Available online at www.sciencedirect.com
ScienceDirect
Cognitive Psychology 54 (2007) 33–61
Cognitive Psychology
www.elsevier.com/locate/cogpsych

A Bayesian view of covariation assessment ☆

Craig R.M. McKenzie *, Laurie A. Mikkelsen

Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA

Accepted 19 April 2006
Available online 9 June 2006

Abstract

When participants assess the relationship between two variables, each with levels of presence and absence, the two most robust phenomena are that: (a) observing the joint presence of the variables has the largest impact on judgment and observing joint absence has the smallest impact, and (b) participants' prior beliefs about the variables' relationship influence judgment. Both phenomena represent departures from the traditional normative model (the phi coefficient or related measures) and have therefore been interpreted as systematic errors. However, both phenomena are consistent with a Bayesian approach to the task. From a Bayesian perspective: (a) joint presence is normatively more informative than joint absence if the presence of variables is rarer than their absence, and (b) *failing* to incorporate prior beliefs is a normative error. Empirical evidence is reported showing that joint absence is seen as more informative than joint presence when it is clear that absence of the variables, rather than their presence, is rare.

Keywords: Covariation assessment; Rationality; Bayesian inference

Although reasoning and decision making errors are often reported (e.g., Evans, Newstead, & Byrne, 1993; Gilovich, Griffin, & Kahneman, 2002; Kahneman & Tversky, 2000), they are often disputed as well. For example, sometimes it is argued that partici-

☆ This research was supported by National Science Foundation Grants SES-0079615, and SES-0242049. The authors thank John Anderson, Michael Liersch, Mike Oaksford, and Shlomi Sher for their helpful comments, and Evan Stanelle for help with the proof.
* Corresponding author. Fax: +1 858 534 7190.
E-mail address: cmckenzie@ucsd.edu (C.R.M. McKenzie).
URL: psy.ucsd.edu/~mckenzie (C.R.M. McKenzie).

0010-0285/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2006.04.004

pants construe tasks differently than experimenters (Hilton, 1995; Schwarz, 1996), that many errors are limited to (or at least exacerbated by) the laboratory environment (Anderson, 1990, 1991; Klayman & Ha, 1987; McKenzie, 2003, 2004a, in press; McKenzie & Mikkelsen, 2000; McKenzie & Nelson, 2003; Oaksford & Chater, 1994, 1996, 2003), and that some purported errors are consistent with an alternative normative standard (Anderson, 1990, 1991; Chase, Hertwig, & Gigerenzer, 1998; Gigerenzer, 1991, 1996; Gigerenzer et al., 1999; McKenzie, 2004a; Sher & McKenzie, in press; Oaksford & Chater, 1994, 1996, 2003). In this article, we invoke all of the above arguments to explain robust "errors" in covariation assessment.

Assessing how variables covary underlies such fundamental behaviors as learning (Hilgard & Bower, 1975), categorization (Smith & Medin, 1981), and judging causation (Cheng, 1997; Cheng & Novick, 1990, 1992; Einhorn & Hogarth, 1986), to name just a few. Crocker (1981) noted that people's ability to accurately assess covariation allows them to explain the past, control the present, and predict the future. It is hard to imagine a more important cognitive activity and, accordingly, much research has been devoted to this topic since the groundbreaking studies of Inhelder and Piaget (1958) and Smedslund (1963; for reviews, see Allan, 1993; McKenzie, 1994).

Despite the important role that covariation assessment plays in people's daily lives, most research over the last four decades examining performance with two binary variables—presumably the simplest possible case—has concluded that people are surprisingly poor at the task. Two robust findings have stood out over the years. One is that participants do not treat the four cells of a 2×2 contingency matrix as equally important, especially when both variables have levels of presence and absence. In particular, observing the joint presence of variables has the largest impact on judgments, and observing their joint absence has the smallest impact (e.g. Levin, Wasserman, & Kao, 1993; Kao & Wasserman, 1993; Lipe, 1990; Schustack & Sternberg, 1981; Wasserman, Dorner, & Kao, 1990). The second robust finding is that participants' prior beliefs about a relationship influence judgments of covariation (e.g. Alloy & Tabachnik, 1984; Chapman & Chapman, 1967, 1969; Crocker, 1981; Jennings, Amabile, & Ross, 1982; Nisbett & Ross, 1980; Peterson, 1980). Both of these findings represent departures from the traditional normative model and have therefore usually been interpreted as shortcomings in participants' ability to assess covariation.

We present a nontraditional view of both the covariation task and participants' behavior. Generally, we argue that participants approach the task from an *inferential* perspective rather than from the traditional *descriptive* (in the statistical sense) perspective (see also Griffiths & Tenenbaum, 2005). The traditional view is that the participant's task is to provide a summary of the presented four cell values and, therefore, any additional real-world knowledge is irrelevant. Our view is that participants naturally adopt an inferential approach and attempt to determine the likelihood that there is a relationship between the variables. Furthermore, participants exploit their knowledge related to the inferential task that makes sense outside the laboratory. More specifically, a Bayesian account can naturally explain why participants are influenced by prior beliefs about the relation to be assessed and why the four cells are seen as differentially informative. The former phenomenon is the hallmark of Bayesian statistics and therefore fits naturally with the current perspective. The latter phenomenon, differential cell informativeness, is due to participants' knowledge, or assumptions in the case of impoverished laboratory stimuli, about which of the four types of observations are rare. From a Bayesian perspective, observing

the conjunction of two events is informative to the extent that the events are rare (e.g. Horwich, 1982; Howson & Urbach, 1989). We will show that, unless there is good reason to believe otherwise, participants assume (probably tacitly) that the presence of variables is rare, and that is why joint presence has the largest influence on behavior and joint absence has the smallest influence. These two Bayesian principles—incorporating prior beliefs and finding conjunctions of rare events more informative than conjunctions of common ones—can account for a large amount of research that asks participants for a summary judgment of covariation between two binary variables.

We want to make clear from the outset that our Bayesian account neither assumes nor implies that people are Bayes-optimal processors of information. In addition to the wealth of evidence indicating that people are not optimal Bayesians (e.g. McKenzie, 1994), Bayesian models are notorious for their enormous complexity when applied to real-world problems, making them unlikely candidates for models of psychological processes. Even Pearl's (1988) Bayesian network approach to modeling human inference, meant to avoid such "scaling up" problems, has been deemed unmanageable for even moderately complex problems (Dagum & Luby, 1993). In contrast to these quantitative approaches, however, our account simply assumes that people take into account their prior beliefs and event rarity in qualitatively appropriate ways when assessing covariation. We will show that there is strong evidence that both of these Bayesian principles are used in lay covariation assessment.

Others have argued for various aspects of a Bayesian account of covariation assessment (or closely related tasks). Most relevant is Anderson (1990), who adopted a Bayesian perspective of inferring causation between two binary variables and was the first (to our knowledge) to note that observing the joint presence of variables might be normatively more informative than observing their joint absence because presence is rare. His insightful analysis is in many ways more detailed than ours, but his account has, we believe, failed to convey the important message that simple Bayesian principles can explain covariation behavior. Part of the reason for this is that his "rational analysis" assumes that organisms are *optimally* adapted to their environment (given minimal assumptions about processing limitations), which led, in our opinion, to two closely related shortcomings: Many find implausible the idea that people are optimal Bayesians (as noted above), and the resulting model was complicated and lacked intuitive appeal. It will become evident that we do believe that cognition is influenced by the structure of the environment, or "how the world usually works," but we do not assume that the adaptation is optimal. This will be reflected in our qualitative, rather than quantitative, Bayesian approach.

Two other important differences between our approach and Anderson's deserve mention. The first is that Anderson (1990, 1991; Anderson and Sheu, 1995) did not bolster the Bayesian account by pointing to the many studies showing that prior beliefs influence covariation judgments, which we will do later in this article. The second is that Anderson's (1990, 1991) rational analysis is concerned with whether cognition is adaptive, whereas we are also concerned with whether it is adapt*able* (Klayman & Brown, 1993; McKenzie, 2005; McKenzie & Mikkelsen, 2000). By "adaptive," we mean that behavior reflects certain (largely invariant) aspects of the environment. We take Anderson's (1990, 1991; Anderson and Sheu, 1995) argument regarding the fact that presence tends to be rare to be an argument that the "bias" for joint presence over joint absence is adaptive. However, despite its importance, Anderson's account remains post hoc and appears mute regarding the important question of whether covariation behavior is adaptable—that is, whether

behavior changes appropriately when it is clear that the situation is atypical. In particular, the question arises as to whether participants will find joint absence more informative than joint presence if it is clear that absence, rather than presence, is rare. The answer to this question—which we provide later—is crucial for establishing the Bayesian perspective as a testable account that leads to new predictions rather than just an intriguing post hoc rationalization of biases in covariation judgment.

Fales and Wasserman (1992) entertained the possibility that a Bayesian approach might explain human causal inference, pointing out, for example, that observed learning curves are consistent with the approach. However, because they contrasted Bayes with Rescorla and Wagner's (1972) model of associative learning, they emphasized situations in which multiple predictor variables compete with each other to explain another outcome variable, whereas the current focus is on just two variables (i.e., one predictor and one outcome variable). Furthermore, Fales and Wasserman were concerned with neither the differential informativeness of the four cells nor the influence of prior probabilities on learning causal relations.

Recently, Griffiths and Tenenbaum (2005) also argued for a Bayesian account of inferring causation between two binary variables. They successfully explained patterns in causal judgments that other theories could not, thereby lending considerable credence to a Bayesian perspective. For example, they showed that simply manipulating sample size (by multiplying all cell frequencies by a constant) affected judgments in the predicted manner. Like Fales and Wasserman (1992), however, Griffiths and Tenenbaum (2005) were not concerned with explaining joint presence bias and the influence of prior beliefs. Furthermore, their model's predictions were based entirely on the frequencies in the 2×2 matrices presented to participants, whereas we will show that participants' knowledge beyond cell frequencies is important. Griffiths and Tenenbaum's model could be augmented to incorporate such knowledge, however.

