Memory & Cognition
2006, 34 (3), 577-588

Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias

CRAIG R. M. McKENZIE
University of California, San Diego, La Jolla, California

Researchers have recently pointed out that neither biased testing nor biased evaluation of hypotheses necessitates *confirmation bias*—defined here as systematic overconfidence in a focal hypothesis—but certain testing/evaluation combinations do. One such combination is (1) a tendency to ask about features that are either very likely or very unlikely under the focal hypothesis (*extremity bias*) and (2) a tendency to treat confirming and disconfirming answers as more similar in terms of their diagnosticity (or informativeness) than they really are. However, in previous research showing the second tendency, materials that are highly abstract and unfamiliar have been used. Two experiments demonstrated that using familiar materials led participants to distinguish much better between the differential diagnosticity of confirming and disconfirming answers. The conditions under which confirmation bias is a serious concern might be quite limited.

Although a wide variety of reasoning and decision-making errors have been reported (e.g., Evans, Newstead, & Byrne, 1993; Gilovich, Griffin, & Kahneman, 2002; Kahneman & Tversky, 2000), they are often disputed. For example, it has been argued that participants often construe tasks differently than do experimenters (Hilton, 1995; Schwarz, 1996), that some purported errors are consistent with an alternative normative standard (Anderson, 1990, 1991; Chase, Hertwig, & Gigerenzer, 1998; Gigerenzer, 1991, 1996; Gigerenzer, Todd, and the ABC Research Group, 1999; McKenzie, 2004a; McKenzie & Mikkelsen, in press; Oaksford & Chater, 1994, 1996, 2003; Sher & McKenzie, in press), and that many errors are limited to (or at least exacerbated by) the laboratory environment (Anderson, 1990, 1991; Klayman & Ha, 1987; McKenzie, 2003, 2004b; McKenzie & Mikkelsen, 2000; McKenzie & Nelson, 2003; Oaksford & Chater, 1994, 1996, 2003). This article takes the latter position on confirmation bias—one of the most widely cited errors in the reasoning literature—and argues that the conditions under which the bias occurs are more limited than previously thought.

Confirmation bias is usually said to occur in tasks that fall under the topic of *hypothesis development* (Klayman, 1995), which is concerned with how people put their ideas or hypotheses to test. For present purposes, this process will be seen in terms of three components: hypothesis generation, testing, and evaluation. In the context of a physician's diagnosing a patient, hypothesis generation occurs when the physician produces possible causes of the patient's symptoms. The testing component refers to the physician's deciding which questions to ask the patient or which tests to run in order to help determine whether a generated hypothesis is correct. Once answers or test results are known, the evaluation component occurs: How strongly do the results support or refute the hypothesis?

Hypothesis development is not limited to relatively formal settings, such as ones in which physicians diagnose illnesses or scientists test theories. People constantly engage in hypothesis development as a means of imposing structure on complex environments (Brehmer, 1980; McKenzie, 2004b). Given the importance of hypothesis development, it is not surprising that it has been the focus of much psychological research over the past several decades (for reviews, see Klayman, 1995; McKenzie, 2004b; Poletiek, 2001). However, widespread interest in the topic did not develop until Wason's (1960) article, which cast lay hypothesis development in a bad light. People were said to be prone to *confirmation bias* because they appeared to be trying to confirm the hypothesis that they were entertaining.

As has been noted by Fischhoff and Beyth-Marom (1983; see also Klayman, 1995; Klayman & Ha, 1987), *confirmation bias* has been used to describe many different phenomena. In this article, the bias will be said to occur if people behave in a way that leads to systematic overconfidence in a focal hypothesis (i.e., the favored hypothesis or the hypothesis being tested). This is consistent with recent views (e.g., Klayman, 1995; Nickerson, 1998)

This research was supported by National Science Foundation Grants SES-0079615 and SES-0242049, and some of the results were presented at the 42nd Bayesian Research Conference, Fullerton, CA, and at the 45th Annual Meeting of the Psychonomic Society, Minneapolis, MN. Michael Liersch, Jonathan Nelson, Mike Oaksford, and Fenna Poletiek provided valuable comments on earlier drafts. Correspondence should be addressed to C. R. M. McKenzie, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, MC 0109, La Jolla, CA 92093-0109 (e-mail: cmckenzie@ucsd.edu).

Note—This article was accepted by the previous editorial team, when Colin M. MacLeod was Editor.

and seems sufficiently broad. Presumably, one would not want to label behavior as *confirmatory* if it does not lead to more confidence in the focal hypothesis than is warranted. For example, people might tend to choose particular types of test, but unless this systematically results in overconfidence, there is no confirmation bias.

Confirmation bias has been said to result from errors at each stage of hypothesis development, especially the testing and evaluation stages (Klayman, 1995; Nickerson, 1998). Recently, however, it has been argued that earlier claims were exaggerated. Indeed, neither testing strategies nor evaluation strategies, by themselves, necessarily lead to confirmation bias, but certain *combinations* do (Klayman, 1995; Poletiek, 2001; Slowiaczek, Klayman, Sherman, & Skov, 1992). The realization that testing and evaluation strategies alone do not necessitate confirmation bias has reduced the conditions under which the bias is believed to occur.

This article examines the evaluation component of one of the three testing/evaluation combinations discussed by Klayman (1995). (The other two are saved for the General Discussion section.) It has been shown previously that participants are insufficiently sensitive to differences in the diagnosticity (or informativeness) of different answers to the same question (Slowiaczek et al., 1992). This evaluation tendency, combined with a certain testing tendency (described below), leads to systematic overconfidence in the focal hypothesis, or confirmation bias (Klayman, 1995; Slowiaczek et al., 1992). However, this insensitivity to differential diagnosticity has been shown with tasks in which abstract and unfamiliar materials have been used (e.g., sampling marbles from urns or choosing questions to ask imaginary creatures from distant planets). Two experiments, reported below, show that using familiar materials increases sensitivity substantially. Given that most real-world hypothesis development presumably involves familiar variables, these findings suggest that the conditions under which confirmation bias is expected to occur are even more constrained than has recently been argued.

The next two sections discuss why hypothesis testing and evaluation strategies, respectively, do not by themselves necessitate confirmation bias. The subsequent section shows how a particular testing/evaluation combination does lead to confirmation bias. Next, some reasons are provided as to why previous research in which this testing/evaluation combination has been examined might have led to overly pessimistic conclusions. The results from two experiments are then presented, and the implications for confirmation bias are discussed.

Hypothesis Testing: Choosing Questions to Ask

Consider Table 1, which shows the percentage of Gloms and Fizos—imaginary creatures on a distant planet—who possess each of eight different features (adapted from Skov & Sherman, 1986; Slowiaczek et al., 1992). For example, Feature 1 might be *drinks gasoline*, and the table shows that 50% of Gloms and 90% of Fizos do so. You have traveled to their planet, have encountered a creature, and want to know whether it is a Glom or a Fizo. There are only these two types of creatures, and they are equally numerous. You get to ask *yes*/*no* questions about their features (e.g., "Do you drink gasoline?"). If you could only ask about a limited number of features, which would you prefer?

