Consider the alternative: The effects of causal knowledge on representing and using alternative hypotheses in judgments under uncertainty

Brett K. Hayes [1]

Guy E. Hawkins [1,2]

Ben R. Newell [1]

1 School of Psychology, University of New South Wales, Australia

2 Amsterdam Brain and Cognition Center, University of Amsterdam, The Netherlands

Author Note

This work was supported by Australian Research Council Discovery Grant DP120100266 to the first and third authors and a Future Fellowship FT110100151 to the third author. We thank Kate Blundell, Jeremy Ngo and Kelly Jones for help with programming and running the experiments. We thank Klaus Fiedler and two anonymous reviewers for their comments on an earlier draft of this paper.

Correspondence concerning this article should be addressed to Brett K. Hayes, School of Psychology, The University of New South Wales, Sydney, NSW, 2052, Australia. E-mail: B.Hayes@unsw.edu.au

Abstract

Four experiments examined the locus of impact of causal knowledge on consideration of alternative hypotheses in judgments under uncertainty. Two possible loci were examined; overcoming neglect of the alternative when developing a representation of a judgment problem and improving utilization of statistics associated with the alternative hypothesis. In Experiment 1 participants could search for information about the various components of Bayes' rule in a diagnostic problem. A majority failed to spontaneously search for information about an alternative hypothesis, but this bias was reduced when a specific alternative hypothesis was mentioned before search. No change in search patterns was found when a generic alternative cause was mentioned. Experiments 2a and 2b broadly replicated these patterns when participants rated or made binary judgments about the relevance of each of the Bayesian components. In contrast, Experiment 3 showed that when participants were given the likelihoods of the data given a focal hypothesis $p(D|H)$ and an alternative hypothesis $p(D|\neg H)$, they gave estimates of $p(H|D)$ that were consistent with Bayesian principles. Additional causal knowledge had relatively little impact on such judgments. These results show that causal knowledge primarily impacts neglect of the alternative hypothesis at the initial stage of problem representation.

A key part of judgment under uncertainty involves evaluating the probability of a hypothesis given data. For example, when a person notices a persistent red blemish on their forearm they may want to judge how likely it is to be a form of skin cancer. Normatively, such judgments should involve consideration of the likelihood that the observed data could have arisen from one or more different causes. This assumption is enshrined in Bayes' theorem. In its most elementary form given below, the theorem assumes that after observing a datum D (e.g., the red blemish), the probability of a focal hypothesis H (e.g., the person has skin cancer), $p(H|D)$, is assessed relative to the likelihood of the datum in the presence of the focal hypothesis $p(D|H)$ *and* the likelihood of the data given alternative hypotheses $p(D|\neg H)$ (e.g., the blemish is an allergic reaction).

$$p(H \mid D) = \frac{p(D \mid H)p(H)}{p(D \mid H)p(H) + p(D \mid \neg H)p(\neg H)} \quad (1)$$

A large body of evidence however suggests that people often underweight or ignore alternative causes of the data in judgments under uncertainty (Beyth-Marom & Fischhoff, 1983; Dougherty, Thomas & Lange, 2009; Feeney, Evans, & Clibbens, 2000; McKenzie, 1998; Mynatt, Doherty & Dragan, 1993). Such departures from normative reasoning can occur for a number of reasons. First, the reasoner may see the alternative sources of the data as irrelevant to the task of estimating $p(H|D)$ and hence give them little consideration (cf. Beyth-Marom & Fischhoff, 1983; Doherty, Chadwick, Garavan, Barr & Mynatt, 1996; Dougherty, et al., 2009). According to this view people generate sub-optimal estimates because they fail to incorporate alternative sources of the data in their initial representation of the problem. Some research attributes this neglect to a meta-cognitive failure to consider the sampling or causal mechanisms that generate observed data (e.g., Fiedler, 2012; Fiedler, Freytag, & Unkelbach, 2011; Juslin, Winman, & Hansson, 2007). Although these accounts differ in many details they share the view that neglect of alternatives occurs at the initial stages of problem representation.

People may also fail to give adequate consideration of alternative hypotheses at a later stage in the judgment process. According to this view even if people grasp the relevance of alternative hypotheses they may fail to understand how statistical information associated with these hypotheses should be combined to provide an estimate of $p(H|D)$. In other words, people may fail to *utilize* statistical information relating to alternative hypotheses such as the relative likelihoods of $p(D|H)$ and $p(D|\neg H)$. Consistent with this account people often confuse the likelihood of the focal hypothesis $p(D|H)$ with the posterior probability $p(H|D)$ (e.g., Gigerenzer & Hoffrage, 1995; Villejoubert & Mandel, 2002), and often make errors in additively combining $p(D|H)$ and $p(D|\neg H)$ (Bearden & Wallsten, 2004; Riege & Teigen, 2013; Tversky & Koehler, 1994). We refer to these as "utilization deficiency" accounts.

These accounts are in no sense mutually exclusive. It is possible that one may give insufficient attention to the alternative hypothesis when generating an initial representation problem *and* misuse the associated statistical information. Moreover, these are not the only reasons for deviations from Bayesian norms in intuitive judgments of probability. It is well known for example, that people also frequently neglect the base-rate of the focal hypothesis when estimating $p(H|D)$ (Bar-Hillel & Fischhoff, 1981; Barbey & Sloman, 2007; Gigerenzer & Hoffrage, 1995; Hawkins, Hayes, Donkin, Pasqualino & Newell, in press). The current work however focuses primarily on why people fail to appropriately use information about the alternative hypothesis.

Is the neglect of alternatives due to an incomplete causal model?

Krynski and Tenenbaum (2007) outlined a novel approach to explaining failures of statistical reasoning in general, and failures to consider alternative hypotheses in particular. They argue that people typically interpret judgment problems as involving networks of probabilistic causes and effects. According to this view the first step in solving a judgment problem is to construct an intuitive causal model linking the problem elements. Problems

arise when the cause of a key statistic is not clearly specified, leading to its omission from the intuitive model. It follows that judgment accuracy should be improved when the relevant causal relations are more transparent (cf. Ajzen, 1977; Tversky & Kahneman, 1980).

As a test of this approach Krynski and Tenenbaum (2007) presented a version of the classic mammogram problem (cf. Eddy, 1992; Gigerenzer & Hoffrage, 1995), where participants are told that a woman has received a positive mammogram (D) and are asked to estimate the posterior probability that she has cancer $p(H|D)$. Participants are given the base rate or prior probability of the focal hypothesis of having cancer $p(H)$, the likelihood of receiving a positive mammogram given cancer $p(D|H)$, and the false positive rate or likelihood of a positive mammogram in the absence of cancer, $p(D|\neg H)$. When no information about the source of false positives was supplied, less than 15% of participants generated a normative estimate. The accuracy of estimates increased markedly when a specific alternative cause of false positives (a benign cyst) was mentioned. Subsequent work has found a less substantial effect of causal explanation of false positives on judgment accuracy (e.g., Hayes, Newell, & Hawkins, 2013; Hayes, Hawkins, Pasqualino, Newell & Rehder, 2014; McNair & Feeney, 2013). Nevertheless this work found some evidence of causal facilitation such that those given the causal explanation generated estimates that were closer to the normatively correct response, although they were often not precisely correct.

A serious limitation in the interpretation of these results however, is that it is difficult to disentangle whether the additional causal information impacted on *early stage neglect* of the alternative hypothesis or helped participants to better *utilize* the statistics supplied about the base rate and likelihood of the alternative. It is also possible that the causal information had a positive effect on *both* aspects of the judgment process. Krynski and Tenenbaum (2007) appear to favor the latter view. They suggest that causal knowledge can both highlight the need to consider the alternative hypothesis and facilitate more normative

utilization of likelihood statistics. In the absence of such knowledge they argue that "Many participants may realize that the false-positive statistic is somehow relevant, and if they cannot fit it into their model, they may look for some simple way to use it to adjust their judgment" (p. 436).

The chief aim of the current studies therefore was to advance our understanding of how causal knowledge about alternative hypotheses impacts on judgments under uncertainty. In particular we examined whether such knowledge primarily impacts early stage neglect, utilization of likelihood statistics or both. To achieve this we devised different versions of a common diagnostic problem that focused either on neglect (Experiments 1, 2a, 2b) or utilization of the alternative hypothesis (Experiment 3). In the first set of studies participants were asked to search for information that was relevant to a diagnostic judgment under conditions where, a) search for information about the alternative hypothesis required minimal effort, and b) no utilization of numerical statistics was required. By contrast in Experiment 3, the alternative hypothesis was more clearly specified and the goal was to utilize information about the relative likelihoods of the focal and alternative hypotheses to estimate $p(H|D)$.