Cheng (1997) also presented a normative (albeit non-Bayesian) view of inferring causation. Though our focus is on covariation rather than causation, her model also indicates that joint presence is normatively more informative than joint absence. Importantly, though, considerations of rarity do not play a role in her conclusion regarding differential cell importance. Hence, unlike our account, Cheng's does not predict that joint absence will be seen as more informative than joint presence when it is clear that absence of the variables is rare.

Also relevant is Alloy and Tabachnik (1984), who reviewed a large number of covariation studies and concluded not only that prior beliefs influenced covariation assessment, but also that this was normatively justifiable. However, despite their focus on prior beliefs, the authors did not adopt a Bayesian perspective, and they ended up treating prior beliefs as a layer to be added to the traditional normative model. This forced the authors to make an awkward distinction between being "accurate"—reporting judgments in accord with the traditional normative model—and being "rational"—incorporating prior beliefs. There is no need to make such a distinction within a Bayesian framework: Incorporating prior beliefs is rational precisely *because* doing so increases accuracy (assuming that the prior beliefs are reasonably accurate). An additional important difference between the current perspective and that of Alloy and Tabachnik's is that the latter did not attempt to account for why the four cells are perceived by participants as differentially informative.

The rest of the article is organized as follows. In the first part, we describe the type of covariation task we intend to explain as well as the traditional normative model. The second part briefly reviews research on covariation assessment with an eye toward conclusions regarding the dominating influence of variables' joint presence over their joint

absence. We then introduce the likelihood portion of Bayes' theorem, showing that it predicts that joint presence will be seen as more informative than joint absence when presence is rare—a condition that, as Anderson (1990) has argued, appears to hold in the real world. Subsequently, we highlight differences between the traditional and inferential views of the task. Next, some recent research is summarized indicating that participants are highly sensitive to the rarity of data when making inferences, as they should be from a Bayesian perspective. Two experiments then test the prediction that joint absence will be seen as more informative when absence, rather than presence, is rare. We subsequently point out the natural fit between the Bayesian viewpoint and the large body of data showing that prior beliefs influence covariation assessment. Finally, we discuss the contribution of our analysis and data to the current debate regarding human rationality.

1. The traditional covariation task and the traditional normative model

In the typical covariation task we are concerned with, there are just two variables, each with levels of presence and absence, creating the familiar 2×2 contingency matrix. Such a matrix is depicted in Fig. 1 for Variables X and Y. Cell A corresponds to the joint presence of the variables, Cell B to the presence of X and the absence of Y, Cell C to the absence of X and the presence of Y, and Cell D to the joint absence of the variables. Imagine, for example, being presented with the following information regarding a treatment and recovery from an illness: 15 people received the treatment and recovered (Cell A), 5 people received the treatment and did not recover (Cell B), 9 people did not receive the treatment and recovered (Cell C), and 3 people did not receive the treatment and did not recover (Cell D). Sometimes the observations are presented sequentially, while other times the information is summarized (as above). Participants might be asked to assess the strength of the relation given the four cell values, or they might be asked which of two matrices shows a stronger relation. A model considered normative in this context is the phi coefficient: $\varphi = (AD - BC)/[(A + B)(C + D)(A + C)(B + D)]^{1/2}$, where A, B, C, and D correspond to the respective cell values. A simpler model, $\Delta p = [A/(A + B)] - [C/(C + D)]$, is often substituted (Allan, 1980). Phi is a special case of Pearson's product-moment correlation coefficient, ranging between -1 and 1. To the extent that the coefficient is close to 1 (-1), there is a strong positive (negative) relation between the variables: Y is more (less) likely to be present when X is present rather than absent. When $\varphi = 0$, as in the above example, X and Y are independent. Note that φ and Δp are *descriptive statistics* in that they are simply *summaries of the presented information*. No information beyond the four cell frequencies is considered relevant. If any additional information or beliefs were to influence judgment, this would be considered an error from the traditional perspective.

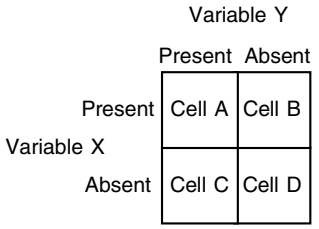

Fig. 1. The four cells of a 2×2 contingency matrix.

2. Past research on differential cell impact

Probably the most robust finding in the covariation literature is that participants do not treat the four cells as equally important. In particular, observing the joint presence of the variables has the largest impact on judgments, and observing their joint absence has the smallest impact. This has been shown by regressing strength judgments onto cell frequencies (Mandel & Lehman, 1998; Schustack & Sternberg, 1981), by asking participants directly which cells are most important (Crocker, 1982; Wasserman et al., 1990), by inferring cell importance and/or which rule participants use based on patterns of responses (Arkes & Harkness, 1983; Kao & Wasserman, 1993; Levin et al., 1993; Shaklee & Mims, 1982; Shaklee & Tucker, 1980; Ward & Jenkins, 1965; Wasserman et al., 1990), and by a meta-analysis of covariation research (Lipe, 1990). The impacts of Cells B and C, which are sometimes equivalent to each other, fall between that of A and D (Crocker, 1982; Kao & Wasserman, 1993; Levin et al., 1993; Lipe, 1990; Wasserman et al., 1990). Because the clearest empirical difference is between the impact of Cells A and D, these two cells are the focus of this article, both theoretically and empirically.

Because the four cells are equally important in calculating φ (more on this below), differential impact of the four cells has been routinely interpreted as nonnormative. For example, Kao and Wasserman (1993) state that, "It is important to recognize that unequal utilization of cell information implies that nonnormative processes are at work" (p. 1365), and Mandel and Lehman (1998) attempted to explain differential cell informativeness in terms of a combination of two reasoning biases. However, based on normative Bayesian principles, the following approach naturally accounts for, among other things, the preference for joint presence over joint absence.

3. A Bayesian view of the four cells' informativeness

Assume for the moment that participants, rather than viewing the task as one of *description*, naturally approach it as one of *inference*, in which they are attempting to use the cell information to distinguish between two mutually exclusive hypotheses about a larger population of instances. For illustrative purposes, assume that one hypothesis, H1, corresponds to a moderate positive contingency between X and Y, $\rho = 0.5$ (where ρ is the population's hypothesized φ), and that the alternative hypothesis, H2, corresponds to noncontingency, $\rho = 0$. That is, assume that participants are trying to determine how likely it is that there is a moderate contingency between the variables rather than none (see also Griffiths & Tenenbaum, 2005).

Now we can ask how informative each of the four possible observations is under these conditions. One way to measure the informativeness of data is to calculate how well they help distinguish between the hypotheses under consideration (see, e.g. Nelson, 2005), which is quantified using likelihood ratios. The numerator of a likelihood ratio corresponds to the probability of observing the data assuming that H1 is true, and the denominator to the probability of observing the same data assuming that H2 is true. A datum is diagnostic to the extent that its likelihood ratio differs from 1. In the case of a Cell A observation, the likelihood ratio is $p(A|\mathrm{H1})/p(A|\mathrm{H2})$. Because H2 is $\rho = 0$—that is, X and Y are independent—$p(A|\mathrm{H2}) = p(X)p(Y)$, where $p(X)$ and $p(Y)$ correspond to the relative frequency, or the subjective probability, of the presence of each of the variables, X and Y. Similarly, $p(B|\mathrm{H2}) = p(X)[1 - p(Y)]$, $p(C|\mathrm{H2}) = [1 - p(X)]p(Y)$, and $p(D|\mathrm{H2}) =$

$[1 - p(X)][1 - p(Y)]$. Assume that $p(X) = p(Y) = .1$; that is, the presence of each variable is rare. The denominator of the likelihood ratio for *A*, *B*, *C*, and *D* observations equals 0.01, 0.09, 0.09, and 0.81, respectively.

What about the numerators? For $\rho = 0.5$ (H1), the respective numerators equal 0.055, 0.045, 0.045, and 0.855.[1] Thus, the likelihood ratios corresponding to *A*, *B*, *C*, and *D* observations are 5.5, 0.5, 0.5, and 1.06, respectively. The fact that the likelihood ratios for *A* and *D* observations are greater than 1 demonstrate that they are evidence in favor of H1, and the likelihood ratios of less than 1 for *B* and *C* observations indicate that they are evidence in favor of H2. We will be dealing with situations in which the qualitative impact of each observation is clear (*A* and *D* observations always favor one hypothesis, and *B* and *C* the other) and will concentrate on how informative a given observation is, regardless of the hypothesis it favors. The measure of informativeness we will use is the absolute log likelihood ratio (|LLR|):

$$|\mathrm{LLR}_j| = \mathrm{Abs}(\log_2[p(j|\mathrm{H1})/p(j|\mathrm{H2})]),$$

where *j* corresponds to A, B, C, or D. The larger |LLR| is, the more informative the observation. When LLR = 0, the observation is completely uninformative. |LLR| equates a likelihood ratio and its inverse and is a commonly used measure of informativeness in Bayesian hypothesis testing (e.g. Evans & Over, 1996; Good, 1983; Klayman & Ha, 1987).

