores) covaried perfectly with the outcome. We took for granted that participants would notice this (and they did). At issue was how they filled out the conditional hypothesis. Note that there are two equally legitimate ways to do so. Participants could write “If applicants have average SATs, then they will be rejected” or “If applicants have high SATs, then they will be accepted.” The former phrasing targets the common events and the latter the rare events.

The second group received the above scenario, but the college was a local one that did not receive many applications. Few applications were rejected; most were accepted. Accordingly, one of the 5 applicants was rejected and 4 were accepted. “Accepted” and “rejected” in the above scenario were simply reversed, and the word “high” was replaced by “low.” Everything else was the same. Thus, acceptance was the rare outcome for the first group, and rejection was the rare outcome for the second group.

Both groups also read a job promotion scenario in which the promotion was either rarely granted or rarely denied. The college application and job promotion scenarios were functionally identical; only the employee, category, and outcome labels were changed, the unusual outcome happened to the fourth, rather than the second, person, and the antecedent part of the conditional hypothesis was “If employees . . .”. The group presented with the college scenario with a low (high) acceptance rate was presented with the job scenario with a high (low) promotion rate.

Results and Discussion

When presented with the college scenario in which acceptance was rare, all 16 participants targeted acceptance. That is, 16 participants filled in the conditional hypothesis with (essentially) “If applicants have high SAT scores, then they will be accepted,” and *none* responded with “If applicants have average SAT scores, then they will be rejected.” In contrast, when acceptance was

common, only 5 of the 17 participants (29.4%) targeted acceptance. The difference in targeting acceptance between the two scenarios is significant ($p < .001$; p -values correspond to Fisher's exact tests unless noted otherwise).

When presented with the job scenario in which promotions were rare, 15 out of 17 participants (88.2%) participants targeted promotion. When promotions were common, *none* of the 16 participants targeted promotion ($p < .001$).

These results indicate that participants spontaneously target rare rather than common events when asked to provide a conditional explanation, or tentative hypothesis, as to what caused different outcomes to occur. Preference for targeting particular events reversed depending on whether they were rare or common.

However, one shortcoming of the experiment is that the common antecedent was always scoring "average" whereas the rare antecedent was always scoring "high" or "low." There might be linguistic reasons (besides issues of rarity) to refrain from phrasing explanations or hypotheses in terms of "average" occurrences, perhaps because they are in the middle of a dimension. This confound was eliminated in our next experiment.

EXPERIMENT 2

Method

Participants were 68 UCSD students drawn from the same population as in Experiment 1. The materials were identical to those of Experiment 1 except that "average" was replaced with "high" ("low") whenever "low" ("high") was the rare category value. Thus, the antecedent for both rare and common outcomes was either "high" or "low."

Results and Discussion

When presented with the college scenario in which acceptance was rare, 30 out of 34 participants (88.2%) targeted acceptance. However, when acceptance was common, only 12 out of 34 participants (35.3%) targeted acceptance ($p < .001$). When presented with the scenario in which promotions were rare, 30 out of 34 (88.2%) participants targeted promotion, and when promotions were common, only 6 out of 34 (17.6%) targeted promotion ($p < .001$).

These results are slightly weaker than in Experiment 1, suggesting that wording of the antecedent had some effect. Nonetheless, the results still show a strong tendency to target rare events. As in Experiment 1, preference for phrasing the conditional hypothesis in terms of particular events reversed depending on whether the events were rare or common.

EXPERIMENT 3

In Experiments 1 and 2, participants were explicitly told which outcomes were rare. In this experiment, we used two scenarios in which participants

could apply their real-world knowledge about rarity. We wanted to see if they would spontaneously target the rare events when not told explicitly about rarity.

Method

Participants were 22 UCSD students recruited from the same population as the earlier experiments, and they were presented with two scenarios similar to those used earlier. One involved applicants being accepted vs. rejected by UCSD. We assumed (and later verified) that our participants knew that acceptance to UCSD is rare. No mention was made of the acceptance rate, but note that the rare event has a positive valence. (We controlled for outcome valence between scenarios.) There were again 5 applicants in the scenario, but only one was accepted. The same 5 categories were used as before, but the possible scores in each category were “above” and “below,” referring to whether the student scored above a high threshold. In terms of the scenario example presented above (under Experiment 1), “below” replaced “average” and “above” replaced “high.” Telling participants that thresholds were set high made it apparent that “above” scores were rare and “below” scores were common.