A second major aim of the current studies was to clarify how the framing of alternative causes impacts their consideration in judgments under uncertainty. Krynski and Tenenbaum (2007) found that specification of an alternative cause of the observed data (e.g., a benign cyst) was sufficient to improve the accuracy of a judgment about the probability of the focal hypothesis (i.e. the likelihood that a woman has cancer given that she has received a positive mammogram). This finding however is at odds with other results that show that mere mention of an alternative cause is insufficient to improve statistical reasoning. Beyth-Marom and Fischhoff (1983) for example, found that when an alternative occupational category was mentioned (e.g., university professor) the perceived relevance of that information for evaluating the likelihood that an individual belonged to a focal category

(business executives) did not differ from a baseline in which no alternative was specified. Increases in the perceived relevance of the alternative were only found when the judgment was re-framed as evaluating the relative evidence for the focal *and* the alternative (i.e., whether it was more probable that an individual was an executive or a professor). Likewise Evans, Venn and Feeney (2002) found an increase in normative search patterns when the final judgment involved deciding between the focal and alternative hypotheses, but not when the alternative was simply mentioned.

This inconsistency in previous findings led us to re-examine the circumstances under which merely mentioning the alternative hypothesis can affect judgments involving $p(H|D)$. Experiments 1 and 2a compared the impact of different levels of specificity of causal information on consideration of an alternative hypothesis. Experiment 2b examined the related issue of the impact of the number of alternative hypotheses suggested. Unlike previous work (e.g., Beyth-Marom & Fischhoff, 1983; Evans, et al., 2002; Krynski & Tenenbaum, 2007) in all experiments the focal and alternative hypotheses were unfamiliar to participants (fictional disease labels). This permitted a more direct assessment of the impact of specificity and number of alternatives, independent of any effects of participants' existing causal knowledge.

Experiment 1

This study aimed to examine whether providing additional information about the cause of an alternative hypothesis reduces early stage neglect of this alternative. There is ample evidence that people fail to spontaneously look for information about alternative hypotheses when the task goal is to evaluate $p(H|D)$. In classic work Wason (1968) showed that when testing a hypothetical rule people tend to focus on searching for information that reconfirms that rule rather than information that would favor an alternative hypothesis (see Evans, 1982, for a review). In judgment tasks people also often fail to give adequate weight

to the alternatives. One form of evidence for this conclusion comes from studies of the phenomenon of pseudodiagnosticity (e.g., Dougherty, et al., 2009; Evans, et al., 2002; McKenzie, 1998; Mynatt, et al., 1993). In this task participants are told that an object has multiple features (D1, D2, etc) and asked to decide whether the object belongs to one of two categories (H1 or H2). To assist in the decision participants can search various likelihood values. The normative strategy from a Bayesian perspective is to search for the likelihood of each feature given each of the competing hypotheses, for example, p(D1|H1), p(D1|H2). Most participants however, prefer the "pseudodiagnostic" strategy of searching for information about the likelihood of additional features associated with a single hypothetical alternative, for example, p(D1|H1), p(D2|H1), and so on. [1]

In an approach closer to the one taken in the current studies, Beyth-Marom and Fischhoff (1983) asked people to search for information about whether a person at a party was more likely to be a university professor (H) or a business executive (¬H) given that the person was known to be a member of a particular club (D). Information about the base rate of professors at the party p(H), and the percentage of professors who were club members p(D|H) were rated as relevant to the judgment by a majority, whereas the likelihood of the alternative hypothesis p(D|¬H) was only rated as relevant by a minority (around 35% of participants).

Such results suggest that people are likely to fail to *spontaneously* look for information concerning alternative causes of the data when assessing the evidence for p(H|D). To examine this prediction we adapted a task used by Doherty and Mynatt (1990), in which searching successive Bayesian components involved minimal effort. Participants took the role of a doctor seeking to diagnose whether a patient with a particular symptom (the datum, D) had a particular disease (the focal hypothesis, H). They were then presented with text descriptions of the four components of Bayes' rule, p(H), p(¬H), p(D|H), p(D|¬H), and asked

to choose the components that would be most helpful in arriving at a diagnosis. Unlike Krynski and Tenenbaum (2007), no numerical values were presented and participants were not required to produce a numerical estimate of probability. Instead, the key outcome measure was the frequency with which information concerning the alternative hypothesis, $p(D|\neg H)$ and/or $p(\neg H)$, was consulted during search. Following previous work on early stage neglect (e.g., Beyth-Marom & Fischhoff, 1993; Dougherty, et al., 2009), it was expected that participants would neglect information about the alternative in favor of information about the focal hypothesis, even in this "minimalist" diagnostic reasoning task. The two more novel questions addressed in this study were 1) can such early stage neglect be reduced by providing additional information about the causes of the observed data?, and 2) do such effects depend on the specificity of this causal information?

Krynski and Tenenbaum (2007) provided participants with a specific causal alternative to the focal hypothesis. However causal knowledge is not always so specific. Keil (2006) for example, suggests that people often possess only very general and often incomplete knowledge about the causes of mechanical and natural phenomena. Nevertheless in most circumstances this level of knowledge is adequate for everyday inference and adaptive functioning. The question therefore arises as to what level of specificity of causal knowledge is required to generate an additional "causal node" in an intuitive causal model.

To answer this we compared the impact on early stage neglect of providing a specific causal alternative with a more generic form of causal knowledge. The way that the alternative hypothesis was presented in the search task was manipulated between groups. The unspecified alternative was a baseline condition where no mention was made of an alternative source of the observed symptoms. In the generic alternative condition, it was suggested that the observed symptom could arise due to other diseases, but no specific disease labels were given. In the specified alternative, a specific disease was identified as an

alternative cause of the observed symptoms. Finally we added a specified alternative – label group to control for labelling and response option differences between the specified alternative and other conditions (see Design for details).

If additional causal information impacts on early stage neglect then we should see more inspection of $p(D|\neg H)$ and/or $p(\neg H)$ in the specified alternative condition than in the unspecified baseline. If generic causal knowledge is sufficient to prime attention to the alternative then we should see a parallel effect in the generic alternative condition.

Method

Participants. A total of 1207 participants (39% female; M_{AGE} = 30.39 years) were recruited via Amazon Mechanical Turk and were paid 0.50c US. In this and all subsequent experiments participation was restricted to respondents from the United States who had at least a 90% approval rating for their previous MTurk work. Three were excluded because they failed to respond correctly to the attention screening item (described below). Participants were randomly allocated to one of four conditions. Sample sizes after exclusion were unspecified alternative, n = 299; generic alternative, n = 303; specified alternative, n = 298 and specified alternative – label control, n = 304. Note that a large sample was required in order to examine the distribution of search patterns, separately in each condition, over 15 possible search combinations, ignoring search order.

Design and Procedure. Participants in all conditions were told to imagine that they were a doctor and that that their aim was to identify the types of information that would assist in determining whether a patient who showed a particular symptom ("red rash on their fingers") had a focal disease ("Buragamo"). Participants were told that that they could search as few or as many options as they wanted. They were also reminded that in real-life diagnosis additional searches are likely to cost time and money, and so only options that were "necessary for the diagnosis" should be chosen (see Figure 1, panel A for full instructions).

Ten seconds later four search options appeared on the screen. These included the base rate of the focal hypothesis $p(H)$, the base rate of the alternative hypothesis $p(\neg H)$, the likelihood of the data (i.e., red rash) given the focal hypothesis $p(D|H)$, and the likelihood of the data given the alternative hypothesis $p(D|\neg H)$. These were presented in four rectangles of different colors, arranged in two rows of two (e.g., Figure 1, panels B and C). The location of each search option was randomized for each participant. Participants clicked on a rectangle to indicate their choice of that option. Once selected the rectangle changed color. A previously selected search option could be de-selected by clicking it a second time. Participants were required to select at least one option. Once selections were finalized participants clicked an on-screen button to advance to the attention screening question. This involved the presentation of four unlabeled rectangles on screen; two were the same color and the other two were different colors. Participants were asked to click on the two identical rectangles.