For the *A* through *D* observations in this example, then, |LLR| equals 2.46, 1.0, 1.0, and 0.08, respectively. Given the above assumptions, a Cell A observation is most informative, a Cell D observation is least informative (indeed, it is virtually uninformative), and Cells B and C fall in between. This is, of course, consistent with the robust empirical finding that $A > B \approx C > D$. However, several assumptions were made in the above analysis, including H1 was $\rho = 0.5$, H2 was $\rho = 0$, and $p(X) = p(Y) = 0.1$. For simplicity, we also treated φ as an unbiased estimator of ρ, although it is not. How sensitive to these assumptions is the result that $|\mathrm{LLR_A}| > |\mathrm{LLR_D}|$? If the marginal probabilities, $p(X)$ and $p(Y)$, do not change under the competing hypotheses, then the competing hypotheses are irrelevant. All that is necessary for a Cell A observation to be more informative than a Cell D observation is that $p(X) < 1 - p(Y)$ (Horwich, 1982; Mackie, 1963; McKenzie & Mikkelsen, 2000). Cell D observations are more informative than Cell A observations when $p(X) > 1 - p(Y)$ (see also Over & Green, 2001; Over & Jessop, 1998). As we show in Appendix A, they are equally informative only when $p(X) = 1 - p(Y)$.[2]

Given that participants' preference for Cell A to Cell D would be rational in an inferential task if *X* and *Y* were rare (see also Anderson, 1990), it is natural to wonder whether

[1] After calculating the cell values for the independent case ($\varphi = 0$), cell values for a different value of φ can be calculated by multiplying that φ value by its denominator (the square root of the product of the four marginal probabilities; see text for calculating φ). Relative to the values in the independent case, add this product to the *A* and *D* values and subtract this product from the *B* and *C* values. Using the example in the text, $0.5 \times (.1^2 \times .9^2)^{1/2} = 0.045$. Thus, the Cell A value for $\varphi = 0.5$ is $0.01 + 0.045 = 0.055$, and the Cell D value $= 0.81 + 0.045 = 0.855$. Subtracting the product from the Cell B and C values for the independent case yields: $0.09 - 0.045 = 0.045$ for both cells. Calculating φ with these new cell values yields 0.5. The new cell values are the only ones that will yield the desired level of φ with these marginal probabilities.

[2] The relative informativeness of Cells B and C, which provide evidence against a positive relationship, is also determined by $p(X)$ and $p(Y)$ and is independent of the competing hypotheses. When $p(X) = p(Y)$, Cells B and C are equally informative. Cell B is more informative than Cell C when $p(X) < p(Y)$, and Cell C is more informative than Cell B when $p(X) > p(Y)$.

the X's and Y's people talk and think about in the real world are indeed rare. That is, does a variable that can either be present or absent tend to be absent more often than present? The answer will depend on the specific circumstances, but we believe that, for the vast majority of variables, the answer is yes. Most things are not red, most things are not mammals, most people do not have a fever, and so on. It is important to note that we are making a claim about how people use language, not about metaphysics. Imagine two terms, "X" and "not-X" (e.g., red things and non-red things, accountants and non-accountants), where there is no simple, non-negated term for not-X. Which would be the larger category, X or not-X? We believe that not-X will be the larger category in the vast majority of cases. It seems plausible that people have learned through a lifetime of experience that the presence of properties is rarer than their absence, and that therefore observing the joint presence of variables is usually more informative than observing their joint absence when determining whether the variables are related. A "bias" for Cell A observations in the laboratory might reflect deeply rooted tendencies that are highly adaptive outside the laboratory.

The four panels in Fig. 2 show the respective cells' informativeness ($|LLR_j|$) as a function of $p(X)$ and $p(Y)$, which were orthogonally manipulated between 0.1 and 0.9 in steps of 0.1 (resulting in 81 data points in each panel). H1 was $\rho = 0.1$ and H2 was $\rho = 0$. (The low ρ value for H1 was used because there are low upper bounds on ρ when one of $p(X)$ or $p(Y)$ is low and the other is high.) The cells' informativeness is determined by $p(X)$ and $p(Y)$. The top left panel shows that a Cell A observation is most informative when $p(X)$

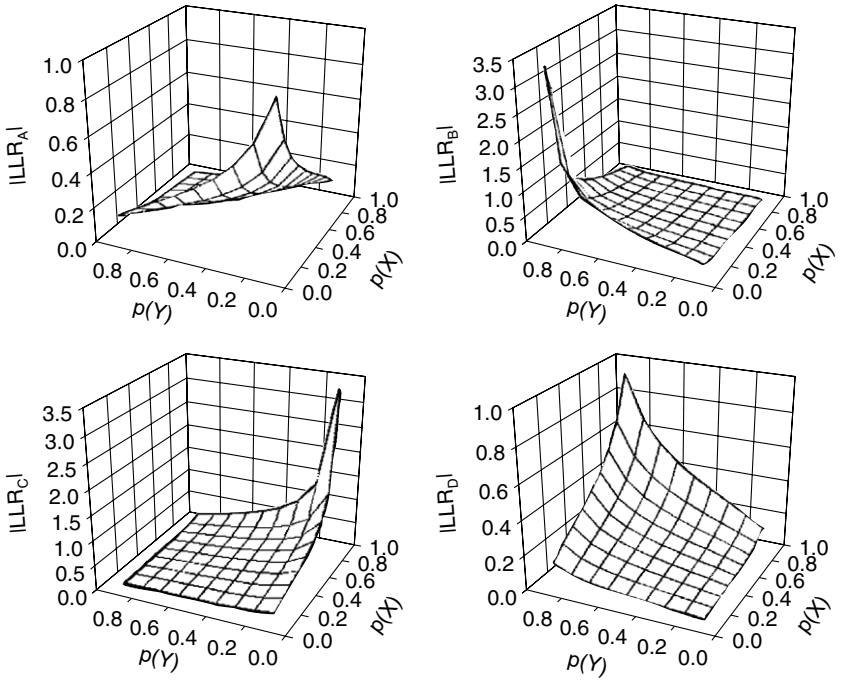

Fig. 2. Simulation results showing the log likelihood ratio ($|LLR_j|$) of each of the four cells as a function of p(X) and $p(Y)$.

and $p(Y)$ are both low; the top right panel shows that Cell B is most informative when $p(X)$ is low and $p(Y)$ is high; the bottom left shows that Cell C is most informative when $p(X)$ is high and $p(Y)$ is low; and the bottom right panel shows that Cell D is most informative when $p(X)$ and $p(Y)$ are both high. Though there are conditions under which each of the cells is most informative, the claim here is that X and Y tend to be rare. Under these conditions, Cell A is most informative, Cell D is least informative, and Cells B and C fall in between (see also Anderson, 1990, pp. 149–160).

Of course, one can make virtually any behavior appear rational by adding post hoc assumptions, and therefore caution is warranted. On the other hand, a simple normative account of a large literature should not be ignored. Furthermore, we do more in this article than posit that participants behave as though they assume that presence is rare. Later, we: (a) review recent findings showing that participants are highly sensitive to event rarity in inference tasks, (b) present new data demonstrating that Cell D is perceived as most informative when it is clear that *absence* is rare, and (c) use the general Bayesian approach to account naturally for the robust finding that prior beliefs influence judgments of covariation. First, though, more needs to be said about differences between the traditional and inferential views.

4. The traditional and inferential views of covariation assessment

Because the traditional view of covariation assessment is concerned with the direction and/or strength of the relation between the two variables, given the four cell frequencies, the usual claim is that all cells are equally important. Even within the traditional framework, however, a new A or D observation can have a different impact on φ (and Δp) depending on the marginal frequencies (relative to $N = A + B + C + D$), or $p(X)$ and $p(Y)$. A new A or D observation has the same impact on φ only when $p(X) + p(Y) = 1$ for the matrix at hand. Furthermore, a new A observation will have more impact than a D observation on φ whenever $p(X) + p(Y) < 1$ and will have less impact when $p(X) + p(Y) > 1$. For example, consider the matrix with A through D values of 1, 9, 9, and 81, respectively. For this matrix, $\varphi = 0$ and $p(X) + p(Y) = .1 + .1 = .2$. Adding one Cell A observation leads to a greater increase in φ than adding one Cell D observation: $\varphi = 0.08$ in the former case and 0.001 in the latter. Even though the traditional normative model can be differentially influenced by adding an observation to different cells when $p(X) + p(Y)$ differs from 1, participants are often shown many different matrices that control for marginal frequencies and, averaging across the matrices, it is found that Cell A has the largest influence on judgments (e.g. Levin et al., 1993; Mandel & Lehman, 1998).

An important difference between the traditional view and our inferential view is that $p(X)$ and $p(Y)$ in the former view depend entirely on the matrix at hand, whereas $p(X)$ and $p(Y)$ in the latter view are based on learning prior to receiving the current matrix of information (though these beliefs could be influenced by the new information). For example, imagine a participant presented with a matrix in which all four cell frequencies were 5, and thus $p(X) + p(Y) = 1$. According to the traditional view, the normative response is that there is no relation between X and Y and, furthermore, a subsequent A or D observation should have equal impact on judgment. This would not necessarily be true from the inferential viewpoint, however, depending on the participant's beliefs (perhaps implicit) about $p(X)$ and $p(Y)$ beyond the current matrix information. If a participant generally believes that presence is rare, and therefore $p(X) + p(Y) < 1$ (in the larger popu-

lation of interest), then the 5 Cell A observations are more informative than the 5 D observations and might well be "overweighted." Similarly, a subsequent A observation will have more impact than a subsequent D observation. In contrast, if the participant believed that the *absence* of *X* and *Y* was rare, then the 5 Cell D observations, and any subsequent Cell D observations, might have the largest impact on behavior.

A complication arises when comparing the traditional and inferential views in that they seem to imply that different dependent measures are appropriate for the task. The traditional view leads to asking participants questions such as "How strong is the relation between *X* and *Y*?", whereas the inferential view leads to questions such as "How strong is this evidence for a relation between *X* and *Y*?", or "How confident are you that *X* and *Y* are related?" The objection might be raised that traditional tasks ask about strength of relation, a descriptive (statistical) question, and participants should respond accordingly. But, of course, one of our main points is that participants view the task differently than experimenters, and what is important is how participants, not experimenters, construe the task (see also Hilton, 1990, 1995; McKenzie, 2003, 2005; Oaksford & Chater, 1994; Schwarz, 1996; cf. Stanovich, 1999; Stanovich & West, 2000). If it is naturally construed in the inferential manner put forth here, then it might be difficult for participants to ignore the differential informativeness of the cells when rating descriptive strength. Consistent with this, Wasserman et al. (1990) found that even participants who claimed that the four cells were equally important exhibited Cell A bias when reporting strength judgments (see also Mandel & Lehman, 1998). That is, these participants appear to have, in a sense, accepted the experimenters' traditional view of the task, but they nonetheless behaved in a manner consistent with the current inferential approach. Indeed, the experiments we report later will show that participants' ostensibly descriptive judgments are influenced by normatively relevant inferential factors.