The second scenario regarded freshman (first-year) UCSD students completing their first year vs. dropping out. Dropping out during freshman year is rare, and participants were assumed to know this. Note that the rare event in this scenario has a negative valence. The scenario was similar to the acceptance scenario, with the following differences. The outcome category label was “first-year outcome,” “dropped out” replaced “accepted” and “completed” replaced “rejected.” The 5 categories were medical problems, family problems, social problems, financial problems, and academic problems. The ratings in each category were either “moderate” or “extreme.” Participants were told that “extreme” meant that the student experienced unusual difficulty in that category, whereas “moderate” meant that the student experienced what one might expect. “Moderate” replaced “below” in the application scenario, and “extreme” replaced “above.” The names were different, and the rare event occurred for the fourth, rather than second, student. “Financial problems” covaried perfectly with the outcome, and the antecedent “If freshmen . . .” replaced “If applicants . . .”.

On the last page of the questionnaire, participants estimated the percentage of applicants to UCSD that are accepted and the percentage of first-year UCSD students that drop out.

Results and Discussion

When presented with the UCSD acceptance scenario, all 22 participants targeted acceptance ($p < .001$, binomial test). Their mean estimate of the acceptance rate was 33.7%, indicating that they perceived acceptance as rarer than rejection. When presented with the UCSD drop-out scenario, 21 out of the 22 participants targeted dropping out ($p < .001$, binomial test). The mean

estimate of the first-year drop-out rate was 11.3%, showing that participants perceived the outcome as rare.

These results indicate that explicitly telling participants that a certain outcome is rare is not necessary to trigger the targeting of rare events. Implicit knowledge regarding rarity was sufficient. This was true for both the positive and the negative rare outcomes.

However, participants were provided with information regarding 5 students, only 1 of whom experienced the rare outcome, and, arguably, were therefore given information about rarity beyond their real-world knowledge. Our next experiment eliminated this source of information as well by presenting only two people in a scenario, each experiencing one of the two outcomes. Furthermore, as an even stronger test of our thesis, some of the antecedent variables used were continuous (e.g., GPA values) rather than dichotomous (e.g., high vs. low GPA). In this way, participants had to categorize particular values as common or rare based entirely on their real-world knowledge.

EXPERIMENT 4

Method

Participants were 166 UCSD students recruited from the same population as the earlier experiments, and they were presented with an admission and a drop-out scenario similar to those in Experiment 3. Half of the participants received the following version of the admission scenario:

The two people listed below, Steve and Melissa, are seniors in high school and both applied for admission to UCSD. Next to each of the applicants are their scores and ratings in three categories. On the far right is listed whether each applicant was accepted or rejected. As you can see, Steve was accepted and Melissa was rejected.

	H.S. GPA	SAT Scores	Extra- curricular	APPLICATION OUTCOME
Steve	4.0	1240	Average	ACCEPTED
Melissa	3.7	1390	Average	REJECTED

Based on the information above, *what explanation of the different outcomes seems most plausible to you?*

IF APPLICANTS _____

THEN THEY WILL BE _____

Because there are only two applicants, there is no explicit information regarding which events are rare. In this scenario, we assumed that participants would focus on high school GPA because the higher GPA led to acceptance (whereas the higher SAT scores led to rejection). The value of 3.7 was used because this was the mean estimate reported by 77 participants in a pilot study asking for (among other things) the average GPA of high school seniors applying to UCSD. Thus, the low value is common and the high value is relatively rare (in that few applicants have GPA's at or above that value). We predicted that participants would fill out the statement with "If applicants have a high GPA, then

they will be accepted” rather than “If applicants have a mediocre GPA, then they will be rejected.” Note that the GPA dimension is continuous, so participants must categorize a particular GPA value as rare or common based on their experience outside the laboratory.

The other half of the participants received the same scenario except that Steve was rejected and Melissa was accepted, leading participants to focus on the SAT scores dimension. Like the GPA values, the SAT scores are continuous and the low value corresponded to the mean estimate of participants in the pilot study. We predicted that participants would fill out the statement with “If applicants have high SAT scores, then they will be accepted” rather than “If applicants have mediocre SAT scores, then they will be rejected.” Again, this prediction is based on participants’ real-world knowledge regarding rarity.