Four experimental conditions differed in their specification of an alternative hypothesis (i.e., a different diagnosis) and labeling of the search alternatives. In the unspecified alternative condition only the focal disease was mentioned in the instructions and search options. The alternative hypothesis was "without Buragamo" (Figure 1, panel B). The generic alternative condition was similar except that it was noted in instructions that the symptom of red rash "could be caused by a number of diseases". In the specified alternative condition participants were told that red rash could also be caused by the disease "Terragaxis" and specified "with Terragaxis" as the alternative hypothesis in the search options (see Figure 1, panel C). The specified alternative – label control had identical instructions to the specified alternative condition but used the same search options as the unspecified and generic alternative conditions. Including this condition allowed us to examine whether simply providing a specific alternative was sufficient to increase search of information relevant to the alternative hypothesis or whether this alternative also had to be

identified on one of the search options. The task took around 5 minutes to complete on average.

Results

An initial analysis found that the total number of options searched differed between experimental conditions, $F(3, 1200) = 14.05$, $\eta^2_P = 0.34$, $p < .001$. Post-hoc comparisons (Tukey's HSD) revealed that those in the specified alternative condition searched more options on the average ($M = 1.89$, $SD = 0.70$) than any of the other three conditions (unspecified alternative, $M = 1.67$, $SD = 0.64$; generic alternative, $M = 1.59$, $SD = 0.57$; specified alternative – label control, $M = 1.60$, $SD = 0.60$; all q's > 5.8, p's < .001), which did not differ from one another (q's < 2.3, p's > .05).

The proportion of participants in each condition using each of the 15 possible search patterns is given in Table 1. From a Bayesian perspective the most informative search patterns were the combination of $p(H)$, $p(D|H)$ and $p(D|\neg H)$, and the selection of all four options. Relatively few participants selected either of these combinations but rates of choice were higher in the unspecified (8%) and specified alternative (7%) groups than in the generic alternative (3%) or label control (1%) groups, χ^2 (3, N = 1204) = 9.55, $p = .02$.

Table 1 shows that all groups saw $p(D|H)$ as important and were likely to search for this information either as their only choice or in combination with other options. In the unspecified alternative, generic alternative and specified alternative – label control groups the three most common strategies in descending order of frequency were $p(D|H)$, the combination of $p(D|H)$ and $p(D|\neg H)$, and $p(H)$ combined with $p(D|H)$. Jointly these accounted for 69-83% of all searches. Those in the specific alternative condition showed a different pattern. In this case the modal choice was $p(D|H)$ combined with $p(D|\neg H)$, selected by more than half the participants, with $p(D|H)$ the next most common choice. The percentage of participants selecting $p(D|H)$ combined with $p(D|\neg H)$ in the specified

alternative condition (50.7%) was reliably higher than in any of the other three conditions; unspecified alternative (22.7%, $z = 7.08$, $p < .001$), generic alternative (25.7%, $z = 6.29$, $p < .001$), specified alternative – label control (22.0%, $z = 7.77$, $p < .001$).

A more inclusive test of differences in consideration of the alternative hypothesis was carried out by aggregating search patterns according to whether or not they included $p(D|\neg H)$ and/or $p(\neg H)$. The proportion of searches that included the alternative hypothesis was higher in the specified alternative group (70.5%) than in the other three conditions (unspecified alternative: 41.1%; generic alternative: 41.9%; specified alternative – label control: 44.7%), $\chi^2(3, N = 1204) = 70.52$, $p < .001$. This result remained robust when four-option searches were excluded, $\chi^2(3, N = 1180) = 60.98$, $p < .001$.

Although our main focus was on search for information about the alternative hypothesis, we also examined group differences in consideration of the base rate of the focal hypothesis. All search patterns were aggregated according to whether they included $p(H)$ or not. The proportion of searches that included $p(H)$ was found to be lower in the specified alternative condition (18.5%) than in the other three conditions (unspecified alternative: 35.8%, generic alternative: 28.4%; specified alternative – label control: 39.5%), $\chi^2(3, N = 1204) = 36.48$, $p < .001$. This result remained robust when four-card searches were excluded, $\chi^2(3, N = 1180) = 32.8$, $p < .001$.

Discussion

In this study participants searched for information that they believed to be relevant for diagnosing a particular disease (the focal hypothesis) given observed data. The two key questions were whether people would spontaneously search for information about an alternative cause of the data (the alternative hypothesis) and whether this search would increase when information about an alternative cause of the data was provided. The first finding was that despite the minimal effort required for information search, most participants

showed profound neglect of information about the alternative hypothesis in their information search. Fewer than half the participants in the unspecified alternative, generic alternative, and specified alternative – label control conditions searched for any information about the alternative hypothesis.

This neglect of the alternative hypothesis in searching diagnostic information is consistent with a range of previous findings showing a preference for examination of statistical information about the focal hypothesis over the alternative (e.g., Beyth-Marom & Fischhoff, 1983; Dougherty, et al., 2009; Mynatt, et al., 1993). The current results re-confirm this finding using a task that made minimal demands on attention and memory and required no numerical calculation. Consistent with early-stage neglect accounts these data show that people rarely spontaneously search for information about alternative causes when evaluating the likelihood of a focal hypothesis.

The more novel finding was that this bias was reduced by mentioning a specific alternative *cause* of the observed data. These results are consistent with previous work demonstrating that the bias to search for information about the focal hypothesis may be reduced when a clear alternative hypothesis is specified (e.g., Doherty, et al., 1996; Trope & Mackie, 1987). This represents an important first step toward clarifying the effects of providing causal information on alternatives reported by Krynski and Tenenbaum (2007). .

This experiment also clarified the level of specificity of causal information necessary to produce increased attention to the alternative. This increase was only found when, a) the alternative cause was given an explicit label, and b) the alternative hypothesis was clearly labeled in the search options. Just mentioning the possibility of alternative causes of the data was insufficient to increase search for information related to the alternative. It appears therefore that specific but not general causal knowledge is likely to increase the likelihood

that people will incorporate alternative explanations of the data into their intuitive causal models.

As in many previous studies (e.g., Bar-Hillel & Fischhoff, 1981; Barbey & Sloman, 2007; Gigerenzer & Hoffrage, 1995) only a minority of participants saw base rate information as relevant to evaluating $p(H|D)$. A somewhat unexpected finding was that increased search for information about the alternative in the specified alternative condition appeared to come at the expense of information about the base rate of the focal hypothesis. In the specified alternative condition, participants were *less* likely to search options that included $p(H)$ than in other conditions.

It may be that priming attention to the alternative hypothesis directed attention away from normatively relevant information about the focal hypothesis. There are also two artifactual explanations of this result that need to be considered. Recall that in the specified alternative condition it was made clear that the two diseases were the only causes of the observed symptom, hence $p(\neg H) = 1 - p(H)$. Those in this group who searched options including $p(\neg H)$ may have seen search for the focal base rate as redundant. However this seems unlikely since searches for the complementary base rates $p(H)$ and $p(\neg H)$ were actually more common in the specified alternative condition (5% of searches) than in the other three conditions (<1% of searches), $z = 6.43$, $p < .001$.

Another possibility is that the tradeoff is an artifact of our search procedure. Table 1 shows that two options was the upper limit of information search for most participants and that one of these was very likely to be $p(D|H)$. This conservative search was likely due, at least in part, to the explicit instruction to only search for the most relevant types of information. The tradeoff may therefore reflect a change in the relative priority of the second option chosen for search. It may be that those in the specified alternative condition still believed that $p(H)$ was relevant to diagnosis but that choice of this option lay beyond the two

choice "stopping rule" adopted by many participants. In Experiment 2a we therefore used a procedure with no perceived cost to examining multiple options.

Experiment 2a

Experiment 2 aimed to further clarify the effects of causal information on early stage neglect of information about the alternative hypothesis. In particular we aimed to make a more detailed assessment of people's beliefs about the neglected options. When these options were not searched it could be that they saw no value in that type of information for estimating $p(H|D)$. Alternately they may have seen the option as relevant but having insufficient weight relative to the options that were searched. In this study therefore participants were asked to i) rate the relevance of each of the four information options, and ii) to make a binary choice about which option was "relevant" to solving the diagnosis problem. Unlike the search task in Experiment 1 which prompted participants to prioritize the four response options, the current ratings and choices allowed participants to give an independent assessment of the relevance of each Bayesian component to assessing $p(H|D)$. If causal knowledge has a robust impact on early stage neglect then we should still see higher relevance ratings and choices in the specified alternative as compared to the unspecified and generic conditions.