5. Evidence regarding sensitivity to event rarity in inference tasks

We mentioned earlier that Anderson (1990) argued that Cell A "bias" is adaptive (from a Bayesian perspective) because the presence of variables is rare. Though important, this account remains a post hoc explanation of differential sensitivity to the cells. To show that the Bayesian account goes beyond that, it needs to be demonstrated that participants' covariation behavior is adaptable with respect to rarity. If there already exists evidence that participants are sensitive to event rarity in the predicted manner when making inferences in other tasks, this would increase the plausibility of the Bayesian account of covariation assessment. Indeed, recent results show that participants are sensitive to the rarity of observations in a variety of inference tasks (Feeney, Evans, & Clibbens, 2000; Green & Over, 2000; Green, Over, & Pyne, 1997; Kirby, 1994; McKenzie, 2004a; McKenzie & Amin, 2002; McKenzie & Mikkelsen, 2000; Oaksford & Chater, 1994, 2003; Oaksford, Chater, Grainger, & Larkin, 1997; Oaksford, Chater, & Grainger, 1999; Oaksford & Wakefield, 2003). Rarity appears to play a crucial role in human inference, and this makes sense given that rarity plays a crucial role in normative theories of hypothesis testing (Poletiek, 2001). In fact, results have indicated (and the experiments reported later in this article will provide further evidence) that even when experimenters present participants with abstract, unfamiliar materials in the hope of eliminating "real world" intrusions, participants merely fall back on default assumptions about what is rare.

One example demonstrating participants' sensitivity to rarity comes from McKenzie and Mikkelsen (2000), who asked participants to test hypotheses of the form, "If $X1$, then $Y1$." Each variable, X and Y, had exactly two levels ($X1$ and $X2$, $Y1$ and $Y2$). For example, some participants tested the hypothesis, "If a person has genotype A, then that person has personality type X," and they were told that everyone has either genotype A or B and either personality type X or Y. Participants were then presented with an $X1$ and $Y1$ observation and an $X2$ and $Y2$ observation (e.g., a person with genotype A and personality type X and a person with genotype B and personality type Y)—both of which support the hypothesis—and asked which provided stronger support. Decades of hypothesis-testing research have shown that participants overwhelmingly prefer confirming observations mentioned in the hypothesis ($X1$ and $Y1$ observations) to confirming observations that are not mentioned ($X2$ and $Y2$ observations). This phenomenon, which might be related to the preference for Cell A to Cell D information, has been referred to in various hypothesis-testing contexts as "confirmation bias," "matching bias," and "positive testing" (e.g. Evans, 1989; Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987; McKenzie, 1994, 2004b; see also McKenzie, 1998, 1999; McKenzie, Wixted, Noelle, & Gyurjyan, 2001). McKenzie and Mikkelsen (2000) also found this phenomenon—but only when the hypothesis being tested regarded unfamiliar variables and there was no information regarding the rarity of the observations. When told that $X1$ and $Y1$ were common relative to $X2$ and $Y2$, more participants correctly selected the $X2$ and $Y2$ observation as more supportive. Even stronger results were found when familiar variables were used and participants presumably knew which levels of the variables were rare. For example, participants testing the hypothesis, "If a person is HIV−, then that person is mentally healthy," were more likely to select the rare, and hence more informative, observation—a person who is HIV+ and psychotic—even though the observation was not mentioned in the hypothesis. The combination of familiar variables and a "reminder" that $X1$ and $Y1$ were common led participants to correctly select the $X2$ and $Y2$ observation more often than the $X1$ and $Y1$ observation, even though they were testing "If $X1$, then $Y1$."

Another example of the influence of rarity on behavior comes from Wason's selection task, in which participants are presented with four cards, each of which has either $X1$ or $X2$ on one side and $Y1$ or $Y2$ on the other. For each card, only one side is showing; $X1$, $X2$, $Y1$, and $Y2$ each face up on one card. Participants have to decide which cards to turn over to test the hypothesis, "If $X1$ appears on one side, then $Y1$ appears on the other." 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 to see if the rule is true or false? According to one interpretation of the rule, propositional logic dictates that the $X1$ and $Y2$ 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 $X1$ and $Y1$ cards (A and 2 in the example; 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; Over and Jessop, 1998) have shown that, from a Bayesian perspective (in which the available cards are sampled from a larger population to which the rule is to be generalized), the $X1$ and $Y1$ cards are the most informative for testing the rule—if one assumes that $X1$ and $Y1$ are rare relative to $X2$ and $Y2$ (the "rarity assumption"). As predicted by this account, participants presented with a reduced array selection task, where only the $Y1$ and $Y2$ cards are present, were more likely

to select the *Y*2 card as it became rarer (Oaksford et al., 1997; see also Green & Over, 2000; Green et al., 1997; Oaksford & Chater, 2003; Oaksford et al., 1999; but see Oberauer, Wilhelm, & Diaz, 1999). Note that, like the present article, Oaksford and Chater (1994) offered a Bayesian account of a task traditionally seen in non-Bayesian normative terms.

McKenzie, Ferreira, Mikkelsen, McDermott, and Skrable (2001) provided empirical evidence supporting the rarity assumption: They found that participants tended to phrase conditional statements (or hypotheses) in terms of rare, rather than common, events. Thus, people might consider mentioned confirming observations most informative, or consider turning over the mentioned cards most informative, because they usually *are* most informative, at least from a Bayesian perspective.

While these results are encouraging with respect to reversing the Cell A bias, it should be kept in mind that the covariation task is different from the selection task and from McKenzie and Mikkelsen's (2000) hypothesis-testing task. In the selection task, participants must select cards to turn over, and they cannot be sure what will be on the other side. McKenzie and Mikkelsen's (2000) hypothesis-testing task is more similar to a covariation task, but an important difference is that the variables to be tested had symmetric levels such as "genotype A" and "genotype B," rather than the traditional asymmetric levels of "present" and "absent" as in most covariation tasks. Much evidence has shown that participants have trouble reasoning with variables that are absent or negated (e.g. Wason & Johnson-Laird, 1972; Evans et al., 1993), making it unclear whether McKenzie and Mikkelsen's (2000) hypothesis-testing results will generalize to covariation assessment. In addition, instructions differ between the tasks. In hypothesis-testing tasks and selection tasks, participants are often asked to test "if-then" statements, whereas in covariation tasks, participants are asked to assess the relationship between variables.

Despite differences between the tasks, people appear highly sensitive to rarity when making inferences, as they should be according to the Bayesian perspective. Evidence from different areas of research support the prediction that manipulating the rarity of the presence vs. absence of the variables in a covariation task will influence perceived cell informativeness. In particular, participants might deem Cell D more informative when absence, rather than presence, is rare.

6. Experiment 1

Our first experiment manipulated the rarity of the presence of two unfamiliar variables through a learning manipulation before presenting participants with various covariation tasks. Half of the participants learned that the presence of both variables was rare, and half learned that their absence was rare. The prediction was that, relative to the presence rare group, the absence rare group would view the variables' joint presence (Cell A) as less informative and their joint absence (Cell D) as more informative. As mentioned earlier, we will focus on behavior with respect to Cells A and D because these two cells have traditionally been shown to have the largest and smallest influence, respectively, on behavior.

6.1. Method

Participants were 160 University of California, San Diego (UCSD) students who received partial credit for psychology courses. The experiment took place on computer. Participants were told to imagine that they were researchers studying whether there was

a relation between a certain genotype and a particular personality trait. They first learned how common the genotype was by seeing 50 people's "profiles." Each profile stated whether the person had the genotype (yes or no), but there was no information regarding whether the personality trait was present (indicated by "?"). Participants were told that the purpose was to learn how common the genotype was, and that they would subsequently be asked how many of the 50 people had the genotype. For half of the participants, the genotype was present in 5 out of 50 people (10%) and for the other half of the participants, the genotype was present in 45 out of 50 people (90%). Each profile (e.g., "Genotype: Yes/Trait: ?") was presented for 2000 ms followed by a 500 ms delay (blank screen) and a tone. After seeing the 50 profiles, all participants estimated the number of people out of 50 who had the genotype. They were subsequently told the correct number of people with the genotype.

Participants then learned about the prevalence of the personality trait. They were again shown the same 50 people's profiles (said to be in a different random order), but this time the presence vs. absence of the personality trait was visible (yes or no), but not the information regarding the genotype ("?"). They were again told that they would estimate the number of people with the trait after viewing the profiles. After providing estimates, participants were told the correct number of people exhibiting the trait. If the genotype was present in 10% of the people, the trait was also present in 10%; if the genotype was present in 90% of the people, the trait was also present in 90%. Thus, the genotype and the trait were either both rare or both common.

After learning about the prevalence of the genotype and trait, participants were told that they were ready to assess the relation between them. Because the genotype and trait could be either present or absent, there were four possible complete profiles. Each of the four was then listed (e.g., "Genotype: Yes/Trait: Yes"). The order of the four profiles was A (yes/yes) through D (no/no) for half of the participants and D through A for the other half.

Participants were then presented with results said to be from two different random samples of 8 people from the group of 50. One sample had A through D values of 5, 1, 1, 1, respectively, and the other had values of 1, 1, 1, 5.[3] Participants selected the sample showing the stronger relation between the genotype and the trait.

The next task presented two individual profiles, one with both the genotype and the trait present (a Cell A observation), and one with both absent (Cell D). Participants selected the profile that provided stronger evidence for a relation between the genotype and trait.

The final task presented participants with what was said to be a random sample of 8 people from the group of 50. Half of the participants in each condition saw a matrix with A through D values of 5, 1, 1, 1, and half saw a matrix with A through D values of 1, 1, 1, 5. They were asked to report how strong the relationship was "*for these 8 people*", which emphasized the descriptive aspect of the question. They reported strength on a scale of 0 to 20, with 0 = "no relation," 10 = moderate relation", and 20 = "perfect relation."

[3] Some readers might notice an oversight on the part of the experimenters: The first sample is impossible for the presence rare group—there are, e.g., 6 people in the sample with the genotype—and the second is impossible for the presence common group. The large value should have been 4, not 5. However, only one participant (out of 160) mentioned this in post-experimental comments.