In addition, half the participants received the following version of the drop-out scenario:

UCSD is trying to understand better why some freshmen complete their first year while others drop out. Frank and Jim, listed below, were freshman at UCSD last year. Next to each of them are their scores and ratings in three categories. On the far right is whether each completed his first year or dropped out. As you can see, Frank dropped out and Jim completed his first year.

	1st-year GPA	Financial problems?	Health problems?	FIRST-YEAR OUTCOME
Frank	2.7	Severe	None	DROPPED OUT
Jim	1.8	Mild	None	COMPLETED YEAR

Based on the information above, *what explanation of the different outcomes seems most plausible to you?*

IF FRESHMEN _____

THEN THEY WILL _____

This scenario was expected to lead participants to focus on financial problems. In the pilot study, 88.3% of the participants chose “mild” when asked to select from among “none,” “mild,” and “severe” financial problems as the most common situation for first-year students. (This antecedent variable was not continuous simply because we could think of only one plausible quantified continuous variable—GPA—for this scenario.) We therefore predicted that participants would fill out the statement with “If freshman have severe financial problems, then they will drop out” rather than “If freshman have mild financial problems, then they will complete their first year.”

The other half of the participants received the same drop-out scenario except that Frank completed his year and Jim dropped out, leading participants to focus on 1st-year GPA. The 2.8 value was the mean estimate in our pilot study, and therefore the 1.7 value is uncommonly low. We predicted that participants would fill out the statement with “If freshman have a low GPA, then they will drop out” rather than “If freshman have an average GPA, then they will complete their first year.”

All participants responded to one version of the admission scenario and one version of the drop-out scenario. We controlled for scenario order.

Results

When presented with the UCSD acceptance scenario in which high school GPA was salient, 76 out of 83 participants (91.6%) targeted the rare events ($p < .001$, binomial test; one response was uncodable). That is, 76 responded with (essentially) “If applicants have a high GPA, then they will be accepted” and only 7 responded with “If applicants have a mediocre GPA, then they will be rejected.” In the other version of the scenario which highlighted SAT scores as the antecedent, 76 out of 82 participants (92.7%) targeted the rare events ($p < .001$, binomial test).

For the drop-out scenario in which finances were the salient antecedent, 76 out of 84 participants (90.5%) targeted the rare events ($p < .001$, binomial test). However, there was only a weak effect for the drop-out scenario where GPA was the salient antecedent. Only 48 out of 81 participants (59.3%) targeted the rare events (“If low GPA, then drop out”) over the common events (“If mediocre GPA, then complete year”; $p = .06$, one-tailed binomial test; one response was uncodable).

Discussion

This experiment provided strong evidence that real-world knowledge regarding rarity influences how conditional hypotheses are phrased. Rare events were targeted even though there was no explicit information in the scenarios regarding rarity, and most of the antecedent variables were continuous, forcing participants to categorize particular values as common or rare based entirely on their real world knowledge.

There was, however, one partial exception. For the drop-out scenario, where the salient antecedent was freshman GPA, participants exhibited only a weak preference for targeting the rare events. We believe that “markedness,” a linguistic notion, accounts for the weakened effect.² In the context of our task, we mean by a markedness effect that participants exhibit a tendency to phrase conditional hypotheses in terms of “If X is high (or large), then . . .” rather than “If X is low (or small), then . . .”, at least when the antecedent variable has a natural ordering. (The “high” end of the variable is the *unmarked* end.) The only condition of the four that resulted in a weak effect was the one in which targeting the rare events required stating “If GPA is *low*, then students will drop out.” Targeting the rare events in the other three conditions required stating “If X is *high*, then . . .”. This finding caught us by surprise and raised an obvious question: To what extent could markedness, rather than rarity, explain all of our findings?

Fortunately, we were able to examine this question thoroughly with the

² For many continuous dimensions, the adjectives used to describe the polar ends are linguistically asymmetrical. Consider the dimension “length,” which has polar ends of “long” and “short.” “Long” is the “unmarked” adjective in this case because only it is used in a neutral way to ask a question (“How long is the table?”), to identify the dimension (“length”), or to quantify an expression (“6 feet long”; e.g., Clark, 1969).