A second modification to the Experiment 1 procedure was the provision of more detailed descriptions of each of the response options. This addressed the possibility that the text descriptions in Experiment 1 may not have clearly differentiated between the information conveyed by the base rate options, $p(H)$ and $p(\neg H)$, as opposed to the likelihoods, $p(D|H)$ and $p(D|\neg H)$. The experimental design was also streamlined by dropping the specified alternative – label control condition.

Method

Participants. A total of 375 participants (45% female; M_{AGE} = 34. 9 years) were recruited via Amazon Mechanical Turk and were paid 0.50c US. Twenty eight were excluded because they failed to respond correctly to the attention screening items, reported having participated in the previous experiment, or rejected all response options as irrelevant in both ratings and choices. Participants were randomly allocated to one of three conditions. Sample sizes after exclusion were, unspecified alternative, n = 116; generic alternative, n = 115; specified alternative, n = 116.

Procedure. The general instructions in each of the three experimental conditions were similar to the corresponding conditions in Experiment 1 except there was no mention of a cost associated with rating or choosing a response option. The major change from Experiment 1 was in the way that participants indicated the response options which they saw as relevant for diagnosis. This proceeded in two stages. In the first stage participants were told to "evaluate each of the options separately and rate its relevance to the diagnosis" using a 7-point Likert scale (1 = "not at all relevant", 7 = "highly relevant"). After completing the ratings, the task instructions and response options were repeated but participants were now asked to choose the response options "you think are relevant for the diagnosis and which are irrelevant". As in the previous study the four response options were presented as labeled rectangles. Participants clicked radio buttons to indicate whether or not they believed each option to be relevant or irrelevant to the diagnostic judgment.

In each stage the descriptions of the response options were more detailed than those used in Experiment 1. The following was added to the response options in panel B of Figure 1 (unspecified and generic alternative conditions): for the $p(H)$ response option – "These are the people in the population who HAVE Buragamo, but may or may not have a red rash"; for the $p(\neg H)$ response option – "These are the people in the population who DO NOT HAVE Buragamo, but may or may not have a red rash"; for $p(D|H)$ – "This is the subset of people

who HAVE Buragamo and have a red rash"; for $p(D|\neg H)$ – "This is the subset of people who DO NOT HAVE Buragamo but have a red rash". The following was added to the corresponding response options in panel C of Figure 1 (specified alternative condition); $p(H)$ – "These are the people in the population who HAVE Buragamo, but may or may not have a red rash"; $p(\neg H)$ – "These are the people in the population who HAVE Terragaxis, but may or may not have a red rash"; $p(D|H)$ – "This is the subset of people who HAVE Buragamo and have a red rash"; $p(D|\neg H)$ – "This is the subset of people who HAVE Terragaxis and have a red rash".

Results and Discussion

Relevance ratings. The mean ratings given to the response options are given in Figure 2. To examine group differences the data were entered into a series of one-way analyses of variance with two planned orthogonal contrasts; the first compared ratings in the specified alternative condition with the other two conditions; the second compared ratings in the unspecified and generic alternative conditions.

Ratings of the relevance of the focal base rate were generally close to the mid-range of the scale and did not differ between groups (F's < 1.3). Ratings of the base rate of the alternative hypothesis were generally below the mid-range of the scale. As shown in Figure 2 however, this option was rated as more relevant in the specified alternative than in the other two conditions, $F(1, 344) = 7.77$, $\eta^2_P = 0.02$, $p = .006$. No differences in ratings were found between the unspecified and the generic alternative, ($F < 1.5$). The likelihood of the focal hypothesis given the data was rated as highly relevant in all conditions, where ratings did not differ across the three conditions (F's < 0.5). Notably ratings of the relevance of the likelihood of the alternative hypothesis were close to the middle of the scale in the unspecified and generic alternative conditions but were significantly higher in the specified

alternative condition, $F(1, 344) = 6.5$, $\eta^2_P = 0.02$, $p = .01$. No differences in these ratings were found between the unspecified and generic alternative conditions ($F < 2.7$).

Relevance choices. The proportion of occasions that a response option was chosen as relevant is given in Figure 3. In many respects, the pattern of choices in each group was similar to the rating data. The likelihood of the focal hypothesis $p(D|H)$ was chosen by most participants in all conditions and there were no group differences ($F < 2.1$). The base rate of the alternative hypothesis and the likelihood of the alternative hypothesis were both more likely to be chosen as relevant by those in the specified alternative condition than the other two conditions ($F(1, 344) = 5.51$, $\eta^2_P = 0.02$, $p = .02$; $F(1, 344) = 5.02$, $\eta^2_P = 0.01$, $p = .026$, respectively), whose rates of choice did not differ (F's <1.0). [2]

In this case the focal base rate was less likely to be chosen by those in specified alternative condition than the other conditions, $F(1, 344) = 4.96$, $\eta^2_P = 0.01$, $p = .027$. The combination of increased choice of $p(D|\neg H)$ and $p(\neg H)$ with reduced choice of $p(H)$ in the specified alternative resembles the tradeoff in attention to components found in Experiment 1.

These data extend the findings of Experiment 1 in a number of ways. The rates of endorsement of $p(D|\neg H)$ as relevant to evaluating $p(H|D)$ were considerably higher than the corresponding search rates in the earlier study. This presumably reflects the removal of the implied costs associated with searching additional Bayesian components. Moreover the rating data show that many participants saw $p(D|\neg H)$ as having some value in assessing $p(H|D)$.

Nevertheless as in Experiment 1, people generally saw information related to the alternative hypothesis as less relevant to diagnosis than information about the focal hypothesis. Crucially, as in Experiment 1 the perceived relevance of both $p(\neg H)$ and $p(D|\neg H)$ increased significantly when a concrete alternative source of the observed symptoms was

specified. By comparison, generic mention of other possible causes had little impact on perceived relevance. In the choice measure we again found evidence of a tradeoff in attention to the alternative hypothesis and the base rate of the focal hypothesis.

These results again indicate that causal knowledge about the alternative hypothesis impacts early stage neglect. Before drawing a decisive conclusion about this issue however, it was important to show that these effects persist under conditions where information irrelevant to $p(H|D)$ was included among the rated options.

Experiment 2b

Experiments 1 and 2a found clear evidence of neglect of alternatives at the early stage of problem representation and that causal knowledge about the alternative could reduce this neglect. A possible concern about these studies however, is that participants were only asked to search or make relevance judgments about the four components of Bayes' rule that are normatively relevant to deriving $p(H|D)$. The absence of normatively irrelevant options could have biased the results in favor of neglecting information about the alternative. Some participants may have believed that they were required to discriminate between the four presented alternatives, leading to an artifactual reduction in search for and ratings of $p(D|\neg H)$ and $p(\neg H)$. To address this issue we replicated the Experiment 2a rating and choice procedure but added two normatively irrelevant components.

This experiment also aimed to further delineate the conditions under which neglect of alternatives can be reduced. Both previous studies found an increase in attention to information about the alternative hypothesis when this hypothesis was given a specific label but not when it was referred to generically. However the wording of these instruction conditions introduced a potential confound; there was clearly only a single specified alternative, whereas in the generic condition it was suggested that the observed symptom could be due to more than one disease. Hence it is unclear whether the advantage found for

the specified condition was due to the specific content of the label or to the identification of a single rather than multiple alternative hypotheses. The latter view is suggested by research suggesting that people often favor explanations based on a small number of causes over multi-causal explanations (Lombrozo, 2007). To resolve this issue a new condition was added to the design, where reference was made to a single generic alternative source of the observed symptom. If it is the specific content rather than the number of explanations that is crucial to reducing neglect, then we should still see less neglect of the alternative hypothesis in the specified alternative as compared to the single generic condition.

Method

Participants. A total of 498 participants (43% female; M_{AGE} = 31. 8 years) were recruited via Amazon Mechanical Turk and were paid 0.50c US. Thirty six were excluded because they failed to respond correctly to the attention screening items, reported having participated in previous experiments, or accepted or rejected all six response options. Participants were randomly allocated to four conditions. Sample sizes after exclusion were, unspecified alternative, n = 125; generic alternative, n = 110; generic single-alternative, n=115; specified alternative, n = 112.

Procedure. The instructions and procedure were those used in Experiment 2a, with the following exceptions. Two options that were normatively irrelevant to computing $p(H|D)$ were added to each condition (i.e., "Percentage of people without any disease"; "Percentage of people without any disease and no red rash"). Hence participants provided relevance ratings and made a binary relevance judgment for each of six components (four relevant Bayes components, two irrelevant distractors), presented on screen in random order.