6.2. Results and discussion

Participants' estimates of the number of times the genotype and the trait were present during learning were quite accurate. For the presence rare group, where the genotype and trait were present in 5 out of 50 profiles, the mean estimates were 5.5 and 5.2, respectively. For the absence rare group, where the genotype and trait were present in 45 out of 50 profiles, the mean estimates were 41.1 and 44.6.

When presented with the two matrices, one with A through D values of 5, 1, 1, 1 and one with values of 1, 1, 1, 5, and asked which showed a stronger relation between the genotype and the trait, 25% of participants in the presence rare group selected the large Cell D matrix. However, 41% of the absence rare group selected the large Cell D matrix. This difference is significant: χ^2 (1, $N = 160$) = 4.8, $p = .029$. From a Bayesian viewpoint, Cell D observations are more informative when the variables' absence is rare, and participants' preferences shifted in the direction of the large Cell D matrix under this condition.

When presented with the two individual profiles, one with the genotype and trait both present (Cell A) and one with them both absent (Cell D), and asked which provided stronger evidence for a relation between the variables, 16% of the presence rare group selected the joint absence profile. However, 39% of the absence rare group selected the joint absence profile: $\chi^2(1, N = 160) = 10.2$, $p = .001$. Preference for the joint absence profile increased when absence was rare, as predicted by a Bayesian approach.

Fig. 3 shows the results for the task in which participants rated the strength of the relation between the genotype and the trait. The solid line corresponds to the presence rare group and the broken line to the absence rare group. Both groups rated the large Cell A matrix (i.e., values of 5, 1, 1, 1) as showing a stronger relation than the large Cell D matrix, but the difference between the ratings is much smaller for the absence rare group. That is, compared to the presence rare group, the absence rare group gave relatively equal weighting to Cells A and D. A Rarity (presence rare vs. absence rare) × Matrix (large Cell

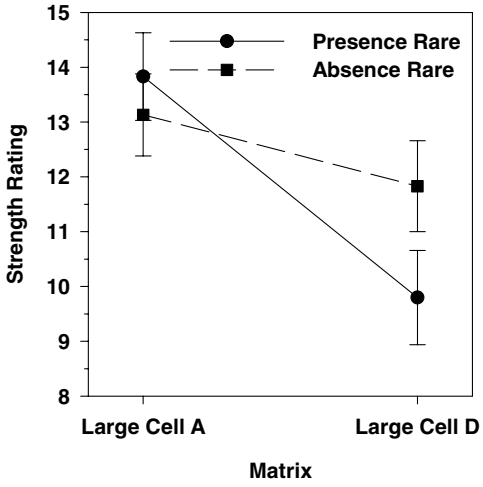

Fig. 3. Experiment 1: Strength-of-relation ratings as a function of whether presence was rare vs. common and whether Cell A vs. Cell D was large. Standard error bars are shown.

A vs. large Cell D) ANOVA on the ratings revealed a main effect of Matrix ($F(1,156) = 10.77$, $p = .001$), but the predicted interaction was only marginally significant ($F(= 2.82$, $p = .095)$).

Note that the predictions for all three tasks were based on participants' beliefs or expectations about rarity beyond the specific matrix or cell information presented to them. That is, a given cell observation or matrix of observations was evaluated in the context of the rarity of the variables' presence more generally. This represents an important departure from the traditional view of covariation assessment.

Nonetheless, one might contend that the results were somewhat weak. Making absence rare did increase the perceived informativeness of Cell D, but a majority of participants still preferred the large Cell A matrix and the Cell A profile, and rated the large Cell A matrix as stronger. However, recall that we are assuming that participants have a general, strong tendency to treat presence as rare because that is the norm outside the laboratory. Thus, it is unlikely that 2-min presentations of unfamiliar variables will completely override a bias for treating presence as rare. One would expect behavior to move in the predicted direction, though, just as we found (see also Oaksford et al., 1997). Our next experiment attempted to demonstrate stronger effects.

7. Experiment 2

Is it possible to get participants to reverse their preference for Cell A over Cell D? The most promising way would be to use variables that participants are familiar with. Ideally, participants would already know how common the levels of each variable are. Tapping into participants' real-world knowledge about rarity can have large effects on behavior in the direction predicted by the Bayesian account (McKenzie, in press; McKenzie & Mikkelsen, 2000). In Experiment 2, one group assessed variables that they were presumably familiar with in terms of how common each level was. For half of these *concrete* participants, the rare level of each variable was labeled "yes" and the common level "no," whereas the opposite was true for the other half. A second group assessed variables that they were unfamiliar with. Half of this *abstract* group also saw one level of each variable labeled "yes" and the other "no", whereas the labels were switched for the other half. The prediction was that the abstract group would consider the observation labeled yes/yes more informative, regardless of which observation it referred to, but that the concrete group would consider the rare observation more informative, regardless of whether it was labeled yes/yes or no/no. In other words, the concrete group is predicted to consider Cell D more informative than Cell A when the former is rare and the latter is common.

7.1. Method

Participants were 306 UCSD students, half of whom were recruited in the same manner as in Experiment 1, and half of whom were recruited by signs posted on campus. The former group received partial course credit and the latter received monetary compensation for participation. Because this experiment did not require a learning session, it took the form of a paper-and-pencil questionnaire rather than a computer program.

Participants in the *concrete condition* were told to imagine that they worked at a large high school and were trying to uncover factors that determine students' "high school outcome": Whether they drop out or graduate. Though very few students drop out, they (the

participants) would like to see all students graduate if possible. By identifying factors that help predict who will drop out and who will graduate, those likely to drop out could be identified and preventive measures could be taken. The factor being examined was students' "emotional status." All students were said to undergo a thorough psychological examination during their freshman year and categorized as either emotionally disturbed or emotionally healthy. They were told that very few students are emotionally disturbed and almost all are emotionally healthy. Though we thought that participants would assume that each of dropping out and being emotionally disturbed is relatively rare, we nonetheless reinforced this in the instructions.

The concrete participants were told that they had access to the records of former students to find out if there was a relationship between students' emotional status and high school outcome. Half of these participants (the presence rare group) were told that each record listed whether the student was emotionally disturbed (yes or no) and whether the student dropped out (yes or no). Thus, the presence (i.e., the "yes" level) of each variable was rare, making a Cell A observation rarer than a Cell D observation. The four possible categories of students were then listed (e.g., Emotionally disturbed: Yes/Dropped out: Yes). Order of the categories was A through D for half of the group and D through A for the other half. The different orders were maintained for all the tasks.

This presence rare group was then presented with the files of two former students (said to be randomly sampled), one of whom was emotionally disturbed and dropped out (Cell A) and one of whom was not emotionally disturbed and did not drop out (Cell D). Participants selected the student that provided stronger support for a relationship between emotional status and high school outcome. A second task presented what were said to be two different random samples of 9 former students. One sample had *A* through *D* values of 6, 1, 1, 1 and the other had values of 1, 1, 1, 6. Participants selected the sample that provided stronger support for a relation between emotional status and high school outcome. The final task presented participants with either the large Cell A matrix or the large Cell D matrix and asked them to rate the strength of the relation *for those 9 people*. A 21-point scale was used, with 0 = "no relationship" and 20 = "perfect relationship."

For the other half of the concrete participants, the students' records listed whether they were emotionally healthy (yes or no) and whether they had graduated (yes or no). Thus, the absence of each of these variables was rare, making Cell A more common than Cell D. These absence rare participants were presented with the same two former students as in the other group, but labeled differently: One was emotionally healthy and graduated (Cell A) and one was not emotionally healthy and did not graduate (Cell D). Note that Cell A in one group was logically equivalent to Cell D in the other; they were simply labeled differently. The participants selected the student that provided stronger support for a relationship between emotional status and high school outcome. The cells in the choice and ratings tasks were also simply relabeled relative to the presence rare group.

The *abstract group* was given an essentially identical task, but the content was changed so that they would not have experience with how rare or common the levels of the variables were. They were told that they were trying to uncover the factors that determine whether people have personality type *X* or personality type *Y*. Everyone was said to have one type or the other. The factor they were currently examining was genotype; everyone was said to have either genotype A or genotype B. They were trying to find out if there was a relationship between genotype and personality type.

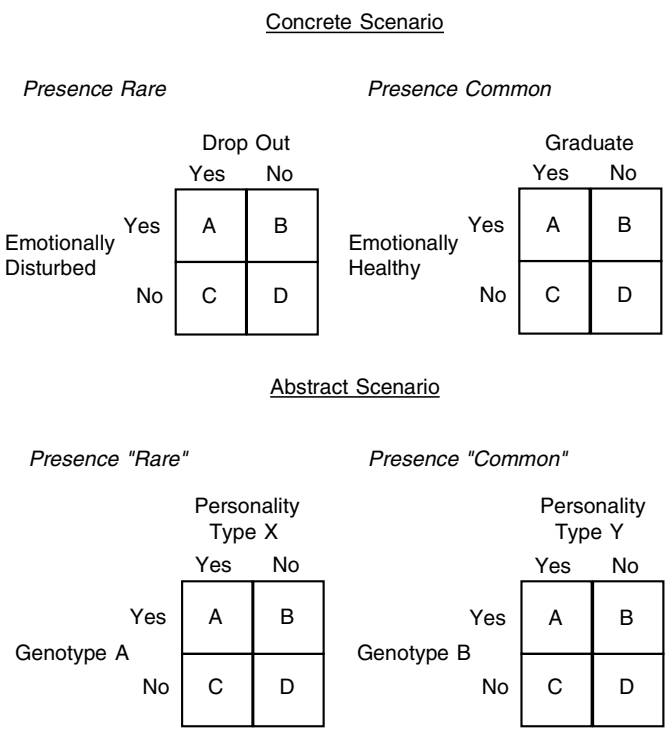

Fig. 4. The design of Experiment 2.