Table 1
Percentage of Gloms and Fizos Possessing Each of Eight Features and the Expected Absolute Log Likelihood Ratio, $E(|\,\mathrm{LLR}\,|)$, Associated with Asking About Each Feature

Feature	Gloms	Fizos	$E(\|\,\mathrm{LLR}\,\|)$
1	50	90	1.3
2	50	60	0.3
3	50	10	1.3
4	90	50	1.3
5	50	2	1.9
6	10	0.1	0.5
7	90	99.9	0.5
8	0.1	10	0.5

Choosing which questions you ought to ask is complicated by the fact that you do not know which answer you will receive, and different answers might be differentially informative. If Feature 1 were asked about, a *yes* answer would favor the Fizo hypothesis (because more Fizos than Gloms have this feature). The Bayesian odds that the creature is a Fizo are $.5/.5 \times .9/.5 = .45/.25$. The first ratio is the prior odds, $p(\text{Fizo})/p(\text{Glom})$, or the odds that the creature is a Fizo before asking the question. The second ratio is the likelihood ratio, $p(\text{Feature 1}\,|\,\text{Fizo})/p(\text{Feature 1}\,|\,\text{Glom})$, which captures how diagnostic the received answer is. The final ratio represents the posterior odds that the creature is a Fizo, rather than a Glom, given the *yes* answer, or $p(\text{Fizo}\,|\,\text{Feature 1})/p(\text{Glom}\,|\,\text{Feature 1})$. These odds are almost 2 to 1. The probability that the creature is a Fizo after the *yes* answer is $.45/(.45 + .25) = .64$.

What if the creature had answered *no*? Such an answer would favor the Glom hypothesis, and the normative odds that the creature is a Glom are $.5/.5 \times .5/.1 = .25/.05$, or 5 to 1, and the probability is $.25/(.25 + .05) = .83$.

Note that the two answers are not equally diagnostic, or informative. The *yes* answer changed confidence in the identity of the creature from 50% to 64%, whereas the *no* answer changed confidence from 50% to 83%. The *no* answer is more diagnostic.

Because different answers to the same question can be differentially diagnostic, you must consider each potential answer's diagnosticity and the probability of receiving each answer when judging the usefulness of a question. One way to do so is to calculate a question's *expected* diagnosticity, weighting each of the answers' diagnosticities by the probability of receiving the answer, which depends on both the prior probability of the hypotheses and the frequency of the feature being asked about (Bassok & Trope, 1984; Klayman & Ha, 1987; Skov & Sherman, 1986; Slowiaczek et al., 1992; Trope & Bassok, 1982). In the case of Feature 1, a *yes* answer occurs 70% of the time, because 50% of the creatures are Gloms, 50% of whom will answer *yes*, and 50% are Fizos, 90% of whom will answer *yes*. Similarly, *no* will occur 30% of the time. One way to measure the expected diagnosticity of asking about Fea-

ture 1 is to calculate the expected absolute log likelihood ratio, or $E(|\,\mathrm{LLR}\,|)$: $.7\,|\log_2(.5/.9)| + .3\,|\log_2(.5/.1)| = 1.3$, which is the value shown in the final column of Table 1 (Klayman & Ha, 1987; Slowiaczek et al., 1992). Normatively speaking, one ought to calculate this (or some related) measure for each of the features in Table 1 and ask about those that have the highest values (for a review of alternative measures of a question's value, see Nelson, 2005).

In a task such as this, participants appear to be primarily concerned with diagnosticity (Bassok & Trope, 1984; Skov & Sherman, 1986; Slowiaczek et al., 1992; Trope & Bassok, 1982, 1983). Participants would prefer asking about Feature 1 to asking about Feature 2. There are, however, at least two other factors that affect choices. One is positivity, which refers to preferring questions for which the probability of a *yes* answer is higher for the focal hypothesis than for the alternate (Klayman & Ha, 1987; Skov & Sherman, 1986). When choosing between Features 1 and 3, for example, participants will tend to prefer Feature 3 if the Glom hypothesis is focal, although $E(|\,\mathrm{LLR}\,|)$ is the same for both questions.

Most important for the present purposes is the third factor: Participants prefer questions for which the probability of a *yes* answer is more extreme (further from .5) under the focal hypothesis than under the alternate hypothesis. For example, if choosing between Features 3 and 4—which have equal $E(|\,\mathrm{LLR}\,|)$ and both of which are positive tests of the Glom hypothesis—participants would prefer Feature 4 if Glom were focal. Skov and Sherman (1986) found that choice of question depended on all three factors: diagnosticity, positivity, and extremity.

None of these testing tendencies leads to confirmation bias. Selecting questions on the basis of any factor other than diagnosticity can lead to inefficiencies, but not necessarily to confirmation bias. As long as one updates confidence appropriately after receiving an answer (as described earlier), there will be no systematic bias favoring the focal hypothesis.

To illustrate this point for extremity bias, imagine that the Fizo hypothesis is the focal hypothesis. Asking about Feature 1 might be an example of an extremity bias, because the feature is very likely under the Fizo hypothesis and only moderately likely under the alternate. A *yes* answer, evidence for the Fizo hypothesis, is likely (70% likely, assuming equal prior probabilities), and a *no* answer, evidence for the Glom hypothesis, is unlikely (30%). However, this does not lead to a confirmation bias, because, although the *yes* answer is likelier than the *no* answer, it is also less diagnostic. As was shown above, the *yes* answer increases confidence in the Fizo hypothesis from 50% to 64%, whereas the *no* answer decreases confidence from 50% to 17%. Indeed, there is always this trade-off when tests of a hypothesis are selected (Poletiek, 2001, chaps. 1 and 2; Poletiek & Berndsen, 2000)—likely outcomes are less diagnostic than unlikely ones—so one cannot select a question beforehand that will systematically favor a hypothesis (at least under the conditions outlined here). If many people independently asked about Feature 1, average confidence in the Fizo hypothesis would be 50%, assuming that confidence was updated appropriately [70% of the people would be 64% confident, and 30% would be 17% confident; $(.7 \times .64) + (.3 \times .17) = .5$]. All else being equal, the extremity bias does not favor a particular hypothesis.

Hypothesis Evaluation: Making Use of Answers

How do people evaluate a hypothesis after receiving an answer to a question or finding out the result of a test? Many studies have been performed to examine this question, and a variety of biases have been reported (e.g., Nickerson, 1998). However, Klayman (1995) and Slowiaczek et al. (1992) have argued that such biases do not necessarily favor the focal hypothesis (see also McKenzie, 2004b). To illustrate, we will focus on the evaluation phenomenon of direct interest: People are insufficiently sensitive to the differential diagnosticity of different answers to the same question. For example, recall that for Feature 1, a person ought to be 64% confident that the creature is a Fizo following a *yes* answer and 83% confident that the creature is a Glom following *no.* However, using materials structurally identical to these, Slowiaczek et al. found that participants reported mean values of 68% and 70% after receiving *yes* and *no* answers, respectively. The participants did not appear to appreciate the extent to which the answers were differentially diagnostic. Slowiaczek et al. replicated this finding, using different Glom and Fizo questions and using tasks involving sampling marbles from urns.