The second change was the addition of a generic single-alternative condition. The instructions for this condition were similar to those for the generic condition except that participants were told to "Note that this symptom could be caused by another disease". To

further reduce the procedural differences between the specified alternative and other conditions the statement that the focal and alternative diseases were the only possible causes of the observed symptom was removed from that condition. [3]

Results and Discussion

Relevance ratings. The mean relevance ratings given to the response options are shown in Figure 4. The Figure shows that participants clearly distinguished between the relevance of the four Bayesian components and the two distractors, with much higher ratings given for the former, $F(1, 462) = 2073.88$, $\eta^2_P = 0.82$, $p<.001$.

Mean relevance ratings for the four Bayes rule components were generally higher in this study ($M = 4.91$) than in Experiment 2a ($M = 4.67$); cross-experimental $t(802) = 2.92$, $p = .003$. Nevertheless, many of the trends in Figure 4 replicated those found in the previous study. Ratings of $p(D|H)$ were close to ceiling in all groups. Ratings of the base rate of the focal hypothesis were again close to the mid-rage and did not differ between groups (all F's<1.0).

As in the previous study, ratings of the base rate of the alternative hypothesis were generally below the mid-range of the scale. Again, however ratings for this component were higher in the specified alternative than in the other three conditions, $F(1, 458) = 7.35$, $\eta^2_P = 0.02$, $p = .007$. Ratings of this component in the generic-single condition did not differ from those in the unspecified condition, $F(1, 458) = 0.46$.

As noted, ratings of the relevance of the likelihood of the alternative hypothesis $p(D|\neg H)$ were somewhat higher than in Experiment 2a. In this study however they were still well below the corresponding ratings for $p(D|H)$, $F(1, 458) = 213.34$, $\eta^2_P = 0.32$, $p<.001$. Unlike the previous study there were no group differences between these ratings (F's <2.7).

Relevance choices. The proportion of occasions that a response option was chosen as relevant is given in Figure 5. In most respects, the pattern of choices in each group was

similar to the rating data. Participants again clearly discriminated between the relevance of the four Bayes components and the distractors, $F(1, 462) = 2775.24$, $\eta^2_P = 0.86$, $p<.001$. The four Bayes components were more likely to be chosen as relevant in this study ($M = 0.80$) compared to Experiment 2a ($M = 0.69$); cross-experimental $t(802) = 7.52$, $p <.001$.

The likelihood of the focal hypothesis $p(D|H)$ was chosen as relevant by most participants in all conditions and there were no group differences (F's <2.2). Unlike the earlier study, there were no group differences in choice of the focal base rate, F's <2.1). The base rate of the alternative hypothesis was more likely to be chosen as relevant by those in the specified alternative than the other three conditions, $F(1, 458) = 13.78$, $\eta^2_P = 0.03$, $p <$ .001. The rate of choice of this component did not differ between the unspecified and generic-single conditions ($F <0.5$).[4]

The likelihood of the alternative hypothesis $p(D|\neg H)$ was judged as relevant less often than $p(D|H)$, $F(1, 458) = 46.03$, $\eta^2_P = 0.09$, $p<.001$. There were no group differences between judgments of the relevance of $p(D|\neg H)$, (F's <1.7).

Overall these results show that people could readily discriminate between components of Bayes theorem that are potentially relevant to estimating $p(H|D)$ and irrelevant distractors. The inclusion of the distractors appeared to increase relevance ratings and judgments for the four Bayesian components, relative to the corresponding measures in the previous experiment. This suggests that ratings in the previous study may have been affected by a demand characteristic to reduce ratings for the Bayes components that were perceived as the least relevant to the diagnosis task.

The corresponding increase in relevance ratings and choices for the four Bayesian components did weaken some of the group differences found in Experiment 2a. In this case the provision of the specific causal information did not increase relevance ratings or choices of $p(D|\neg H)$ above the unspecified baseline. Nevertheless specific causal information did

increase the perceived relevance of the base rate of the alternative hypothesis. Notably relevance ratings and judgments for the single generic alternative did not differ from those given when no alternative was specified. Hence it appears that it is the specificity of content rather than the number of alternative causes of the data that is crucial to reducing neglect.

In this case we found no evidence of a tradeoff between the perceived relevance of the focal base rate and Bayes components associated with the alternative hypothesis. This suggests that the tradeoff is restricted to conditions like those in the previous experiments where participants may have felt the need to discriminate between normatively relevant alternatives. Under such conditions it appears that providing causal information about one component of Bayes rule can lead to neglect of other components (see Fischhoff & Bar-Hillel, 1984, for a related finding regarding the mixed outcomes associated with training to attend to the focal base rate).

Taken together, the results of Experiments 1, 2a and 2b re-confirm earlier findings of profound neglect of information about the alternative hypothesis when assessing $p(H|D)$ (e.g., Beyth-Marom & Fischhoff, 1983; Dougherty, et al., 2009; Mynatt, et al., 1993). This myopic view was most evident at the earliest stage of problem representation when participants simply had to search for or evaluate the potential value of the various components of Bayes' rule. The most important novel finding in these studies is that this early stage neglect could be reduced by providing a specific alternative cause of the observed data. By comparison provision of generic causal information (either about a single or multiple causes) had little effect on early stage neglect. These results refine Krynski and Tenenbaum's (2007) claims about the effects of causal knowledge on Bayesian reasoning, suggesting that such knowledge promotes the incorporation of information about the alternative hypothesis into an intuitive representation of the problem.

Experiment 3

The final experiment examined the impact of causal knowledge on the utilization of statistical information associated with the alternative hypothesis. In this experiment participants were provided with the likelihood statistics associated with a focal hypothesis $p(D|H)$ and an alternative $p(D|\neg H)$, and were asked to estimate $p(H|D)$. In this context people are unlikely to neglect the alternative because evaluating the strength of the evidence favoring the two alternatives is an explicit goal of the task (e.g., Beyth-Marom & Fischhoff, 1983; Feeney, et al., 2000; Klayman & Brown, 1993).

The key question was whether utilization of such statistical information is affected by the way the alternative hypothesis is specified. Krynski and Tenenbaum (2007) argue that people generally have difficulty reasoning about data observed in the *absence* of a cause (e.g., the likelihood of red rash in the absence of Buragamo). Hence providing a specific alternative cause (e.g., the likelihood of red rash given Terragaxis) should improve utilization of the likelihood statistics.

In contrast, it may be that once early stage neglect is overcome, the provision of a specific causal alternative has little additional effect on utilization. Such a view is suggested by accounts which see a meta-cognitive failure to consider alternative causes as the primary source of errors in statistical reasoning (e.g., Fiedler, 2012; Juslin, et al., 2007). Such accounts suggest that once neglect of an alternative cause is overcome people are often quite accurate in their intuitions about the implications of relevant statistical information.

To test these contrasting predictions, participants were presented with a diagnostic problem in which they were sequentially presented with relevant statistical information. In the first stage they were given the base rate of a focal hypothesis (the disease Buragamo) and then asked to estimate $p(H|D)$. In the second stage they were presented a new datum (red rash) and with the likelihoods of $p(D|H)$ and $p(D|\neg H)$. Participants were asked to re-estimate

p(H|D) in the light of this information. As in the previous experiments the alternative hypothesis was either unspecified (i.e., was described as the absence of the focal disease) or was presented as a specified alternative (i.e., a different disease).

Normatively, when the focal and alternative hypotheses are equally likely given the data then no revision in initial estimates should take place. If the likelihood of p(D|H) is substantially greater than p(D|¬H) then estimates should be revised upwards. If the likelihood of p(D|H) is substantially lower than p(D|¬H) then estimates should be revised downwards. The details of these Bayesian predictions are summarized in Table 2. Our main interest was in whether participants given information about a specified alternative cause showed more or equally appropriate qualitative changes in probability estimates than those given no specified alternative.

Method

Participants. A total of 959 participants (45% female; M_{AGE} = 32.98 years) were recruited via Amazon Mechanical Turk and were paid $1.00 US. Eleven participants were excluded because they failed the attention check question, gave zero probability estimates at either stage 1 or 2, or reported having previously completed a similar study. Participants were randomly allocated to one of eight conditions of roughly equal size (n's = 117-123).

Design and Procedure. The experiment was carried out in two stages. In stage one, participants were told that they were to take the role of a doctor and to use the information given to assess the probability that a new Patient X has the disease Buragamo. They were then given the disease base rate ("80% of patients that you have seen had Buragamo. The remaining patients did not have Buragamo"), and asked "What do you think is the probability that Patient X will have Buragamo?" Answers were given as a percentage.