Half of the abstract group saw individuals' records that stated whether each person had genotype A (yes or no) and whether each had personality type *X* (yes or no). For the other half, the records were in terms of whether each person had genotype B (yes or no) and personality type *Y* (yes or no). Thus, for example, a person with genotype A and personality type *X* was a yes/yes observation for the former group of abstract participants and was a no/no observation for the latter group. The subsequent tasks were identical to those of the concrete group: They chose between a Cell A and a Cell D observation, chose between the large Cell A matrix and the large Cell D matrix, and they rated the strength of the relationship between genotype and personality type for either the large Cell A matrix or the large Cell D matrix.

To maintain consistency across the concrete and abstract scenarios, and to provide benchmarks for comparison, we will refer to the genotype A and personality type *X* levels as "rare" and the genotype B and personality type *Y* levels as "common," although the labels are arbitrary. We will put the terms in quotation marks when discussing the abstract group as a reminder to the reader. The design of the experiment is illustrated in Fig. 4.

7.2. Results

Preliminary analyses showed that neither the order of the cell information (A to D vs. D to A) nor how participants were compensated (course credit vs. monetary payment) had

significant effects on the dependent measures. These variables were not included in the analyses below.

The percentage of participants selecting a Cell D observation as more informative than a Cell A observation is shown in the top panel of Fig. 5. For the concrete group, few considered Cell D stronger evidence of a relation when presence was rare, but most preferred Cell D when absence was rare. That is, the emotionally disturbed/dropped out observation was considered most informative regardless of whether it was labeled "Emotionally Disturbed: Yes/Dropped Out: Yes" or "Emotionally Healthy: No/Graduated: No." In contrast, few abstract group participants selected Cell D in either condition. They simply tended to select the observation labeled yes/yes. For example, most participants in the abstract group selected the "Genotype A: Yes/Personality Type *X*: Yes" observation as most informative in one condition, but few selected the logically equivalent "Genotype B: No/Personality Type *Y*: No" observation in the other condition.

A Scenario (concrete vs. abstract) × Rarity (presence rare vs. absence rare) log-linear analysis was performed on the number of participants selecting Cell A vs. D. There were

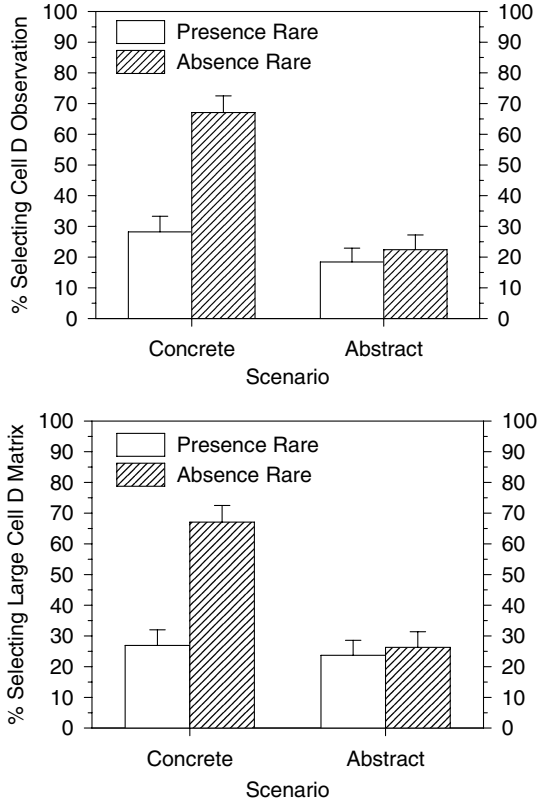

Fig. 5. Experiment 2: The top panel shows the percentage of participants selecting a Cell D observation over a Cell A observation as providing stronger support of a relation as a function of whether the scenario was concrete vs. abstract and whether presence was rare vs. common. The bottom panel shows the percentage of participants selecting the large Cell D matrix over the large Cell A matrix as indicating stronger support of a relation. Standard error bars are shown.

main effects of Scenario, with more concrete participants selecting Cell D ($p < .001$), and Rarity, with more participants selecting Cell D when absence was rare ($p < .001$). Both of these effects are explained by the interaction ($p = .009$): Concrete participants tended to select Cell A when presence was rare and Cell D when absence was rare, whereas abstract participants tended to always select Cell A.

The results for the task in which participants were asked whether the large Cell A matrix (6, 1, 1, 1) or the large Cell D matrix (1, 1, 1, 6) provided stronger evidence of a relation between the variables are shown in the bottom panel of Fig. 5. The concrete group rarely selected the large Cell D matrix when presence was rare, but often selected it when absence was rare. In contrast, the abstract group rarely selected the large Cell D matrix in either condition. A Scenario (concrete vs. abstract) × Rarity (presence rare vs. absence rare) log-linear analysis showed main effects of Scenario, with more concrete participants selecting the large Cell D matrix ($p < .001$), and Rarity, with more participants selecting the large Cell D matrix when absence was rare ($p < .001$). Most important was the interaction ($p = .002$): Concrete participants tended to select the large Cell A matrix when presence was rare and select the large Cell D matrix when absence was rare, whereas abstract participants tended to always select the large Cell A matrix.

The results of the ratings task are shown in Fig. 6. The left panel shows the results for the concrete group. The left pair of columns in the panel shows the ratings when the large cell corresponded to the rare observation (an emotionally disturbed drop out). The first column corresponds to labeling the rare level as "yes" and the second to labeling it as "no." In other words, these two matrices were logically identical but labeled differently. The virtual equivalence between the columns shows that labeling had essentially no effect on the ratings. The same is largely true for the matrices in which the large cell corresponded to the common observation (an emotionally healthy graduate), shown by the next pair

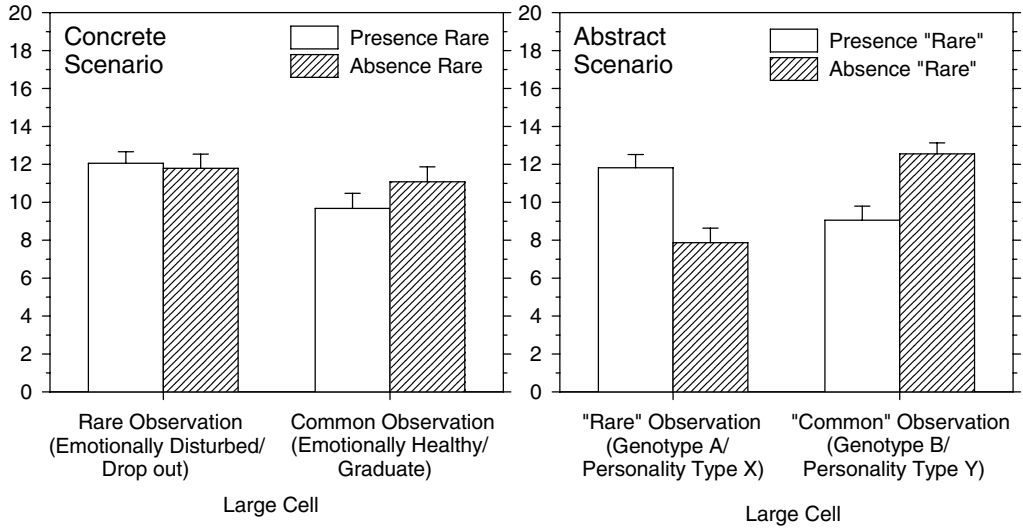

Fig. 6. Experiment 2: The left panel shows strength-of-relation ratings for the concrete scenario as a function of whether presence was rare vs. common and whether the matrix's large cell was a rare vs. common observation. The right panel shows analogous results for the abstract scenario. Standard error bars are shown.

of columns. Ratings were similar regardless of whether the common observation was labeled as no/no or yes/yes (first and second columns, respectively).

The right panel in Fig. 6 shows the ratings for the abstract group. The first pair of columns in this panel corresponds to the two matrices in which the "rare" observation (genotype A/personality type *X*) was the large cell. Whether this observation was labeled yes/yes (the first column) or no/no (second column) had a large effect on ratings, with higher ratings when the large cell was yes/yes. The final pair of columns shows ratings for the two matrices in which the "common" observation (genotype B/personality type *Y*) was the large cell. The first of these columns corresponds to the no/no labeling and the second to the yes/yes labeling. Again, labeling the same observation differently affected ratings, with the yes/yes labeling leading to higher ratings.

A Scenario (concrete vs. abstract) × Rarity (presence rare vs. absence rare) × Matrix (large Cell A vs. large Cell D) ANOVA on the ratings revealed three interactions. The first, Scenario × Matrix, was due to concrete participants rating the large rare observation matrix higher (the first pair of columns) than the large common observation matrix (second pair; 11.9 vs. 10.4, respectively), whereas the opposite was true for the abstract participants (9.8 vs. 10.8; $F(1,298) = 6.06$, $p = .014$). The interaction per se is not meaningful given the arbitrariness of "rare" and "common" in the abstract group. However, the difference in ratings for the concrete group is meaningful (and reliable: $t(152) = 2.10$, $p = .037$): From the current normative perspective, the matrices in which the rare observation is the large cell provide stronger support for a relation than do the matrices in which the common observation is the large cell. This finding provides further evidence that the concrete participants responded in a qualitatively Bayesian manner (even in this ostensibly descriptive strength-rating task). The second interaction was between Rarity and Matrix, essentially indicating that, collapsing across scenarios, large Cell A matrices were given higher ratings than large Cell D matrices ($F = 20.10$, $p < .001$). Most important was the predicted 3-way interaction ($F = 8.14$, $p = .005$): Concrete participants were largely unaffected by whether a given observation was labeled yes/yes vs. no/no, but the abstract group was affected. In terms of the figure, the two columns making up each pair are about the same height for the concrete group, but they are different for the abstract group. Concrete participants did not attend to labeling, but instead attended to whether the large cell observation was rare or common. In contrast, the abstract group's ratings were driven by whether the large cell was labeled yes/yes or no/no. Contrasts revealed no differences between the heights of the two columns making up each pair for the concrete group (both ps $> .2$), but the columns making up each of the two abstract pairs differed from each other (both ts > 3.6, ps $< .001$).