Nonetheless, insensitivity to differential answer diagnosticity does not necessarily favor the focal hypothesis. If Fizo were focal and Feature 1 were asked about, confirmation bias would result: Confidence in the focal hypothesis would be higher than it should be after the *yes* answer ($68\% > 64\%$), and confidence in the alternate would be too low following the *no* answer ($70\% < 83\%$). However, if Glom were focal, the opposite pattern would result: Confidence in the focal hypothesis would systematically be too *low*. Thus, all else being equal, this evaluation bias would be as likely to favor the alternate hypothesis as it would the focal hypothesis.

Extremity Bias + Insensitivity to Answer Diagnosticity = Confirmation Bias

Although neither a preference for *extreme* questions nor insensitivity to differential answer diagnosticity implies confirmation bias, together they do (Klayman, 1995; Slowiaczek et al., 1992), because extremity bias leads the answer supporting the focal hypothesis to be less diagnostic than the answer supporting the alternate. Insensitivity to differential diagnosticity means, in this context, that people will overestimate the diagnosticity of the answer supporting the focal hypothesis and underestimate the diagnosticity of the answer supporting the alternate. The result is systematic overconfidence in the focal hypothesis, or confirmation bias.

Consider once again the case in which Feature 1 is asked about. Assuming an extremity bias, this feature might be asked about if the Fizo hypothesis were focal, because it is

very likely under this hypothesis. As was shown above, insensitivity to the differential diagnosticities of the answers leads to systematic overconfidence in the Fizo hypothesis. (Note that this feature would not be asked about if the Glom hypothesis were focal and an extremity bias were operating.)

To see that the same result occurs when the feature is very *un*likely under the focal hypothesis, consider Feature 3 (Table 1). Extremity again implies that such a feature might be chosen if the Fizo hypothesis were focal. Now, *no* supports the focal hypothesis (confidence = 64%), and *yes* supports the alternate (confidence = 83%). Because the answer supporting the focal hypothesis is less diagnostic than the answer supporting the alternate, insensitivity to this difference will tend to favor the focal hypothesis, which is what Slowiaczek et al. (1992) found.

Participants' Sensitivity to the Rarity of Data

Why, assuming extremity, are the answers that support the focal hypothesis always less diagnostic than the answers supporting the alternate? The reason is that the answers supporting the focal hypothesis are always *more common* than those supporting the alternate for extreme questions. Answers, or test outcomes, are diagnostic to the extent that they are rare or surprising. Most of the creatures possess Feature 1, making the *yes* answer, which supports the focal (Fizo) hypothesis, less diagnostic than the *no* answer, which supports the alternate. Most of the creatures do not possess Feature 3, making the *no* answer, which supports the focal (Fizo) hypothesis, less diagnostic than the *yes* answer, which supports the alternate. Extremity guarantees that the answer supporting the focal hypothesis will be more common and, hence, less diagnostic than the answer supporting the alternate.

It is crucial, though, that it has been shown in other contexts that participants are highly sensitive to the rarity of data, or test outcomes, when both testing and evaluating hypotheses. For example, McKenzie and Mikkelsen (2000) argued that participants, when testing conditional hypotheses ("If X, then Y"), tend to consider confirming observations that are mentioned in hypotheses (the conjunction of X and Y) to provide the strongest support, because they assume that the mentioned observations are rare (and hence really do provide strong support, at least from a Bayesian perspective). When it was clear that the mentioned observations were common, the "bias" disappeared. In a similar vein, Oaksford and Chater (1994) explained, in normative Bayesian terms, a wide variety of results in Wason's (1966, 1968) selection task by adopting the *rarity assumption*—namely, that participants assume that the cards (or the events that they represent) mentioned in the rule to be tested are rare. Several selection task experiments have now shown that making clear to participants that the cards mentioned in the rule are common changes behavior in the predicted direction (Oaksford & Chater, 2003). Both of the accounts above postulate that participants generally assume that the events or features mentioned in the conditional hypothesis are rare. In accord with this view, McKenzie, Ferreira, Mikkelsen, McDermott, and Skrable (2001) found that participants did indeed prefer to phrase conditional hypotheses in terms of rare, not common, events. Thus, participants might treat observations mentioned in conditional hypotheses as rare because they usually are rare. (For additional evidence regarding participants' sensitivity to rarity, see Green & Over, 2000; McKenzie & Amin, 2002; McKenzie & Mikkelsen, in press; Over & Jessop, 1998.)

These findings regarding participants' sensitivity to the rarity of data appear to be at odds with the findings showing insensitivity to differentially diagnostic answers, given that differences in the answers' diagnosticities are due to differences in how rare the answers (or data) are. A plausible explanation of the discrepancy lies in the materials used in the different tasks. In particular, McKenzie and Mikkelsen (2000) found that participants' sensitivity to rarity was significantly enhanced when the participants were familiar with the materials with respect to which features or events were rare (see also McKenzie & Mikkelsen, in press). When abstract and unfamiliar materials were used, the participants were not nearly as capable of making use of rarity information, even when the information was explicitly provided in the form of statistics.

With this in mind, then, the discrepancy between the recent rarity findings and the findings of Slowiaczek et al. (1992) might be due to the fact that the latter used tasks involving abstract and unfamiliar materials, such as sampling marbles from urns and asking creatures from distant planets whether they wear hula hoops. The two experiments below were performed to examine whether participants would show increased sensitivity to differentially diagnostic answers when they were familiar with the rarity of the different answers. To the extent that this was true, it would provide further constraints on the conditions under which confirmation bias would be expected to occur.

EXPERIMENT 1

Method

The participants were 119 University of California at San Diego (UCSD) students who received partial credit for psychology courses by filling out a questionnaire. The participants in the *abstract + statistics* condition read the following:

> Imagine that you have traveled to a distant planet, where there are two types of creature, Gloms and Fizos. You will see four creatures' answers ("yes" or "no") to a question and, based on each creature's answer, guess whether it is a Glom or a Fizo. Keep in mind that 50% of the creatures on the planet are Gloms and 50% are Fizos. You should assume that each creature answered the question honestly.

They then responded to questions on four separate pages. For Creature A, they were told that 50% of the Gloms and 2% of the Fizos wear a hula hoop (Feature 5 in Table 1). This creature was asked, "Do you wear a hula hoop?" and answered "yes" (see Table 2). The participants responded with a best guess as to whether the creature was a Glom or a Fizo and then reported their confidence in their best guess by circling a number on a scale ranging between 50 (*blind guess*) and 100 (*completely certain*) in increments of 5. For Creature B, the participants read that 10% of the Gloms and 0.1% of the Fizos play the harmonica (Feature 6 in Table 1). This creature answered "no" when asked, "Do you play the harmonica?" (Table 2). Creature C was identical to A but answered "no" to the hula hoop question. Creature D was identical to B but answered "yes" to the harmonica question.