current data. In Experiments 1 and 2, we (fortuitously) manipulated markedness and rarity independently when we manipulated which outcome was rare in each scenario. For example, when admission was rare in these experiments, targeting the rare events led to reporting “If GRE scores are *high*, then applicants will be admitted,” whereas when rejection was rare, targeting the rare events led to reporting “If GRE scores are *low*, then applicants will be rejected.” The same is true for the promotion scenario. The results were almost identical across the two scenarios and experiments, so we pooled the data. We examined the number of responses that targeted the rare vs. common events and that targeted the unmarked vs. marked antecedent. Because rarity and markedness were orthogonal to each other, we were able to examine the independent influence of each. The results are shown in the top half of Table 1. The marginal frequencies are most important. For example, the first column marginal frequency shows that 169 of the 202 total responses (83.7%) targeted rare over common events, regardless of markedness. The first row marginal frequency shows that 114 of the 202 responses (56.4%) targeted the unmarked over the marked antecedent, regardless of rarity. Thus, rarity, not markedness, accounts almost entirely for the results of Experiments 1 and 2.

One shortcoming of the above analysis is that markedness and outcome valence were confounded in Experiments 1 and 2. Rare negative outcomes (e.g., being rejected) were always associated with “If *X* is low, then . . .”, whereas rare positive outcomes (e.g., being admitted) were always associated with “If *X* is high, then . . .”. If there is a tendency to target negative outcomes, then the influence of markedness might be underestimated by these two experiments. In Experiment 4, however, we (again fortuitously) manipulated rarity and markedness while holding outcome valence constant. In the drop-out scenario where financial problems were the salient antecedent variable, targeting the rare events required reporting “If financial problems are severe (i.e., high), then students will drop out.” However, for the drop-out scenario where GPA was salient, targeting the rare events resulted in “If GPA is *low*, then students

TABLE 1

Frequency of Responses Targeting Rare vs. Common Events and Unmarked vs. Marked Antecedents in Experiments 1 and 2 (Combined) and in Experiment 4 (Drop-out Scenario)

Antecedent	Events		(Total)
	Rare	Common	
	Experiments 1 and 2 (combined)		
Unmarked	91	23	(114)
Marked	78	10	(88)
(Total)	(169)	(33)	(202)
	Experiment 4 (drop-out scenario)		
Unmarked	76	33	(109)
Marked	48	8	(56)
(Total)	(124)	(41)	(165)

will drop out.” Thus, we were able to compare the influence of rarity and markedness, holding constant the valence of the rare outcome (negative, in this case). The results are shown in the bottom half of Table 1. The first column marginal shows that 124 of the 165 total responses (75.1%) targeted the rare over common events, independent of markedness. The first row marginal shows that 109 of the 165 responses (66.1%) targeted the unmarked antecedent, independent of rarity. Thus, there was again evidence that rarity and markedness independently affected how conditional hypotheses were phrased, and the effect of rarity was stronger than that of markedness. Furthermore, it is reasonable that, relative to Experiments 1 and 2, the effect of rarity was slightly weaker and the effect of markedness was slightly stronger because there was no explicit information regarding rarity in Experiment 4. To the extent that rarity is not salient, it makes sense that other factors, such as markedness, will come into play.

In short, independent of markedness, there were effects of rarity, and the effects were stronger than those of markedness. Though markedness does appear to play a role, it does not explain the results of our experiments. We found the weak effect of rarity in one condition in this experiment because rarity and markedness were competing with each other and we had eliminated explicit information regarding rarity. Even then, however, there was a weak preference for targeting the rare events. Furthermore, it is worth mentioning that markedness is only relevant when the antecedent variable has a natural ordering. When nominal-level variables (e.g., gender, handedness, group affiliation) are used to explain outcomes, markedness will not apply.

Finally, we want to reiterate that Experiment 4 was a strong test of whether conditional hypotheses are phrased in terms of rare events. No explicit information regarding rarity was provided in the materials and 3 of the 4 antecedent variables were scaled, meaning that participants had to spontaneously categorize particular values as unusually low, average (i.e., common), or unusually high. Rarity appears to be a powerful factor in how conditional hypotheses are phrased.

EXPERIMENT 5

In all of our experiments thus far, participants were asked to provide the best *explanation* of why the different outcomes occurred. Is there something special about asking for explanations that leads to targeting rare events? This experiment asked some participants to *describe* rather than explain the outcomes using a conditional hypothesis.