7.3. Discussion

The results for the abstract condition replicated previous findings in the literature that Cell A is seen as much more informative than Cell D. Similarly, the same pattern of results was found for the concrete group when they knew that presence was rare. However, when the labels of the levels of the concrete variables were reversed such that *absence* was rare, participants considered Cell D most informative. This was true both when individual cells and when different matrices were compared in terms of informativeness. To our knowledge, this is the first demonstration of a reversal of Cell A "bias." The results provide strong evidence for the hypothesis that the robust Cell A "bias" demonstrated over the

past four decades stems from (a) participants' inferential approach to the task, and (b) their default assumption (perhaps implicit) that presence is rare. When there is good reason to believe that absence is rare, Cell D is deemed more informative, just as the Bayesian approach predicts.

The results regarding the strength-of-relation ratings showed that even explicitly *descriptive* questions about the matrix at hand are influenced by *normatively relevant inferential information* that goes beyond the information contained in the matrix. This is important because traditional covariation research, which has assumed a descriptive (statistical) normative model, has found Cell A bias when asking participants descriptive questions. Our results indicate that participants take an inferential approach to covariation tasks, and that this influences responses to even purely descriptive questions.

It is possible that participants in the concrete condition mentally recoded joint absence in terms of joint presence when absence was rare. That is, these participants might have recoded "not emotionally healthy and did not graduate" as "emotionally disturbed and dropped out", effectively transforming Cell D into Cell A. This would not be evidence against our position that sensitivity to rarity is driving perceived cell informativeness (since such recoding would presumably take place only for rare, jointly absent events), but it does suggest an interesting process by which the demonstrated Cell D "bias" might have occurred.

Finally, note that we are explaining both the concrete and the abstract group in terms of their sensitivity to rarity: The former exploited real-world knowledge about which observations were rare, and the latter exploited knowledge about how labeling indicates what is (usually) rare.

8. Prior beliefs and covariation assessment

Our focus thus far has been on the role of rarity in covariation assessment, which led to the emphasis on individual cells and the likelihood portion of Bayes' theorem. However, the hallmark of a Bayesian approach is to incorporate prior beliefs that the hypotheses are true (e.g., that X and Y are related vs. unrelated). For example, to the extent that one does not believe that there is a relationship between the variables in question, the stronger the subsequent evidence required before believing that there is a relationship.

The following is Bayes' theorem in odds form:

$$P(E|\mathrm{H1})/p(E|\mathrm{H2}) \times p(\mathrm{H1})/p(\mathrm{H2}) = p(\mathrm{H1}|E)/p(\mathrm{H2}|E),$$

where E corresponds to evidence, or data. In this context, the data are Cell A, B, C, and/or D frequencies. The first ratio, $p(E|\mathrm{H1})/p(E|\mathrm{H2})$, is the likelihood ratio discussed earlier. The second ratio, $p(\mathrm{H1})/p(\mathrm{H2})$, is the prior odds, consisting of the prior probabilities (i.e., before receiving E) that each hypothesis is true. Multiplying them together results in the posterior odds, $p(\mathrm{H1}|E)/p(\mathrm{H2}|E)$, shown on the right. This ratio represents the normative odds that H1 rather than H2 is true, given E. It is perhaps worth pointing out that, though we discussed the likelihood ratio in the context of a single observation, E can also correspond to multiple observations (e.g., an entire matrix). The equation makes explicit that both the likelihood ratio discussed earlier and the prior probabilities discussed above are important from a Bayesian perspective.

In fact, much evidence indicates that prior beliefs influence judgments of covariation (Alloy & Tabachnik, 1984; Chapman & Chapman, 1967, 1969; Crocker, 1981; Jennings

et al., 1982; Nisbett & Ross, 1980; Peterson, 1980). Alloy and Tabachnik (1984) reached this conclusion after reviewing a large number of experiments on both humans and nonhuman animals. For example, Peterson (1980) argued that participants tend to report a positive relationship between noncontingent variables in laboratory experiments because they have strong prior beliefs that there will be a relationship present in the experimental materials. He showed that mentioning in the instructions the possibility that there might be no relationship between the variables, which presumably increased participants' prior probability that no relationship would occur, greatly increased the chances that participants reported that there was no contingency.

Chapman and Chapman (1967) showed especially strong effects of prior beliefs in covariation assessment. They presented participants with 45 Draw-a-Person pictures, each randomly paired with statements about the symptoms exhibited by the patient who supposedly drew the picture. The participants were subsequently asked which picture characteristics and patient characteristics were associated. Interestingly, these naïve participants claimed to see the same relationships that clinicians claimed to see (e.g., participants falsely claimed that "suspicious" patients drew people with atypical eyes). The authors showed further that participants tended to see positive relationships between certain patient and picture characteristics even when the objective relationships were negative, and Chapman and Chapman (1969) showed that strong preconceptions hindered participants' ability to detect other (unexpected) relationships that really were there.

Alloy and Tabachnik (1984) discussed these and many other examples indicating that participants' judgments of covariation are influenced by their prior beliefs about the nature of the relationship presented to them. Seen only in the context of a given experiment, where the information is usually contrived, incorporating prior beliefs might seem odd. Seen in its broader context, though, where current information is part of a constant flow, past experience is relevant, and incorporating prior beliefs is necessary for being accurate. It might be the case that the prior beliefs are inaccurate (as implied by Chapman & Chapman, 1967, 1969), but the normative issue for our purposes is how one should behave *given* prior beliefs and new information. The Bayesian perspective claims that both are relevant; to ignore prior beliefs is an error.

Interestingly, Alloy and Tabachnik (1984) were in the position of having to argue that it makes good sense to incorporate prior beliefs when assessing covariation. Due to the entrenchment of φ as the normative model, previous authors had claimed that the influence of prior beliefs was nonnormative (Crocker, 1981; Nisbett & Ross, 1980). As mentioned earlier, however, Alloy and Tabachnik did not take an explicitly Bayesian stance and were left distinguishing between being "accurate"—behaving in accord with φ or Δp—and being "rational"—incorporating prior beliefs. Working within the traditional framework, incorporating prior beliefs is an awkward layer to be added to the normative model. In contrast, the Bayesian view accommodates prior beliefs naturally; they are *expected* to influence behavior. Thus, this is another instance where knowledge beyond the specific matrix information influences responses that are ostensibly descriptive, and in a manner predicted by the Bayesian account. Participants appear to approach covariation tasks in an inferential, not a descriptive, manner. A qualitative Bayesian approach can parsimoniously account for a variety of covariation findings. Robust "biases" are no longer anomalies to be explained, but are instead the result of normative principles.

9. Human rationality

Because our analysis and results indicate that what have been traditionally viewed as covariation "errors" can be seen as the result of normative principles, the article contributes to the current debate regarding human rationality (e.g. Anderson, 1991; Cohen, 1981; Gigerenzer, 1991, 1996; Kahneman & Tversky, 1996; Oaksford & Chater, 1994, 1996; Stanovich, 1999; Stanovich & West, 2000). The topic has generated heated exchanges, and we want to be clear about what we believe our analysis and data do—and do not—imply.

Our general claim is that people's behavior in covariation tasks is most usefully seen as stemming from an inferential, Bayesian approach. The two most robust findings in the covariation literature—a preference for joint presence over joint absence and the influence of prior beliefs—along with our two experiments, support this idea. These results, together with those reported by Griffiths and Tenenbaum (2005), provide considerable support for a Bayesian perspective.

Furthermore, Bayesianism is considered by many (though not all) statisticians and philosophers to be the optimal approach to updating beliefs in light of new information (e.g. Earman, 1992; Horwich, 1982; Howson & Urbach, 1989). The combination of Bayesianism's formidable normative status and its ability to explain covariation behavior is both interesting and important. However, some might argue that, even if a normative Bayesian approach does explain covariation behavior, participants are nonetheless making errors by incorporating prior beliefs and exhibiting a bias for joint presence. That is, if participants are asked to report the degree of association between the variables, exhibited only by the data in the matrix, then they should behave accordingly. To behave differently is to behave irrationally. We feel that this line of reasoning is presumptuous and, ultimately, counterproductive to the goal of *understanding* human behavior. Our analysis indicates that what have traditionally been viewed as errors need not be viewed as such. One could continue calling the preference for joint presence in a typical, abstract laboratory task an error, but we see little point in doing so. *Should* one approach a covariation task in the traditional manner assumed by psychologists? Our cognitive system apparently approaches covariation a different way—a way that puts the task into a larger inferential framework, is influenced by how the world usually works, and is normatively justifiable.

Although we think it is difficult to make normative claims about which normative theory should be applied to a given real-world task, there might nonetheless be situations in which people give more weight to joint presence than joint absence, or are influenced by prior beliefs, when it is generally agreed that to do so is a mistake. This is not so paradoxical. In the courtroom, for example, it is considered inappropriate for prior beliefs about a defendant's guilt to influence jurors, even though this goes against Bayesian principles (Tribe, 1971; see also Koehler, 1992; Koehler & Shaviro, 1990). However, our analysis indicates that it should not be assumed that people need "help" when assessing whether and how variables are related. "Debiasing" (i.e., encouraging people to calculate φ or Δp) might have no effect or even lead to worse performance if the circumstances are such that the Bayesian approach is reasonable.

We have already mentioned that we are not claiming that people are Bayes-optimal processors of information, neither in general nor in covariation tasks in particular. Nonetheless, the process people do use to assess covariation is "Bayes-like" in that people are

sensitive to prior beliefs and to the rarity of data. Indeed, we suspect that following these two important principles alone will go a long way toward mimicking a Bayesian response (e.g. McKenzie, 1994). Though we do not specify a process model of covariation assessment, our account is useful because it provides a unified account of an important behavior. It answers some otherwise unanswered questions.