Table 2
Experiment 1: Details About the Materials Used in the Three Conditions

Condition	Question	Answer	Hypothesis Supported	Normative Confidence (%)
Abstract + Statistics				
Creature A	Do you wear a hula hoop?	yes	Glom	96
Creature B	Do you play the harmonica?	no	Fizo	53
Creature C	Do you wear a hula hoop?	no	Fizo	66
Creature D	Do you play the harmonica?	yes	Glom	99
Concrete (+ Statistics)				
Student A	Are you taller than 5'10" (178 cm)?	yes	male	96
Student B	Are you taller than 6'1" (185 cm)?	no	female	53
Student C	Are you taller than 5'10" (178 cm)?	no	female	66
Student D	Are you taller than 6'1" (185 cm)?	yes	male	99

Note—The participants in the abstract + statistics condition were told that (1) 50% of Gloms and 2% of Fizos wear a hula hoop and (2) 10% of Gloms and 0.1% of Fizos play the harmonica. The participants in the concrete + statistics condition were told that (1) 50% of males and 2% of females are taller than 5 ft 10 in. and (2) 10% of males and 0.1% of females are taller than 6 ft 1 in. (which is roughly true of North American adults). The participants in the concrete condition were not given any statistical information. For them, "normative confidence" assumes that they believe the statistical information provided to the concrete + statistics participants.

Note that the different answers to both questions are differentially diagnostic. A *yes* answer to "Do you wear a hula hoop?" implies a 96% chance that the creature is a Glom, whereas *no* implies only a 66% chance that it is a Fizo. The asymmetry is even larger for the other question. A *yes* answer to "Do you play the harmonica?" leads to a 99% probability that the creature is a Glom, whereas *no* leads to only a 53% probability that it is a Fizo.

The *concrete + statistics* group read the following:

> Imagine that UCSD students have filled out a survey on a variety of topics. You will see four students' answers ("yes" or "no") to a question and, based on each student's answer, guess the student's gender. Keep in mind that 50% of the students who filled out the survey were male and 50% were female. You will see answers to questions that happen to be about the students' height. It might seem odd that multiple questions about height were on the survey, but redundant questions are often used in surveys to make sure that the respondent is consistent. At any rate, you should assume that each student answered the question honestly.

For Student A, they read that 50% of males and 2% of females are taller than 5 ft 10 in. When asked, "Are you taller than 5 ft 10 in. (178 cm)?" Student A responded "yes" (Table 2). For Student B, the participants read that 10% of males and 0.1% of females are taller than 6 ft 1 in. This student was asked, "Are you taller than 6 ft 1 in. (185 cm)?" and responded "no" (Table 2). Student C was identical to Student A but responded "no," and Student D was identical to B but responded "yes."

Heights were chosen in order to correspond roughly to the statistics provided in the abstract + statistics conditions. The height statistics (and hence the resulting normative answers) were based on the assumptions that the mean male and female heights are 5 ft 10 in. (178 cm) and 5 ft 5 in. (165 cm), respectively, and that the standard deviation of both distributions is 2.5 in. (6 cm), which approximate North American adult heights.

What is important about the concrete + statistics condition is that these participants are familiar with the rarity of the different answers. For example, they know that it is rare for people, especially females, to be taller than 6 ft 1 in., and hence a *yes* response is more diagnostic than a *no* response. The expected result is high confidence that the student is a male in the former case and low confidence that the student is a female in the latter.

The third and final condition, the *concrete* group, was identical to the concrete + statistics group but was not provided with any statistical information. These participants were expected to have learned, to a reasonable approximation, the statistics explicitly provided to the concrete + statistics group. Although it was expected that both concrete groups would outperform the abstract + statistics groups, it was unclear which of the two concrete groups would perform best.

Finally, half of the participants in each of the three groups answered the questions in the order shown in Table 2, and half answered the questions in the reverse order.

Results

The dependent measure used in the analysis was confidence in the normatively favored hypothesis. The participants first made a *best guess* as to the creature's identity or the student's gender on the basis of the creature's/student's answer and then reported confidence in that best guess (on a scale of 50–100). These best guesses were not always normatively correct, however, so for these cases reported confidence in the incorrect hypothesis was transformed to confidence in the correct hypothesis by subtracting the reported confidence from 100. For example, if a participant incorrectly guessed that the student was most likely male and was 70% confident in this incorrect guess, 30% confidence in the normatively correct female hypothesis was used in the analysis. (Potential problems with this methodology will be addressed later.)

The results are shown in the top two panels of Figure 1. The *y*-axis represents mean confidence in the normatively favored hypothesis. The top left panel shows the results for the question (shown below the panel) for which the difference between the diagnosticities of the two answers was relatively small. The left side of the panel corresponds to the *no* answer, and the right side to the *yes* answer. The normative confidence reports after receiving the *no* and *yes* answers to this question are 66% and 96%, respectively (Table 2); these values are also shown in the figure ("Bayes"). A relative lack of sensitivity to the differential diagnosticities of the two answers to the same question is indicated by a relatively flat line. Although the slope of the lines for both concrete groups are flatter than they ought to be, they are nonetheless steeper than that of the abstract + statistics group, who showed almost no sensitivity to the differential diagnosticities of the answers.