In stage two participants in the unspecified alternative condition were told that Patient X has been found to have a red rash and that the following was known from medical records

"Y% of patients that you have seen WITH Buragamo have a red rash and that Z% of patients that you have seen WITHOUT Buragamo have a red rash". Those in the specified condition were told "Y% of patients that you have seen WITH Buragamo have a red rash and that Z% of patients that you have seen WITH another disease Terragaxis have a red rash."

Four different combinations of the likelihoods of $p(D|H)$ and $p(D|\neg H)$ (corresponding to the Y% and Z% terms in the instruction frame) were administered to different groups. The details are given in Table 2. The Table shows that there were two conditions (high/high and low/low) where the likelihoods of $p(D|H)$ and $p(D|\neg H)$ were identical, and two conditions where they differed (high/low, low/high).

The likelihood manipulation was crossed with the framing of the alternative hypothesis. In the unspecified condition, the alternative was described as the absence of the focal disease. In the specified condition the alternative was a different disease (see alternate instructions given above). Hence, there were a total of eight between-subjects conditions.

After being presented with the likelihood information, participants were given their stage 1 probability estimate and asked "What do you now think is the probability that Patient X has Buragamo?" Answers were again given as a percentage. This task took around 15 minutes to complete on average.

Results and Discussion

Stage 1. Seventy eight per cent of participants gave the normatively correct estimate of "80" when provided only with the base rate of the focal hypothesis. However there was a marginal trend for these initial estimates to differ between groups, $F(1, 940) = 4.04$, $\eta^2_P = 0.004$, $p = .045$. Initial estimates in the specified alternative condition tended to be higher ($M = 75.89\%$, $SD = 11.54$) than those in the unspecified condition ($M = 74.25\%$, $SD = 13.48$). Hence in the analyses of Stage 2 estimates, individual estimates given during Stage 1 were used as a covariate.

Stage 2. The key dependent measure was the change in estimates of $p(H|D)$ between stages 1 and 2 (see Figure 6). Because the predicted direction of change differed for high/high vs. high/low likelihood and low/low vs. low/high likelihood conditions, these conditions were analyzed separately. If people considered the likelihoods of both the focal hypothesis $p(D|H)$ and the alternative $p(D|\neg H)$, then there should have been little change in estimates from stage 1 to stage 2 in the high/high condition but an increase in estimates in the high/low condition (see Table 2). A 2 (likelihood condition) x 2 (framing of alternative) analysis of covariance (ANCOVA) revealed a reliable difference in the change in estimates in the high/low as compared with high/high conditions, $F(1,473) = 91.96$, $\eta^2_P = 0.16$, $p < .001$. Figure 6 shows that estimates in the high/high condition generally did not change from stage 1 to 2. However estimates in the high/low condition were revised upwards, consistent with Bayesian predictions. The specification of the alternative hypothesis had no reliable effect on these results (F's < 1.0). These results remained unchanged when the covariate was removed.

If people are considering information about both the focal and the alternative hypothesis then there should be little change in the low/low condition from stage 1 to 2 but a decrease in estimates in the low/high condition. As shown in Figure 6, this prediction was confirmed. There was a reliable difference in the amount of change in probability estimates in the low/high as compared to the low/low condition, $F(1,465) = 139.07$, $\eta^2_P = 0.23$, $p < .001$. The Figure shows some decrease in estimates in the low/low condition from stage 1 to 2. Nevertheless, there was a much larger decrease in estimates in the low/high condition. There was also a small but reliable interaction between the specification of the alternative and the contrast comparing the low/low and low/high conditions, $F(1,465) = 4.88$, $\eta^2_P = 0.01$, $p = .03$. The difference between the revision of estimates in the low/low and low/high conditions was larger in the specified than the unspecified condition. Note however that when the

covariate was removed the difference between revision of estimates in the low/low and low/high likelihood conditions remained robust ($p < .001$) but the interaction between likelihood condition and framing of the alternative became marginal ($p = .06$).[5]

As shown in Table 2, given a stage 1 probability estimate of 80%, Bayes' rule predicts that absolute change in stage 2 estimates in the high/low condition should be smaller (normatively ≈ 17%) than in the low/high condition (normatively ≈ 49%). This prediction was also confirmed, $F(1,473) = 231.51$, $\eta^2_P = 0.33$, $p < .001$. However Figure 6 shows that updating in each case was conservative, with change in estimates generally smaller than the values prescribed by Bayes' theorem.

The differences in belief revision between groups with varying likelihood combinations were maintained when the data was examined at the individual level. Individual stage 2 estimates were classified according to whether they showed no change (defined as giving exactly the same answer at stages 1 and 2), increased, or decreased relative to the estimate given in stage 1. The majority of participants in the high/high condition (57.8%) did not change their estimates whereas the majority of those in the high/low condition (83%) increased their estimates, χ^2 (2, N = 478) = 193.7, $p < .001$. The majority of those in the low/low condition (51%) also showed no change whereas the majority in the low/high condition (81%) decreased their estimates, χ^2 (2, N = 470) = 106.8, $p < .001$. Patterns of change in individual estimates in the high/low and low/high conditions did not differ according to whether the alternative hypothesis was specified, p's > 0.2.

When provided with $p(D|H)$ and $p(D|\neg H)$ people revised their estimates of $p(H|D)$ in a manner that was consistent with qualitative predictions from Bayes' rule. Notably, this was the case even when no specific disease was given as an alternative hypothesis. Unlike the previous studies the specification of an alternative effect only had a small effect on a subset of the required judgments (i.e., only in the low/high condition).

Overall these results show that people generally understood the implications of the relative likelihoods of the focal and alternative hypotheses and used this information in a manner that, at least at a qualitative level, conformed to Bayesian principles. Providing additional specification about the alternative cause had relatively little impact on estimation of $p(H|D)$. This pattern differs from that found in the earlier experiments where providing a specific alternative cause reduced early stage neglect of the alternative hypothesis.

General Discussion

These experiments aimed to identify the locus of impact of causal information on consideration of an alternative hypothesis in judgments under uncertainty. Previous work on causal facilitation of judgments (e.g., Krynski & Tenenbaum, 2007) left open three possibilities; i) that causal knowledge assists by highlighting the relevance of the alternative hypothesis when developing a representation of the problem at hand; ii) that causal knowledge improves utilization of statistical information about the alternative hypothesis, and iii) that causal knowledge has a positive effect on both problem representation and utilization.

The results of four experiments strongly support the first of these accounts. Consistent with the notion of early stage neglect, people frequently failed to consult information about the alternative hypothesis when asked to evaluate $p(H|D)$ (Experiment 1) or rated it as having only modest relevance to this problem (Experiments 2a, b). In all of these studies however, consideration of the alternative hypothesis increased when it was presented as a specific alternative cause of the observed data.

In contrast, when the likelihoods of the data given the focal and alternative hypothesis were stated explicitly, most participants utilized this information in a manner that was broadly consistent with a normative approach (Experiment 3). Providing a specific

alternative cause had relatively little impact on normative utilization of the likelihood statistics.

These results clarify the locus of the effect of causal information on consideration of alternative hypotheses. There was a profound effect on initial problem representation, but only a small effect on understanding of the implications of likelihood statistics. This suggests an important refinement to Krynski and Tenenbaum's (2007) claim that people have difficulty reasoning about statistics that do not have specified causes. It is true that in the absence of a clearly specified alternative cause people underweighted the relevance of the alternative for judgments of $p(H|D)$. However in Experiment 3, when the "alternative cause" was simply presented as the absence of the focal cause, people had little difficulty seeing the implications of the relevant likelihood statistics.

Overall, the current results are only partly consistent with the causal Bayesian approach to judgments under uncertainty suggested by Krynski and Tenenbaum (2007). They propose that the first stage in making such a judgment is construction of an intuitive causal model of the problem. Our results support this view, by showing that specific causal information about an alternative hypothesis promotes construction of a more complete representation of a judgment problem. The second stage proposed by Krynski and Tenenbaum involves assigning given statistics to the correct components of the mental model. The current studies suggest that this stage poses less of a problem for intuitive statistical reasoning. Once it was made clear that the data could occur in the absence of the focal hypothesis, participants grasped the implications for evaluating likelihood statistics and additional causal knowledge provided little further benefit.