Finally, it is important to reiterate that people's covariation behavior in the laboratory seems strongly influenced by the conditions under which they usually operate. Following Anderson (1990), we have argued that a variable's presence is usually rare, and that is why participants generally consider Cell A more informative than Cell D. It is of interest that such effects have occurred despite experimenters' attempts to decontextualize tasks to eliminate real-world influences. Experimenters generally attempt to use tasks that will not invoke participants' idiosyncratic differences. Our research indicates that using impoverished stimuli merely leads participants to fall back on default assumptions about important task parameters and, furthermore, these default assumptions appear to coincide with what one would reasonably expect in the real world. Taking into account real-world conditions, combined with normative principles that make sense under these conditions, can help explain why people behave as they do.

Appendix A.

In this appendix, we show that the joint presence of two binary variables (X and Y) is more informative than their joint absence with respect to determining whether the variables are related if $P(X) < 1 - P(Y)$. The measure of informativeness we use is the absolute log likelihood ratio: $|\mathrm{LLR}_j| = \mathrm{Abs}\,(\log_2[p(j|\mathrm{H1})/p(j|\mathrm{H2})])$, where j corresponds to Cell A, B, C, or D. Let H1 and H2 be mutually exclusive and exhaustive hypotheses about the degree of relatedness between X and Y, which have levels of presence and absence. $P(X)$ and $P(Y)$ are the probability of the respective variable being present (rather than absent) in the population of interest and do not differ between H1 and H2. Let $P(\mathrm{A})$ be the probability of a Cell A observation (joint presence of X and Y), $P(\mathrm{B})$ be the probability of a Cell B observation (the presence of X and the absence of Y), and $P(\mathrm{D})$ be the probability of a Cell D observation (joint absence of X and Y). We start with the equivalence of $|\mathrm{LLR_A}|$ and $|\mathrm{LLR_D}|$

$$|\mathrm{LLR_A}| = |\mathrm{LLR_D}|,$$

$$1 = |\mathrm{LLR_A}|/|\mathrm{LLR_D}|,$$

$$1 = |\mathrm{LLR_A}/\mathrm{LLR_D}|.$$

We can drop the absolute value bars because $\mathrm{LLR_A}$ and $\mathrm{LLR_D}$ will either both be positive or both be negative—given our assumptions that H1 and H2 are mutually exclusive and exhaustive and that $P(X)$ and $P(Y)$ do not change under H1 and H2—so their resulting ratio will always be positive.

$$1 = \mathrm{LLR_A}/\mathrm{LLR_D}.$$

Let the likelihood ratio for Cell j (LR_j) be $p(j|\mathrm{H1})/p(j|\mathrm{H2})$

$$1 = \log_2(\mathrm{LR_A})/\log_2(\mathrm{LR_D}),$$

$$0 = \log_2(\mathrm{LR_A}/\mathrm{LR_D}),$$

$$1 = \mathrm{LR_A}/\mathrm{LR_D}.$$

By definition

$$1 = [P(A|\mathrm{H1})/P(A|\mathrm{H2})]/[P(D|\mathrm{H1})/P(D|\mathrm{H2})].$$

Simple algebra leads to

$$1 = P(A|\mathrm{H1})P(D|\mathrm{H2})/[P(D|\mathrm{H1})P(A|\mathrm{H2})].$$

Using the identities $p(A|\mathrm{H}) = P(X) - P(B|\mathrm{H})$ and $P(\mathrm{D}|\mathrm{H}) = 1 - P(Y) - P(B|\mathrm{H})$ and again assuming that $P(X)$ and $P(Y)$ do not change under H1 and H2, we can express the conditional cell probabilities in the following way:

$$1 = ([P(X) - P(B|\mathrm{H1})][1 - P(Y) - P(B|\mathrm{H2})])/([1 - P(Y) - P(B|\mathrm{H1})][P(X) - P(B|\mathrm{H2})]).$$

Multiplying leads to

$$\begin{aligned} 1 = {} & [P(X) - P(X)P(Y) - P(X)P(B|\mathrm{H2}) - P(B|\mathrm{H1}) + P(Y)P(B|\mathrm{H1}) \\ & + P(B|\mathrm{H1})P(B|\mathrm{H2})]/[P(X) - P(X)P(Y) - P(X)P(B|\mathrm{H1}) - P(B|\mathrm{H2}) \\ & + P(Y)P(B|\mathrm{H2}) + P(B|\mathrm{H1})P(B|\mathrm{H2})]. \end{aligned}$$

Setting the numerator equal to the denominator and subtracting like terms results in

$$\begin{aligned} 1 = {} & [-P(X)P(B|\mathrm{H2}) - P(B|\mathrm{H1}) + P(Y)P(B|\mathrm{H1})]/[-P(X)P(B|\mathrm{H1}) - P(B|\mathrm{H2}) \\ & + P(Y)P(B|\mathrm{H2})]. \end{aligned}$$

The rest is simple algebra

$$0 = -P(B|\mathrm{H1})[1 - P(Y)] - P(X)P(B|\mathrm{H2}) + P(B|\mathrm{H2})[1 - P(Y)] + P(X)P(B|\mathrm{H1}),$$

$$0 = [1 - P(Y)][P(B|\mathrm{H2}) - P(B|\mathrm{H1})] + P(X)[P(B|\mathrm{H1}) - P(B|\mathrm{H2})],$$

$$0 = [1 - P(Y)]/P(X) + \{[P(B|\mathrm{H1}) - P(B|\mathrm{H2})/[P(B|\mathrm{H2}) - P(B|\mathrm{H1})]\},$$

$$1 = [1 - P(Y)]/P(X),$$

$$P(X) = 1 - P(Y).$$

Thus, we have shown that a Cell A observation and a Cell D observation have the same |LLR| if $\mathrm{P(X)} = 1 - P(Y)$. Furthermore, it follows that $|\mathrm{LLR_A}| > |\mathrm{LLR_D}|$ if $P(X) < 1 - P(Y)$ and that $|\mathrm{LLR_A}| < |\mathrm{LLR_D}|$ if $P(X) > 1 - P(Y)$. We can show this by beginning each proof with the desired inequality rather than the equality as above.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15*, 147–149.

Allan, L. G. (1993). Human contingency judgments: rule based or associative? *Psychological Bulletin, 114*, 435–448.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: the joint influence of prior expectations and current situational information. *Psychological Review, 91*, 112–149.

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

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

Anderson, J. R., & Sheu, C-. F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition, 23*, 10–524.

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General, 112*, 117–135.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271–280.

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

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological Review, 104*, 367–405.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*, 545–567.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99*, 365–382.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4*, 317–370.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*, 272–292.

Crocker, J. (1982). Biased questions in judgment of covariation studies. *Personality and Social Psychology Bulletin, 8*, 214–220.

Dagum, P., & Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence, 60*, 141–153.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

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.

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: epistemic utility versus uncertainty reduction. *Psychological Review, 103*, 356–363.

Fales, E., & Wasserman, E. A. (1992). Causal knowledge: what can psychology teach philosophers? *Journal of Mind and Behavior, 13*, 1–28.

Feeney, A., Evans, J. St. B. T., & Clibbens, J. (2000). Background beliefs and evidence interpretation. *Thinking and Reasoning, 6*, 97–124.

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: 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.

Good, I. J. (1983). *Good thinking*. Minneapolis: University of Minnesota Press.

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

Green, D. W., Over, D. E., & Pyne, R. A. (1997). Probability and choice in the selection task. *Thinking and Reasoning, 3*, 209–235.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334–384.

Hilgard, E. R., & Bower, G. H. (1975). *Theories of learning* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*, 65–81.

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

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

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

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: From childhood to adolescence*. New York: Basic Books.

Jennings, D. L., Amabile, T. M., & Ross, L. R. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582–591.

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

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

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition, 51*, 1–28.

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

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

Koehler, J. J. (1992). Probabilities in the courtroom: an evaluation of the objections and policies. In D. K. Kagehiro & W. S. Laufer (Eds.), *Handbook of psychology and law* (pp. 167–184). New York: Springer-Verlag.

Koehler, J. J., & Shaviro, D. N. (1990). Veridical verdicts: Increasing verdict accuracy through the use of overtly probabilistic evidence and methods. *Cornell Law Review, 75*, 247–279.

Levin, I. P., Wasserman, E. A., & Kao, S.-F. (1993). Multiple methods for examining biased information use in contingency judgments. *Organizational Behavior and Human Decision Processes, 55*, 228–250.

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

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

Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General, 127*, 269–285.

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

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

McKenzie, C. R. M. (1999). (Non)Complementary updating of belief in two hypotheses. *Memory and Cognition, 27*, 152–165.

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 and 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. (2005). Judgment and decision making. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of cognition* (pp. 321–338). London: Sage Publications.

McKenzie, C. R. M. (in press). Increased sensitivity to differentially diagnostic answers using familiar materials: implications for confirmation bias. *Memory and Cognition*.

McKenzie, C. R. M., & Amin, M. B. (2002). When wrong predictions provide more support than right ones. *Psychonomic Bulletin and 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 and 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 and Review, 7*, 360–366.

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 and Review, 10*, 596–602.

McKenzie, C. R. M., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes-no and forced-choice tasks. *Journal of Experimental Psychology: General, 130*, 14–155.

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

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

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

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 re-evaluation. *Psychonomic Bulletin and Review, 10*, 289–318.

Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5*, 193–243.

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.

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling: probabilities do matter. *Memory and Cognition, 31*, 143–154.

Oberauer, K., Wilhelm, O., & Diaz, R. R. (1999). Bayesian rationality for the Wason selection task? A test of optimal data selection theory. *Thinking and Reasoning, 5*, 115–144.

Over, D. E., & Green, D. W. (2001). Contingency, causation, and adaptive inference. *Psychological Review, 108*, 682–684.

Over, D., & 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.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufman Publishers.

Peterson, C. (1980). Recognition of noncontingency. *Journal of Personality and Social Psychology, 38*, 727–734.

Poletiek, F. (2001). *Hypothesis-testing behavior*. East Sussex: Psychology Press.

Rescorla, R. A., & Wagner, A. R. . (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Block & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110*, 101–120.

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

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 204–224.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory and Cognition, 8*, 459–467.

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

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 165–173.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645–726.

Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review, 84*, 1329–1393.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19*, 231–241.

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.

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