Method

There were 27 participants from the population described earlier. All were presented with a scenario similar to the college acceptance vs. rejection scenario used in Experiment 2, but with only two antecedents, SAT scores and GPA. Both antecedents were characterized as either “above” or “below” a high threshold,

making “above” rare and “below” common. They were told that 5 high school seniors had applied to UCSD, where few applicants are accepted. Accordingly, only one of the five applicants was accepted. The specific information provided is shown below:

	SAT Scores	GPA	APPLICATION OUTCOME
Alice	Below	Below	REJECTED
Bill	Above	Below	ACCEPTED
Cindy	Below	Below	REJECTED
Dennis	Below	Below	REJECTED
Emily	Below	Above	REJECTED

As before, half of the participants were asked to explain the outcomes: “[W]hat explanation of the different outcomes seems most plausible to you?” The other half of the participants were asked to *describe* the events: “[W]hat do you think is the best way to describe how the different outcomes occurred?” Both groups then completed the statement, “If applicants _____, then _____”.

Results and Discussion

The rare events (high SAT scores and acceptance) were targeted by 12 of the 13 participants asked for an explanation and by 13 of the 14 participants asked for a description. The group percentages (92.3 and 92.9) are not different from each other, and both significantly target rare over common events (both $p_s < .01$, one-tailed binomial tests). Targeting rare events in conditional hypotheses was not limited to explanations. Whether asked to describe or to explain the events, participants targeted the rare events.

EXPERIMENT 6

The strong tendency to mention rare events in Experiments 1–5 raises the question of whether there is something more general than the phrasing of conditional hypotheses driving the effect. Perhaps there is something about our scenarios that leads participants to mention rare events in general and our results say little about how conditional hypotheses per se are phrased. This would imply that participants mention rare events regardless of the task presented to them.

When might one expect *common* events to be mentioned? Consider how people typically *summarize* variables. The primary means for summary is to report a central tendency (mean, median, or mode), capturing what is common, or average. In this experiment, we asked participants, in addition to filling in a conditional hypothesis, to summarize each variable using a single phrase. We tried to constrain participants to choose between responding with, for example, “few applicants were accepted” versus “most applicants were rejected.” The former phrasing targets the rare event and the latter the common one. We expected participants to target rare events in their conditional hypotheses and to target common events in their summaries.

Method

There were 132 participants from the same population as the earlier studies. They were presented with a single scenario similar to that used in Experiment 5. In the current scenario, there were 10 (rather than 5) high school students applying to a college, and the rare level of each variable occurred for 2 (rather than 1) of the applicants. Half of the participants were presented with the “prestigious college” scenario in which high SAT scores, high GPA’s, and acceptance were all rare. The other half of the participants were presented with the “local college” scenario in which low SAT scores, low GPA’s, and rejection were all rare. Only SAT scores covaried perfectly with the application outcome.

In addition to filling in the conditional hypothesis, participants were asked to summarize, in a single phrase, each of the three variables. They were told to use the word “some” or “most” and not to use “and.” The instructions for the SAT scores read, “Please summarize in a single phrase the 10 applicants’ SAT scores.” Half of the participants seeing each scenario filled out the conditional hypothesis first, and half summarized the three variables first.

Results

Of the 132 participants, 7 failed to use a single phrase to describe at least 2 of the 3 variables and were eliminated from the analysis. Figure 1 shows the percentage of participants who targeted rare events depending on (a) task (conditional vs. summary), (b) which task came first, and (c) scenario (rarely accepted vs. rarely rejected). Generally, and most important, targeting rare events was relatively common in the conditional task (solid columns) and rare in the summary task (hatched columns). This was true for all four groups. (Participants were counted as targeting the rare events in their summaries if they did so for at least 2 of the 3 variables.)

The leftmost pair of columns corresponds to the group presented with the rarely accepted scenario and asked for the conditional hypothesis first. The

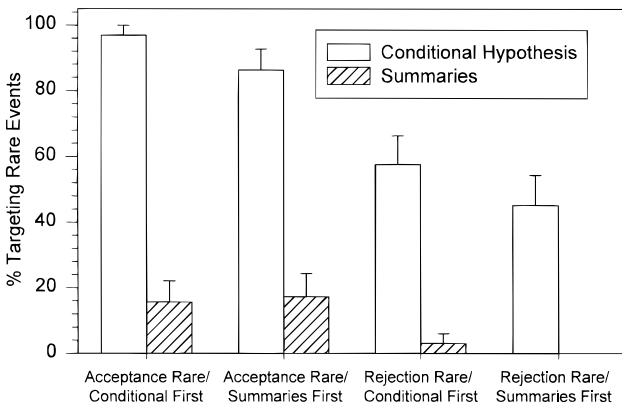


FIG. 1. Percentage of participants targeting the rare events as a function of (a) task (conditional hypothesis vs. summary), (b) which task came first, and (c) scenario (rarely accepted vs. rarely rejected). Standard error bars are shown.