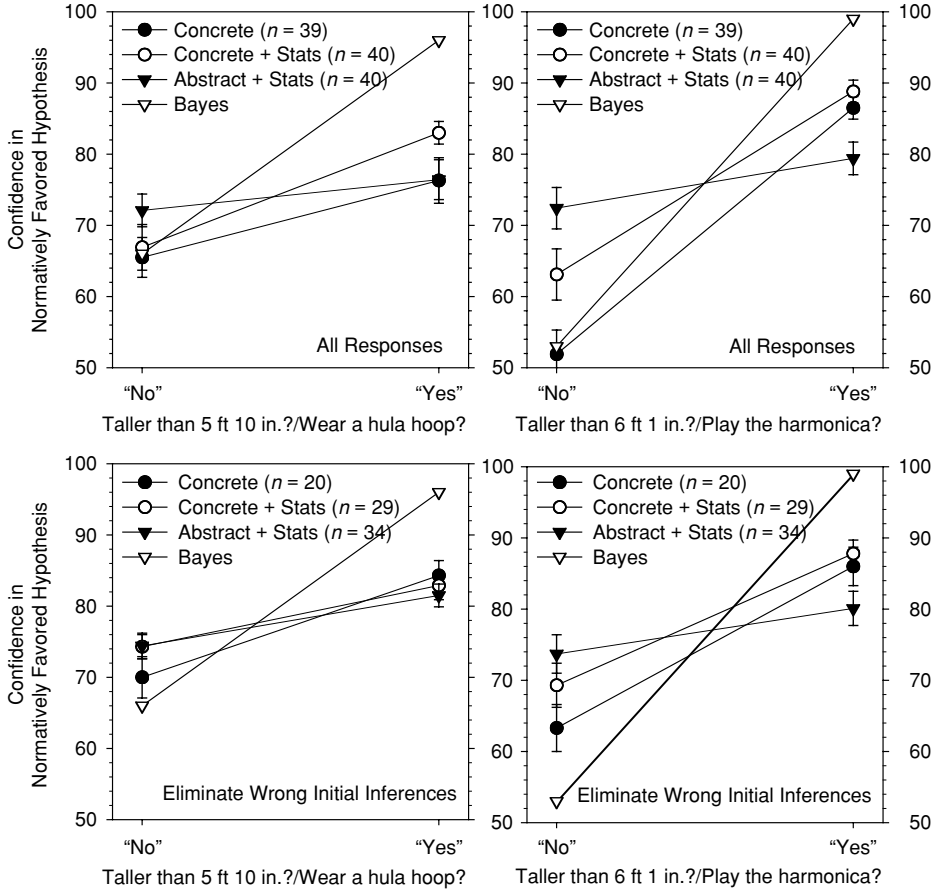

Figure 1. Experiment 1: Confidence in the normatively supported hypothesis as a function of scenario, question, and answer. Bayesian responses and standard error bars are also shown. The top two panels show the results for all of the participants, and the bottom two show the results only for the participants who did not report any incorrect *best guesses*. In all four panels, the line corresponding to the abstract + statistics group is the flattest, showing that this group was least sensitive to the differential diagnosticity of the different answers.

The top right panel shows the results for the question for which the difference between the diagnosticities of the two answers was relatively large (53% vs. 99%). It is again the abstract + statistics group's line that stands out as especially flat.

A 3 (scenario: concrete, concrete + statistics, or abstract + statistics) × 2 (question: small or large difference in answer diagnosticity) × 2 (answer: *no* or *yes*) × 2 (order of questions/answers) mixed model ANOVA was performed on confidence in the normatively supported hypothesis, with the first and fourth variables between subjects and the second and third within subjects. (Bayesian responses shown in Figure 1 are theoretical and were not part of the analysis.) There was a main effect of scenario [$F(2,113) = 3.8, p = .026$], due to the concrete group's reporting overall lower confidence ($M = 70$) than did the concrete + statistics and abstract + statistics groups (Ms = 75). There was also a main effect of answer, with lower confidence following the *no* answer than following the *yes* answer [Ms = 65 and 82; $F(1,113) = 104.5, p < .001$]. This effect can be expected on normative grounds and is easily seen in Figure 1. Importantly, there was a scenario × answer interaction [$F(2,113) = 11.4, p < .001$]: The difference between confidence in responses to the *yes* and *no* answers was smaller for the abstract + statistics group (a difference of 6) than that for the concrete + statistics group (21) and the concrete group (23). There was also a question × answer interaction, which one would expect for normative reasons [$F(1,113) = 20.1, p < .001$]: The slopes are relatively flat in the left panel, as compared with those in the right panel (where the difference between responses to the two answers ought to be larger). Also important is that there was a scenario × question × answer interaction [$F(2,113) = 5.3, p = .006$]. This occurred because, whereas the slope for the abstract + statistics group remained relatively flat across the top two panels, the slopes for the concrete and the concrete + statistics groups increased quite a bit in the top right panel (where the slopes ought to be steeper). There were no other effects.

As has been mentioned, the analysis above was performed using transformed confidence in cases in which

the participants' best guesses were incorrect. On the one hand, this makes use of all the data in a reasonable way. It also allows confidence reports to range between 0 and 100 (rather than 50 and 100), which increases the likelihood that "true" mean judgments close to 50% could result, instead of being "pushed" away from 50% due to the truncation of the scale and the resulting skewed distribution of responses. On the other hand, one might wonder whether the "good" performance (a relatively steep slope) by the two concrete groups (especially in the top right panel) is being driven by confused participants' making incorrect initial inferences following the *no* answers.

To address this, the data were reanalyzed without any participants who made at least one incorrect *best guess*. Note that overall confidence would necessarily increase and at least some of the lines were likely to become flatter. At issue was whether the same pattern of results would hold.

The results are shown in the bottom panels of Figure 1. Note first that 36 of the 119 participants (30%) were excluded from this analysis: 19 from the concrete condition, 11 from the concrete + statistics condition, and 6 from the abstract + statistics condition. (The fact that most of the excluded participants were in the concrete condition will be addressed in the Discussion section.) The same pattern of results nonetheless held up. The line for the abstract + statistics group is the flattest one in both panels. A 3 (scenario) × 2 (question) × 2 (answer) × 2 (order) mixed model ANOVA on confidence in the normatively supported hypothesis revealed a pattern similar to the one with all the participants. There was a main effect of answer [$F(1,77) = 65.7, p < .001$], with *no* answers resulting in lower confidence than did *yes* answers (Ms = 71 vs. 83). Most important is that there was a scenario × answer interaction [$F(2,77) = 4.3, p = .017$]: There was a smaller difference between confidence reports in responses to *yes* and *no* answers for the abstract + statistics group (7) than for the concrete + statistics group (14) and the concrete group (19). As one would expect, there was also a question × answer interaction [$F(1,77) = 7.0, p = .01$]; the slopes are generally steeper in the bottom right panel than in the bottom left panel. The scenario × question × answer interaction approached, but did not reach, significance [$F(2,77) = 2.8, p = .067$]: The abstract + statistics line remains somewhat flat across the two bottom panels, whereas the lines for both the concrete and the concrete + statistics groups become steeper in the bottom right panel. The only other significant effect was the four-way interaction [$F(2,77) = 3.2, p = .047$], which did not offer a straightforward interpretation.

Discussion

The primary result of Experiment 1 was that the abstract + statistics group exhibited little sensitivity to the differential diagnosticity of the different answers to the same question, whereas both concrete groups exhibited much more sensitivity. This finding is relevant because in previous research in which such sensitivity has been examined, groups similar to the abstract + statistics group have been used. Using abstract materials might have been misleading in terms of how well people operate in their natural environment. Familiar materials led to much better performance, even when there was no statistical information. Interestingly, the concrete group tended to show greater sensitivity than did the concrete + statistics group, suggesting that the explicit statistical information hindered rather than helped them (but see McKenzie & Mikkelsen, 2000).

Sensitivity to the differential diagnosticities of answers or test results is important because a lack of sensitivity, combined with an extremity bias, leads to systematic overconfidence in the focal hypothesis, or confirmation bias. Of course, even the concrete participants did not exhibit perfect sensitivity, leaving the door open for at least some degree of confirmation bias. Nonetheless, the present results suggest that to the extent that the materials used are familiar, rather than abstract, confirmation bias is less likely to result.

Recall that most of the participants who were dropped from the analysis that excluded wrong initial *best guesses* were from the concrete group. At the same time, the claim here is that this group is performing best. This is explained by the fact that most of the eliminated concrete participants (17 of 19) made the wrong initial inference when the student was asked "Are you taller than 6 feet 1 inch?" and responded "no." For this question, normative confidence in the normatively supported female hypothesis is 53%, which is virtually a toss-up, and it is not overly surprising that many of the concrete participants sometimes selected the option that was 47% likely. It appears that the participants in this group were more likely to choose the incorrect hypothesis because they were sensitive to this particular answer's *lack* of diagnosticity, leaving them close to 50% confident.