That said it should be noted that although the majority of participants in both unspecified and specified conditions of Experiment 3 revised their judgments in the direction predicted by Bayes rule, almost no one in the high/low and low/high conditions (less than 4%

of participants) made the normatively correct probability estimates. As in many other studies of intuitive statistical reasoning (Corner, Harris & Hahn, 2010; Edwards, 1968; Rottman & Hastie, 2014) people tended to be conservative in their updating of probability estimates when presented with likelihood data. This also raises a question over the third and final stage of the judgment process proposed by Krynski and Tenenbaum (2007). They suggest that once an accurate mental model is constructed and the relevant statistics incorporated in to the model, people will naturally integrate this information in a Bayesian manner. Our results suggest this is only true in a qualitative sense. We suspect that making judgments that are closer to quantitative norms will be difficult without additional instruction in Bayesian theory (cf. Sedlmeier & Gigerenzer, 2001).

Relationship to other accounts of judgment and reasoning involving hypothetical alternatives

The current findings are relevant to accounts of judgment and reasoning which propose that people tend to focus on one hypothesis at a time and avoid the consideration of uncertain alternatives (Evans, 2006; Hayes & Newell, 2009; Mynatt, et al., 1993; Murphy & Ross, 2007). This tendency is summarized in Evans' (2006) *singularity principle* which states that "people construct only one mental model at a time with which to represent a hypothetical situation" (p. 379).

The current work suggests that this principle is only partly true. Consistent with the singularity principle we found that neglect of the alternative largely occurs in the initial representation of the problem. However, our work goes beyond singularity by showing that such neglect can be reduced by making the alternative more salient so that is seen as a competing causal explanation of the observed data. Moreover in contrast to neglect of alternatives at the early stages of problem representation, we have shown that people are readily able to *use* the likelihood statistics associated with each hypothesis, understanding

their implications for $p(H|D)$. The latter result is particularly impressive since in Experiment 3, the base rate probability of the focal hypothesis was high. According to the singularity hypothesis, under such conditions attention should be directed towards the focal hypothesis and away from the alternative (cf. Mynatt, et al., 1993).

Dougherty and colleagues (Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, & Buttaccio, 2014) have developed a detailed model, known as HyGene (short for "hypothesis generation"), of the circumstances under which people generate and use alternative hypotheses in judgments under uncertainty. According to HyGene only hypotheses that are actively maintained in working memory can influence judgments of event probability. The generation of such hypotheses is constrained by the capacity limits of working memory, such that an alternative diagnosis is only generated if it provides a better match to observed data than the poorest-matching hypothesis that is currently active in memory. Hypotheses that have been successfully applied to data in the past are more likely to be stored in long-term memory and to be activated first when a new problem is encountered.

Some aspects of the current results are consistent with the HyGene model. It was certainly the case that the alternative hypothesis was only considered when it was made salient in the problem description. However some of the current findings are potentially problematic for HyGene. There were few differences between the working memory demands on participants in the generic and specified alternative conditions of the first three experiments. Nevertheless robust differences in search for information about the alternative hypothesis were found in each case. Moreover the modest relevance ratings given to the alternative hypotheses in Experiments 2a-b suggest that even when people are aware of these alternatives they give them less weight than the focal hypothesis. These findings suggest that working memory limitations may not be the only factor determining consideration of

alternatives. Instead, the results are more consistent with explanations suggesting that people often misunderstand the in-principle relevance of an alternative hypothesis even when it is available in memory (e.g., Fiedler, 2012).

It should be noted that not all previous studies of diagnostic reasoning have found neglect of alternative causes. Fernbach, Darlow and Sloman (2010) for example, asked people to estimate the likelihood of a cause given an effect. In the standard condition no mention was made of an alternative cause, whereas in the no-alternative condition it was made clear that there were no alternative causes for the observed effect. Diagnostic likelihood ratings were higher in the no-alternative than the standard condition. This implies that those in the standard condition were spontaneously considering alternative causes of the effect when making likelihood ratings (see Oppenheimer, Tenenbaum & Krynski, 2013, for related findings). Notably no difference between likelihoods in the standard and no-alternative conditions was found when the task involved prediction (estimating the likelihood of an effect given a cause) rather than diagnosis.

Fernbach et al.'s results (2010) may however reflect participants' specific knowledge about the target causes and effects. Most of their scenarios involved cause and effect relations that would be relatively familiar to many participants (e.g., rating the likelihood that observed weight loss was due to exercise). In these cases one might expect that alternative causes of the effect (e.g., dieting) are readily available in memory and therefore could influence ratings of the target causal relation. In contrast, the current studies examined intuitions about the relevance of alternative hypothesis in scenarios where the influence of prior causal knowledge was minimized. Our strategy was to examine these intuitions in a diagnostic task where the general structure was likely to be familiar to participants (evaluating whether a given symptom was evidence of a disease) but the focal and alternative causes (i.e. disease labels) were entirely novel (also see Footnote 3).

Our conclusions regarding early-stage neglect may also seem to conflict with work that has found evidence of sensitivity to alternative causes in supervised contingency learning (e.g., Meder, Mayrhofer & Waldmann, 2014; Waldmann, 2000; Waldmann & Hagmayer, 2005). In such studies participants learn about the statistical relationships between effect features (e.g., disease symptoms) and causes (e.g., diseases) through trial by trial observation of cases where the effect and/or the cause are present or absent. A consistent finding is that people have little difficulty learning the conditional probability of a cause given the effect features, and that they readily factor alternative causes into such learning.

These results are important but we need to be cautious in comparing them to the current findings. First, unlike Experiments 1-2b, in contingency learning participants are confronted with evidence both of the effect given the focal cause and the effect in the absence of the cause/presence of an alternative. Second, contingency learning involves explicitly predicting the likelihood of a cause (or causes) given effect features (i.e., $p(H|D)$) whereas the Bayesian problems studied here involved using information about the likelihood of the data given the focal and alternative causes to infer $p(H|D)$. In short, contingency learning studies show that people can learn relevant diagnostic relationships even when multiple causes need to be considered. In contrast, our studies show that people often lack a meta-cognitive appreciation of the relevance of considering alternative causes when estimating $p(H|D)$.

Conclusions

These studies show that the main locus of effect of causal knowledge on judgments under uncertainty is at the early stage of representing the problem. When a clearly specified alternative cause of the data was provided there was an increase in the perceived relevance of this alternative for evaluating $p(H|D)$ and in spontaneous search for information about the alternative. In contrast, when statistical information about an explicit alternative hypothesis was provided, people generally utilized this information in a way that was qualitatively

consistent with normative prescriptions. Additional causal information about the alternative had only a small positive effect on such judgments. These findings clarify and refine previous claims about the beneficial effects of causal knowledge on judgments under uncertainty.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35(5),* 303-314.

Bar-Hillel, M., & Fischhoff, B. (1981). When do base rates affect predictions? *Journal of Personality and Social Psychology*, *41*, 671-680.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*, 241-254.

Bearden, N. J., & Wallsten, T. S. (2004). MINERVA-DM and subadditive frequency judgments. *Journal of Behavioral Decision Making*, *17*(5), 349-363.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology*, *45*(6), 1185-1195.

Corner, A., Harris, A. J., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1625-1630). Austin, TX: Cognitive Science Society.

Crupi, V., Tentori, K., & Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review, 116*, 971-985.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D., & Mynatt, C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory & Cognition*, *24*(5), 644-654.

Doherty, M. E., & Mynatt, C. R. (1990). Inattention to P (H) and to P (D|~ H): A converging operation. *Acta Psychologica*, *75*(1), 1-11.

Dougherty, M., Thomas, R., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. In B. H. Ross (Ed.), *The psychology of learning and motivation*, (Vol. *52*, pp. 299-342). London: Academic Press.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge, England: Cambridge University Press.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed). *Formal representation of human judgment (pp. 17-52).* New York: Wiley.

Evans, J. S. B. (1982). *The psychology of deductive reasoning*. London: Routledge & Kegan Paul

Evans, J. S. B. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378-395.

Evans, J. S. B., Venn, S., & Feeney, A. (2002). Implicit and explicit processes in a hypothesis testing task. *British Journal of Psychology*, *93*(1), 31-46.

Feeney, A., & Evans, J. S. B., & Clibbens, J. (2000). Background beliefs and evidence interpretation. *Thinking and Reasoning*, *6(2)*, 97-124.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*(3), 329-336.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 1–46). London: Academic Press.

Fiedler, K., Freytag, P., & Unkelbach, C. (2011). Great oaks from giant acorns grow: How causal-impact judgments depend on the strength of a cause. *European Journal of Social Psychology*, *41*(2), 162-172.