figure shows that most participants filled in the conditional hypothesis with “If participants have high SAT scores, then they will be accepted,” which targets rare events, but most summarized the applicants’ SAT scores (for instance) with “most have low SAT scores,” which targets common events. The next pair of columns shows that asking for summaries first reduced slightly the number of conditional hypotheses targeting rare events for the same scenario. The right two pairs of columns show that the rarely accepted scenario reduced the tendency for conditional hypotheses to target rare events. Note that the number of summaries targeting rare events was also reduced. Finally, as with the other scenario, the tendency for conditional hypotheses to target rare events was reduced slightly when the summaries came first.

With the proportion of participants targeting rare events as the dependent measure, a 2 (scenario) \times 2 (task order) \times 2 (task) ANOVA, using the latter variable as a repeated measure, corroborated the clear patterns in Fig. 1. Most important was the main effect of task, with fewer participants targeting rare events in their summaries ($F(1,121) = 218.7, p < .001$). There was also a main effect of scenario, with fewer participants targeting rare events when rejection was rare ($F = 35.8, p < .001$). In addition, task and scenario interacted: Relative to the rarely accepted scenario, there was a larger reduction in targeting rare events in the rarely rejected scenario for the conditional task than for the summary task ($F = 8.9, p = .003$). This interaction appears due to a floor effect for the summary task.

Discussion

Participants were more likely to target rare events when filling in the conditional hypothesis than when asked to summarize each variable in a single phrase. These results show that our findings regarding conditional hypotheses are not due to something about the materials that generally leads to mentioning rare events. Whether rare or common events were mentioned depended on the specific task.

It was also found that participants were less likely to target rare events in conditional hypotheses for the rarely rejected scenario. This is consistent with our earlier finding regarding markedness. Indeed, there was even evidence of a markedness effect for the summaries: Participants were less likely to respond with, for example, “most applicants have *low* SAT scores” when low scores were common than they were to respond with “most applicants have *high* SAT scores” when high scores were common.

There was only a small (and nonsignificant) effect of task order. Providing summaries first, which led to targeting common events, slightly reduced the tendency to target rare events in the subsequent conditional hypotheses. Providing the conditional hypothesis first, however, seemed to have no effect on subsequent summaries. The lack of any large effect of task order indicates that the proclivities for targeting rare events in conditional hypotheses and for targeting common events in summaries are both strong.

GENERAL DISCUSSION

People have a tendency—often a very strong one—to phrase conditional hypotheses in terms of rare rather than common events. In Experiments 1 and 2, preference for targeting a given outcome reversed depending on whether the context made the outcome rare vs. common. In Experiments 3 and 4, where participants' real-world knowledge allowed them to infer which events were rare, they targeted rare events without being explicitly told about rarity. Experiment 5 demonstrated that conditional hypotheses targeted rare events regardless of whether participants were asked to describe how the different outcomes occurred or were asked to explain the different outcomes. Finally, Experiment 6 showed that the results of the Experiments 1–5 were not due to a general tendency to mention rare events: When asked to summarize each variable in a single phrase, participants virtually always mentioned the common events.

These findings are clearly consistent with the idea that confirming observations mentioned in conditional hypotheses are generally considered more supportive than unmentioned observations because such hypotheses tend to be phrased in terms of rare events. In other words, mentioned confirming observations generally *are* more supportive, and participants' behavior in the laboratory reflects this. We doubt that it is mere coincidence that (a) rare observations are normatively more informative than common ones (Horwich, 1982; Howson & Urbach, 1989), (b) people are sensitive to the rarity of observations (McKenzie & Amin, 2000; McKenzie & Mikkelsen, 2000; Oaksford & Chater, 1994; Oaksford *et al.*, 1997), (c) people generally consider observations mentioned in conditional hypotheses most informative (Evans, 1989; Fischhoff & Beyth-Marom, 1983; Johnson-Laird & Tagart, 1969; Klayman & Ha, 1987; McKenzie, 1994), and (d) people tend to phrase conditional hypotheses in terms of rare events. Instead, we believe that the constellation of these empirical findings and normative principle is evidence of an efficient, well-adapted approach to testing hypotheses. Our claim is not that lay hypothesis testing is optimal, but that it is an interesting and rich process that deserves more credit than it usually receives (for exceptions, see Friedrich, 1993; Klayman & Ha, 1987; McKenzie & Mikkelsen, 2000; Oaksford & Chater, 1994). Taking into account the environmental conditions under which people typically operate—together with normative principles that make sense given these conditions—can help explain why people behave as they do. What seems irrational and maladaptive in the laboratory might reflect adaptation to conditions outside the laboratory (see also, e.g., Hogarth, 1981; McKenzie, 1994).