EXPERIMENT 2

This experiment was an attempt to replicate and extend the findings from Experiment 1 in a variety of ways. First, the participants reported confidence in a prespecified hypothesis (on a scale of 0–100), rather than reporting confidence in a best guess (on a scale of 50–100). Second, *yes* and *no* answers were no longer confounded with being strongly and weakly diagnostic, as they were in Experiment 1. Finally, only the question (and a variant) with the most differentially diagnostic answers was used, and the concrete + statistics condition was dropped. The latter changes left only the conditions that were most clearly different in Experiment 1. The end result was a more streamlined and balanced design for addressing the question of whether participants' sensitivity to differential diagnosticity would be enhanced with familiar materials.

Method

The participants were 215 UCSD students who received partial credit for psychology courses and were randomly assigned to the abstract + statistics condition or the concrete condition. The initial instructions were identical to those in Experiment 1. The participants in the abstract + statistics condition read, for Creature A, that 90% of Gloms and 99.9% of Fizos wear a hula hoop (Feature 7

in Table 1). When asked "Do you wear a hula hoop?" Creature A responded "Yes," and the participants should have been 53% confident that the creature was a Fizo. They reported their confidence in a specific hypothesis by circling a number on a scale that ranged from 0 (*certain not [hypothesis]*) to 100 (*certain is [hypothesis]*) in intervals of 5; 50 on the scale was labeled *blind guess*. Half of the abstract + statistics participants reported confidence in the Glom hypothesis for all four creatures, and half reported confidence in the Fizo hypothesis. For Creature B, the participants read that 0.1% of Gloms and 10% of Fizos play the harmonica (Feature 8 in Table 1). When asked, "Do you play the harmonica?" this creature responded "no," and the participants should have been 53% confident that the creature was a Glom. Creature C was identical to Creature A but responded "no" to the hula hoop question, and the participants should have been 99% confident in the Glom hypothesis. Creature D was identical to Creature B but responded "yes" to the harmonica question, and normative confidence in the Fizo hypothesis was 99%. Note that the different answers to the same question were differentially diagnostic and that a *yes* answer to one question was just as diagnostic as a *no* answer to the other.

After reading the introductory information, the participants in the concrete condition reported confidence after each of four hypothetical students answered a question about their height. Student A was asked, "Are you taller than 5 ft 2 in. (157 cm)?" and responded "yes." Student B was asked "Are you taller than 6 ft 1 in. (185 cm)?" and answered "no." Student C was identical to Student A but answered "no." Student D was identical to Student B but answered "yes." Half of the participants reported their confidence that each of the four students was male, and half reported their confidence that each was female. Assuming that the participants believe that 10% of males and 0.1% of females are taller than 6 ft 1 in. and that 99.9% of males and 10% of females are taller than 5 ft 2 in. (which is roughly true of North American adults), normative confidence that Student A is male is 53%, that Student B is female is 53%, that Student C is female is 99%, and that Student D is male is 99%.

Both groups answered the questions in one of two orders, either the one described above or the reverse order.

Results

The dependent measure was again confidence in the normatively supported hypothesis. If the participants were told to report confidence in the hypothesis that was not normatively supported (which was true for two of the four confidence reports for every participant), their confidence was subtracted from 100. Five participants failed to report confidence for at least one of the creatures or students and were excluded from the analysis, leaving 210 participants.

The results are illustrated in the four panels in Figure 2. The top two panels show the results for the participants who reported confidence in the female hypothesis (concrete group) or in the Glom hypothesis (abstract + statistics group), and the bottom two panels show the results for the participants who reported confidence in the male or the Fizo hypothesis. (Note, though, that the *y*-axis corresponds to confidence in the normatively supported hypothesis, regardless of which hypothesis was focal.) The two left panels show the results for the question for which the *no* answer was most diagnostic, and the right two correspond to the question for which the *yes* answer was most diagnostic. Also shown are the theoretical Bayesian responses. Most important is that in all four panels, the line for the abstract + statistics group is flatter than the one for the concrete group. That is, the concrete group was much more sensitive than the abstract + statistics group to the differential diagnosticity of the different answers to the same question.

A 2 (scenario: concrete or abstract + statistics) × 2 (hypothesis: female/Glom or male/Fizo) × 2 (question: *no* or *yes* answer most diagnostic) × 2 (answer: low or high diagnosticity) mixed model ANOVA was conducted on confidence in the normatively supported hypothesis, with the first two variables between subjects and the last two within. (The Bayesian responses played no role in the empirical analysis.) A main effect occurred for answer [$F(1,206) = 236.2$, $p < .001$], which one would expect for normative reasons: Answers high in diagnosticity led to higher confidence than did answers low in diagnosticity (Ms = 79 and 57). Most important is that there was a scenario × answer interaction [$F(1,206) = 46.4$, $p < .001$]: The concrete participants showed more sensitivity than did the abstract + statistics participants to the differentially diagnostic answers. This is the key finding.

There were other significant effects that were not of theoretical interest (and were less reliable). There was a hypothesis × question interaction [$F(1,206) = 9.9$, $p = .002$], which, in terms of Figure 2, is shown by the fact that an increase in overall confidence is evident when one moves from the top left panel to the top right panel, whereas a decrease in confidence is shown between the bottom left and the bottom right panels. Relatedly, there was a scenario × hypothesis × question interaction [$F(1,206) = 9.9$, $p = .002$]. For the concrete group, overall confidence in each of the four panels is essentially unchanged, whereas for the abstract + statistics group, overall confidence increases between the top left and the top right panels and decreases between the bottom left and the bottom right panels. There was also an interaction between hypothesis and answer [$F(1,206) = 12.4$, $p < .001$]. Collapsing across groups, there was a smaller difference in confidence for responding to the weakly versus the strongly confirming answer for the male/Fizo hypothesis than for the female/Glom hypothesis. Figure 2 shows that the slopes in the top two panels are, on average, steeper than the slopes in the bottom two panels. In addition, there was a scenario × question × answer interaction [$F(1,206) = 4.7$, $p = .031$]. Collapsing across the top and bottom panels in Figure 2 (i.e., across hypothesis tested), the concrete participants showed increased sensitivity to the weakly and strongly confirming answers, moving from the left to right panels, whereas the abstract + statistics participants showed slightly decreased sensitivity. Finally, there was a significant hypothesis × question × answer interaction [$F(1,206) = 4.1$, $p = .044$]. In terms of Figure 2, and collapsing across groups, less sensitivity to the differentially diagnostic answers is shown in the bottom left panel. The reasons for these interactions are unclear.