We also conducted pilot studies, not reported here, that taught us a related lesson. We used highly abstract scenarios in some studies in an attempt to exert greater experimental control, and the clear preference for targeting the rare outcome (any outcome, actually) disappeared. For example, in one scenario, participants read about a computer game in which an object appeared on a screen and then moved either up or down. The object differed in terms of, for instance, shape and color, and we manipulated whether moving up or down was rare. We attempted such experiments because we could keep scenarios

identical except for which outcome was rare. (Such control was not possible in our reported experiments using more realistic materials. In Experiments 1 and 2, for example, we could not have the high SAT score lead to acceptance in one scenario and lead to rejection in the other—the latter scenario would make no sense—so we used a low SAT score in the latter case.) Not only did participants not have strong preferences for outcomes in these abstract scenarios, they were often confused by the task. It seems, then, that concrete scenarios need to be used, ideally ones in which participants have knowledge about the rarity of the events. This is consistent with results found by McKenzie and Mikkelsen (2000): Even when provided with explicit information about rarity, participants were more likely to correctly select rare observations as more supportive when they tested hypotheses about variables that were concrete and familiar rather than abstract and unfamiliar.

As for why conditional hypotheses target rare events, it might be related to the fact that the conjunction of rare events is normatively more informative with respect to the relation under consideration. In the case of the rarely accepted college scenario, stating that “If you have high SAT scores, then you will be accepted” says much about the relation between SAT scores and application decisions. It implies further that the speaker has knowledge about the relatively rare and valuable observation. In contrast, if low SAT scores and rejection are common, their combination will also be relatively common, such observations say little about a potential relation, and listeners are likely to already know about them.

In sum, our studies indicate that people have a strong tendency to spontaneously phrase conditional hypotheses in terms of rare, rather than common, events. This provides a simple explanation for participants’ default strategy of judging confirming observations mentioned in conditional hypotheses more supportive than unmentioned confirming observations: Mentioned observations usually *are* more supportive because they are rare. We believe that our findings make an important empirical contribution to the emerging picture of lay hypothesis testing as much more sensible—and interesting—than previously thought.

REFERENCES

- Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, **76**, 387–404.
- 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.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hillsdale, NJ: Erlbaum.
- Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis testing from a Bayesian perspective. *Psychological Review*, **90**, 239–260.
- Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review*, **100**, 298–319.
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, **93**, 75–88.

- Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, **90**, 197–217.
- Horwich, P. (1982). *Probability and evidence*. Cambridge, UK: Cambridge University Press.
- Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.
- Johnson-Laird, P. N., & Tagart, J. (1969). How implication is understood. *American Journal of Psychology*, **82**, 367–373.
- Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, **93**, 136–153.
- Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge, UK: Cambridge University Press.
- Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, **94**, 211–228.
- 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: 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., & Amin, M. B. (2000). *When wrong predictions provide more support than right ones*. Submitted for publication.
- McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin and Review*, **7**, 360–366.
- Miller, D. T., Taylor, B., & Buck, M. L. (1991). Gender gaps: Who needs to be explained? *Journal of Personality and Social Psychology*, **61**, 5–12.
- Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning*, **2**, 1–31.
- 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. (1996). Rational explanation of the selection task. *Psychological Review*, **103**, 381–391.
- 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, and Cognition*, **23**, 441–458.
- Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, **121**, 133–148.
- Sanford, D. H. (1989). *If P, then Q: Conditionals and the foundations of reasoning*. London: Routledge.
- Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–161). Harmondsworth, England: Penguin.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, **20**, 273–281.
- Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.
- Weiner, B. (1985). "Spontaneous" causal thinking. *Psychological Bulletin*, **97**, 74–84.

Received November 19, 1999