Discussion

The results of Experiment 2 replicated those of Experiment 1: The participants presented with abstract materials were much less sensitive than were those presented with familiar materials to the differential diagnosticity of

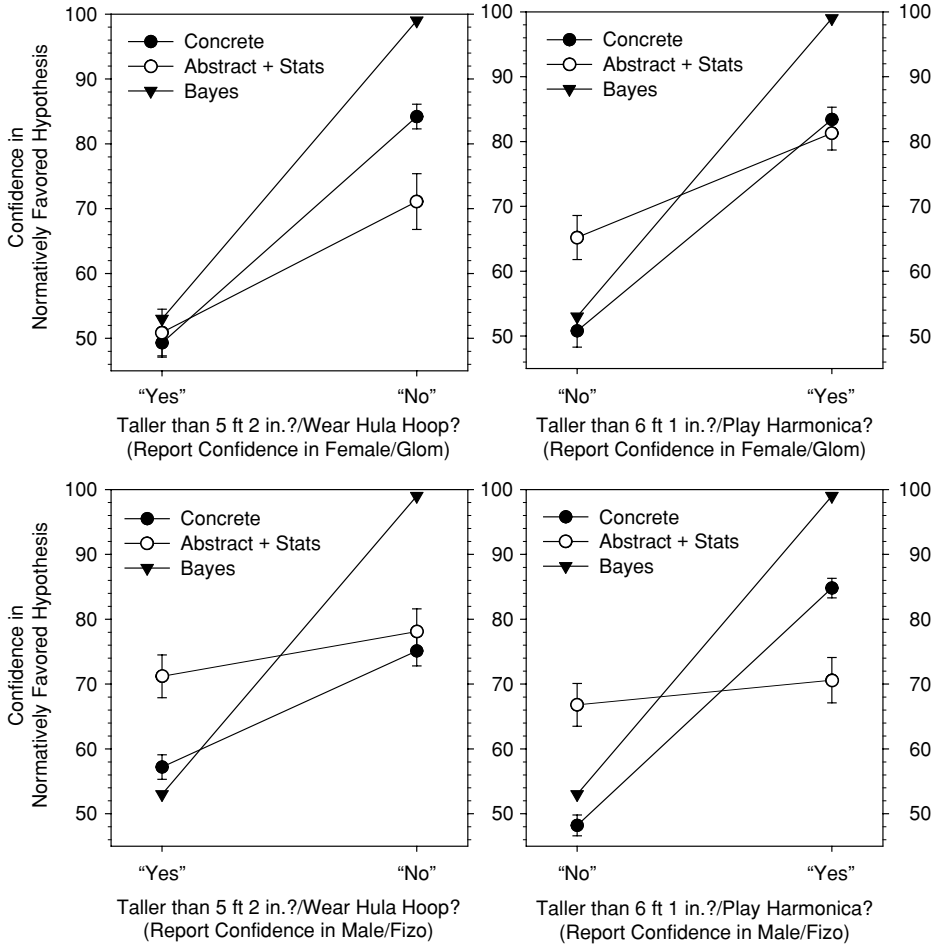

Figure 2. Experiment 2: Confidence in the normatively supported hypothesis as a function of scenario, question, answer, and focal hypothesis. Bayesian responses and standard error bars are also shown. In all four panels, the line corresponding to the abstract + statistics group is the flattest, showing that this group was least sensitive to the differential diagnosticity of the different answers.

different answers to the same question. The finding was replicated when the participants reported confidence in a prespecified focal hypothesis (rather than reporting confidence in an initial best guess) and when both *yes* and *no* answers were the most diagnostic. Previous research in which only abstract materials were used appears to have underestimated participants' sensitivity to differential diagnosticity and, therefore, overestimated the extent to which confirmation bias occurs.

Thus far, the results have been presented in terms of confidence in the normatively supported hypothesis. In order to make a more direct connection to confirmation bias, the results from Experiment 2 are shown in a slightly different form in Figure 3, where the *y*-axis corresponds to confidence in the focal hypothesis when a preference for choosing "extreme" questions is assumed. The data in Figure 3 have been collapsed across the four panels in Figure 2, and the data points on the right side were subtracted from 100. The normative responses in Figure 3 are 53% for the low-diagnosticity answer (left side) and 1% for the high-diagnosticity answer (right side). Now the height, rather than the slope, of each group's line, relative to the Bayesian line, is most important. In particular, the further a group's line is above the Bayesian line, the greater the overconfidence in the focal hypothesis, or confirmation bias. The concrete group's confidence is virtually ideal for the low-diagnosticity answer but is too high for the high-diagnosticity answer. Averaging across the two points, overconfidence is less than 8%. By contrast, the abstract + statistics group's confidence is too high for both answers. It is also higher than the concrete group's confidence in both cases. The abstract + statistic group's average overconfidence is more than 17%.

GENERAL DISCUSSION

Previous research has demonstrated both a hypothesis-testing tendency to ask *extreme* questions and a hypothesis evaluation tendency to be insufficiently sensitive to the differential diagnosticity of different answers to the same

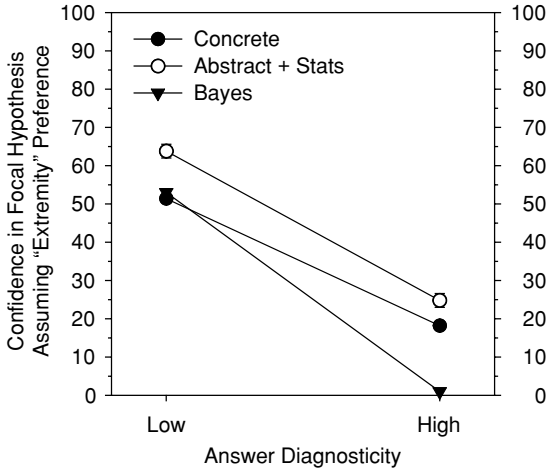

Figure 3. Experiment 2: Confidence in the focal hypothesis when a preference for extreme questions is assumed. Bayesian responses and standard error bars are also shown. The concrete group exhibited overconfidence only for the high-diagnosticity answer, whereas the abstract + statistics group exhibited overconfidence for both answers, as well as greater (over)confidence than did the concrete group for both answers.

question. Neither of these tendencies alone necessitates confirmation bias, but together they do. Two experiments were performed to examine the evaluation component and showed that sensitivity was substantially increased when familiar materials were used, instead of abstract and unfamiliar materials, which have usually been used in such experiments. Recent insights that confirmation bias does not necessarily arise from testing or evaluation strategies by themselves suggested a large reduction in the conditions under which confirmation bias is expected to occur (Klayman, 1995; Poletiek, 2001; Slowiaczek et al., 1992). The main implication of the experiments reported here is that the conditions are even more constrained.

Of course, the participants presented with familiar materials exhibited less than perfect sensitivity, and therefore, confirmation bias may occur nonetheless. However, the extent to which it will occur appears to be small. In Figure 3, for example, overconfidence in the focal hypothesis (assuming an extremity preference) for the concrete participants was less than 8%. If the original experiments demonstrating insensitivity to differential answer diagnosticity had shown the degree of sensitivity demonstrated here with familiar materials, the issue of confirmation bias might never have been raised.

Even the 8% value probably represents an upper limit for the concrete participants in this task. In Figure 3, these participants responded nearly optimally to the low-diagnosticity answer (they were actually 2% underconfident). All of their overconfidence was caused by the high-diagnosticity answer, to which the participants ought to have responded with 1% confidence. However, random noise in the participants' responses would have pushed them away from the end of the scale toward the middle. This is not to say that the overconfidence was due only to noisy responses, but it is reasonable to assume at least some noise. To the extent that random error was responsible for the concrete participants' overconfidence, a biased processing account is not needed to explain even their relatively small degree of confirmation bias.[1]

At this point, it is unclear why familiarity improves performance. One possibility is that using familiar materials facilitates thinking of concrete examples, such as how many males and females one knows who are taller than 6 ft 1 in. Another is that familiarity allows for tapping into more abstract knowledge, in which case responses may be the result of a more immediate "gut feeling" learned from experience. Interestingly, Markovits (1986) found that performance in a reasoning task improved with familiar materials independently of participants' ability to generate concrete instances. Nonetheless, it is perhaps worth mentioning that familiarity could lead to worse performance if experience were at odds with the task at hand. Although a theoretical account of why familiarity is important is not offered here, the present findings have both theoretical and practical value: In our daily lives, we presumably make inferences about variables we are familiar with, and confirmation bias is relatively unlikely to occur under these conditions.

Moreover, recall that the *testing* part (a preference for extremity) of the testing/evaluation combination examined here has been taken for granted. That is, it has been assumed that participants consistently exhibit an extremity bias when choosing questions to ask (or tests to conduct). To the extent that a preference for extremity is not strong or consistent, confirmation bias is less likely to occur.

Klayman (1995) discussed three testing/evaluation combinations that lead to confirmation bias, only one of which has been addressed here. A second combination was positivity (asking questions that you expect a *yes* answer to, if the focal hypothesis is correct) and the fact that respondents are biased to answer *yes* to questions in social settings (*acquiescence bias*; Zuckerman, Knee, Hodgins, & Miyake, 1995). Interviewers apparently do not take respondents' biases into account, leading to overconfidence in the focal hypothesis. Although interesting, note that this example of confirmation bias is limited to asking *yes*/*no* questions in social settings and does not suggest a widespread bias.

The final example discussed by Klayman (1995) is the combination of positivity—in this case, asking about features expected to be present if the focal hypothesis is true—and the fact that participants are more affected by the presence of features than by their absence (e.g., feature-positive effects; Jenkins & Sainsbury, 1969, 1970; Newman, Wolff, & Hearst, 1980). Because positive testing implies that the presence of features confirms the hypothesis and their absence disconfirms the hypothesis, and feature-positive effects imply that presence has more impact than does absence, evidence favoring the hypothesis will have the most impact. However, feature-positive effects have been found primarily in discrimination learning tasks, where it seems reasonable for participants to assume that it is more useful to attend to what something is

rather than to what it is not (e.g., you can more easily learn about elephants by being told what they are than by being told what they are not). Using more traditional hypothesis-testing tasks, Slowiaczek et al. (1992) found only mixed results with respect to participants' being influenced more by *yes* than by *no* answers to questions about the presence of features. It remains to be seen to what degree feature-positive effects occur in hypothesis evaluation.

In sum, confirmation bias might not be as widespread as previously thought, not only because it requires testing and evaluation biases working together, but also because (at least) one such testing/evaluation combination leads to considerable confirmation bias only when abstract and unfamiliar materials are used. When familiar materials are used, the degree of confirmation bias is largely, although not entirely, erased. Other testing/evaluation combinations await close examination in order to determine the degree to which confirmation bias occurs under general, or only under very limited, circumstances.

REFERENCES

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

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral & Brain Sciences*, **14**, 471-517.

Bassok, M., & Trope, Y. (1984). People's strategies for testing hypotheses about another's personality: Confirmatory or diagnostic? *Social Cognition*, **2**, 199-216.

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

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences*, **2**, 206-214.

Evans, J. S. 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.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, **2**, 83-115.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, **103**, 592-596.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.) (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Green, D. W., & Over, D. E. (2000). Decision theoretical effects in testing a causal conditional. *Current Psychology of Cognition*, **19**, 51-68.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, **118**, 248-271.

Jenkins, H. M., & Sainsbury, R. S. (1969). The development of stimulus control through differential reinforcement. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 123-161). Halifax: Dalhousie University Press.

Jenkins, H. M., & Sainsbury, R. S. (1970). Discrimination learning with the distinctive feature on positive or negative trials. In D. Mostofsky (Ed.), *Attention: Contemporary theory and analysis*. New York: Appleton-Century-Crofts.

Kahneman, D., & Tversky, A. (Eds.) (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning & Motivation*, **32**, 385-418.

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

Markovits, H. (1986). Familiarity effects in conditional reasoning. *Journal of Educational Psychology*, **78**, 492-494.

McKenzie, C. R. M. (2003). Rational models as theories—not standards—of behavior. *Trends in Cognitive Sciences*, **7**, 403-406.

McKenzie, C. R. M. (2004a). Framing effects in inference tasks—and why they are normatively defensible. *Memory & Cognition*, **32**, 874-885.

McKenzie, C. R. M. (2004b). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 200-219). Oxford: Blackwell.

McKenzie, C. R. M., & Amin, M. B. (2002). When wrong predictions provide more support than right ones. *Psychonomic Bulletin & Review*, **9**, 821-828.

McKenzie, C. R. M., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., & Skrable, R. P. (2001). Do conditional hypotheses target rare events? *Organizational Behavior & Human Decision Processes*, **85**, 291-309.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, **7**, 360-366.

McKenzie, C. R. M., & Mikkelsen, L. A. (in press). A Bayesian view of covariation assessment. *Cognitive Psychology*.

McKenzie, C. R. M., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review*, **10**, 596-602.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, **112**, 979-999.

Newman, J., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 630-650.

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

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

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, **103**, 381-391.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, **10**, 289-318.

Over, D. E., & Jessop, A. (1998). Rational analysis of causal conditionals and the selection task. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 399-414). Oxford: Oxford University Press.

Poletiek, F. [H.] (2001). *Hypothesis-testing behaviour*. Hove, U.K.: Psychology Press.

Poletiek, F. H., & Berndsen, M. (2000). Hypothesis testing as risk behaviour with regard to beliefs. *Journal of Behavioral Decision Making*, **13**, 107-123.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Mahwah, NJ: Erlbaum.

Sher, S., & McKenzie, C. R. M. (in press). Information leakage from logically equivalent frames. *Cognition*.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, **22**, 93-121.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, **20**, 392-405.

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality & Social Psychology*, **43**, 22-34.

Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis testing. *Journal of Experimental Social Psychology*, **19**, 560-576.

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. 135-151). Harmondsworth, U.K.: Penguin.

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

Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set. *Journal of Personality & Social Psychology*, **68**, 52-60.

NOTE

1. Another factor that can artificially decrease participants' apparent sensitivity to differential diagnosticity is to eliminate participants who make an incorrect *best guess*, as was done in the reanalysis of Experiment 1 (compare the top and the bottom panels in Figure 1). Indeed, this is what Slowiaczek et al. (1992) did in their Experiments 1A and 2A. However, as can be seen in Figure 1, excluding these participants had little impact on the results for the group presented with the abstract + statistics scenario, which is most similar to the scenario used by Slowiaczek et al.

(Manuscript received October 27, 2004;
revision accepted for publication April 1, 2005.)