Fischhoff, B., & Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. *Memory & Cognition*, *12*(4), 402-410.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684-704.

Hawkins, G. E., Hayes, B. K., Donkin, C., Pasqualino, M., & Newell, B.R. (in press). A Bayesian latent mixture model analysis shows that informative samples reduce base rate neglect. *Decision,* http://dx.doi.org/10.1037/dec0000024

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition*, *133*, 611-620.

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition, 37*, 730-743

Hayes, B. K., Newell, B. R., & Hawkins, G. E. (2013). Causal model and sampling approaches to reducing base rate neglect. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*, 678-703.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57*, 227-254.

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*(1), 97-122.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430–450.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*, 232-257.

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

McNair, S., & Feeney, A. (2013). When does information about causal structure improve statistical reasoning? *Quarterly Journal of Experimental Psychology*. *67,* 625-645.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277-301.

Murphy, G. L., & Ross, B. H. (2007). Use of single or multiple categories in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental and computational approaches* (pp. 205-225). New York: Cambridge University Press.

Mynatt, C. R., Doherty, M. E., & Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. *The Quarterly Journal of Experimental Psychology*, *46*(4), 759-778.

Oppenheimer, D. M., Tenenbaum, J. B., & Krynski, T. R. (2013). Categorization as causal explanation: Discounting and augmenting in a Bayesian framework. *Psychology of Learning and Motivation*, *58*, 203-231.

Riege, A. H., & Teigen, K. H. (2013). Additivity neglect in probability estimates: Effects of numeracy and response format. *Organizational Behavior and Human Decision Processes*, *121*(1), 41-52.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*, 109-139.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380-400.

Thomas, R., Dougherty, M. R., & Buttaccio, D. R. (2014). Memory constraints on hypothesis generation and decision making. *Current Directions in Psychological Science*, *23*(4), 264-270.

Trope, Y., & Mackie, D. M. (1987). Sensitivity to alternatives in social hypothesis testing. *Journal of Experimental Social Psychology*, *23*(6), 445-459.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49-72). Hillsdale, NJ: Erlbaum.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547-567.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes' theorem and the additivity principle. *Memory & Cognition*, *30*(2), 171-178.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53-76.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216-227.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly journal of Experimental Psychology*, *20*(3), 273-281.

Footnotes

1. Crupi, Tentori, and Lombardi (2009) argue that the pseudodiganostic search strategy can be considered adaptive from the point of view of maximizing information gain. Note however that this alternative interpretation does not apply to the diagnostic tasks used in the current experiments.

2. The key findings remained robust when relevant choice was analyzed as a categorical variable. In the specified alternative condition rates of choice of $p(\neg H)$ (45%) and $p(D|\neg H)$ (80%) were reliably higher than in the other conditions ($p(\neg H)$: 32%, χ^2 (1) = 5.47, p =.02; $p(D|\neg H)$: 69%, χ^2 (1) = 4.99, p =.03).

3. After the main experiment was completed participants were also asked "Roughly speaking, what percentage of adults in your community do you think would have experienced a red rash on their fingers?" (0-100%). They were instructed to answer on the basis of their "past experience and knowledge". The aim was to address an interesting possibility suggested by a Reviewer that participants were using their prior knowledge of the "red rash" symptom to estimate $p(D)$ and to substitute this for the denominator in Equation 1. This strategy would eliminate the need to consider information about the alternative hypothesis. The data showed that people had highly variable priors about this value (M = 21.3%, SD = 23.3). Given such noisy priors it seems unlikely that participants would use them in preference to the presented response options for diagnostic reasoning. Moreover these priors did not differ between the four experimental conditions in Experiment 2b, $F(3, 460) = 0.60$. Hence differential use of these priors cannot explain the obtained group differences in the rated relevance of $p(\neg H)$.

4. The higher rate of choice of the base rate of the alternative hypothesis in the specified alternative condition (61%) relative to other conditions (42%), remained robust when choice was analyzed as a categorical variable (χ^2 (1) = 12.24, p<.001).

5. The key findings (increase in high/low estimates relative to high/high; decrease in estimates for low/high relative to low/low) remained robust when the data were reanalyzed only with those participants who gave the correct estimate of 80 in stage 1 (n's = 376, 364 for the unspecified and specified conditions respectively).

*Table 1. Experiment 1. Frequency (f) and proportion of search patterns in each condition. The modal search pattern in each condition is shown in **bold**.*

Search Patterns	Unspecified alternative n = 299		Generic alternative n = 303		Specified alternative n = 298		Specified alternative label control n = 304	
	f	*Proportion*	*f*	*Proportion*	*f*	*Proportion*	*f*	*Proportion*
p(H)	6	0.020	4	0.013	7	0.023	20	0.066
p(¬H)	1	0.003	0	0.000	2	0.007	1	0.003
p(D|H)	103	**0.344**	112	**0.370**	58	0.194	81	**0.266**
p(D|¬H)	15	0.050	20	0.066	8	0.027	34	0.112
p(H), p(¬H)	0	0.000	1	0.003	16	0.054	1	0.003
p(H), p(D|H)	67	0.224	60	0.198	23	0.077	67	0.220
p(H), p(D|¬H)	12	0.040	11	0.036	4	0.013	25	0.082
p(D|H), p(D|¬H)	68	0.227	78	0.257	151	**0.507**	62	0.204
p(¬H), p(D|H)	1	0.003	4	0.013	4	0.013	1	0.003
p(¬H), p(D|¬H)	1	0.003	2	0.007	1	0.003	1	0.003
p(H), p(¬H), p(D|H)	0	0.000	0	0.000	1	0.003	1	0.003
p(H), p(¬H), p(D|¬H)	0	0.000	0	0.000	1	0.003	0	0.000
p(H), p(D|H), p(D|¬H)	22	0.074	10	0.033	3	0.010	6	0.020
p(¬H), p(D|H), p(D|¬H)	1	0.003	0	0.000	2	0.007	0	0.000
p(H), p(¬H), p(D|H), p(D|¬H)	2	0.007	1	0.003	17	0.057	4	0.013

Table 2. Summary of the likelihood conditions in Experiment 3 and Bayesian predictions.

Likelihood Condition	Stage 1 base rate $p(H)$	Stage 2 $p(D\|H)$	Stage 2 $p(D\|\neg H)$	Stage 2 $p(H\|D)$ from Bayes rule	Predicted direction of change in Stage 2 estimates
High/High	0.8	0.9	0.9	0.8	No change
Low/Low	0.8	0.1	0.1	0.8	No change
High/Low	0.8	0.9	0.1	0.97	Increase
Low/High	0.8	0.1	0.9	0.31	Decrease

Note: *The stage 2 predictions of p(H|D) were derived from Equation 1.*

Diagnostic Reasoning Task

Panel A: Instructions

Imagine you are a doctor. A patient comes to you with a red rash on his fingers. [*Note that this symptom could be caused by a number of diseases*]. [**Note that this symptom could be caused by two different diseases, Buragamo or Terragaxis**. **For this problem assume that the rash can only be caused by these diseases.**]

What information would you want in order to determine whether the patient has the disease 'Buragamo'?

To help with your diagnosis you can search a computer database for some medical information. Assume that the 4 cards below are different types of information that you can get from the database. Your job is to select all cards necessary to make the diagnosis. But remember that each time you do a search (i.e. select a card) it will cost you time and money. So as a busy doctor you should only select cards that are necessary for the diagnosis.

Click on those cards that you want to select.

Panel B. Response options

% OF PEOPLE WITH BURAGAMO	% OF PEOPLE WITHOUT BURAGAMO
% OF PEOPLE WITH BURAGAMO WHO HAVE A RED RASH	% OF PEOPLE WITHOUT BURAGAMO WHO HAVE A RED RASH

Panel C: Response Options

% OF PEOPLE WITH BURAGAMO	% OF PEOPLE WITH TERRAGAXIS
% OF PEOPLE WITH BURAGAMO WHO HAVE A RED RASH	% OF PEOPLE WITH TERRAGAXIS WHO HAVE A RED RASH

Figure 1. Instructions and search options used in Experiment 1. Panel A shows the instructions given to each experimental group. The instructions in regular font were presented to all participants. The additional instructions in italics were presented in the generic alternative condition. The additional instructions in bold were administered to the specified alternative and specified alternative – label control conditions. Panel B shows the search options used in the unspecified alternative, generic alternative conditions and specified alternative – label control conditions. Panel C shows the search options used in the specified alternative condition.

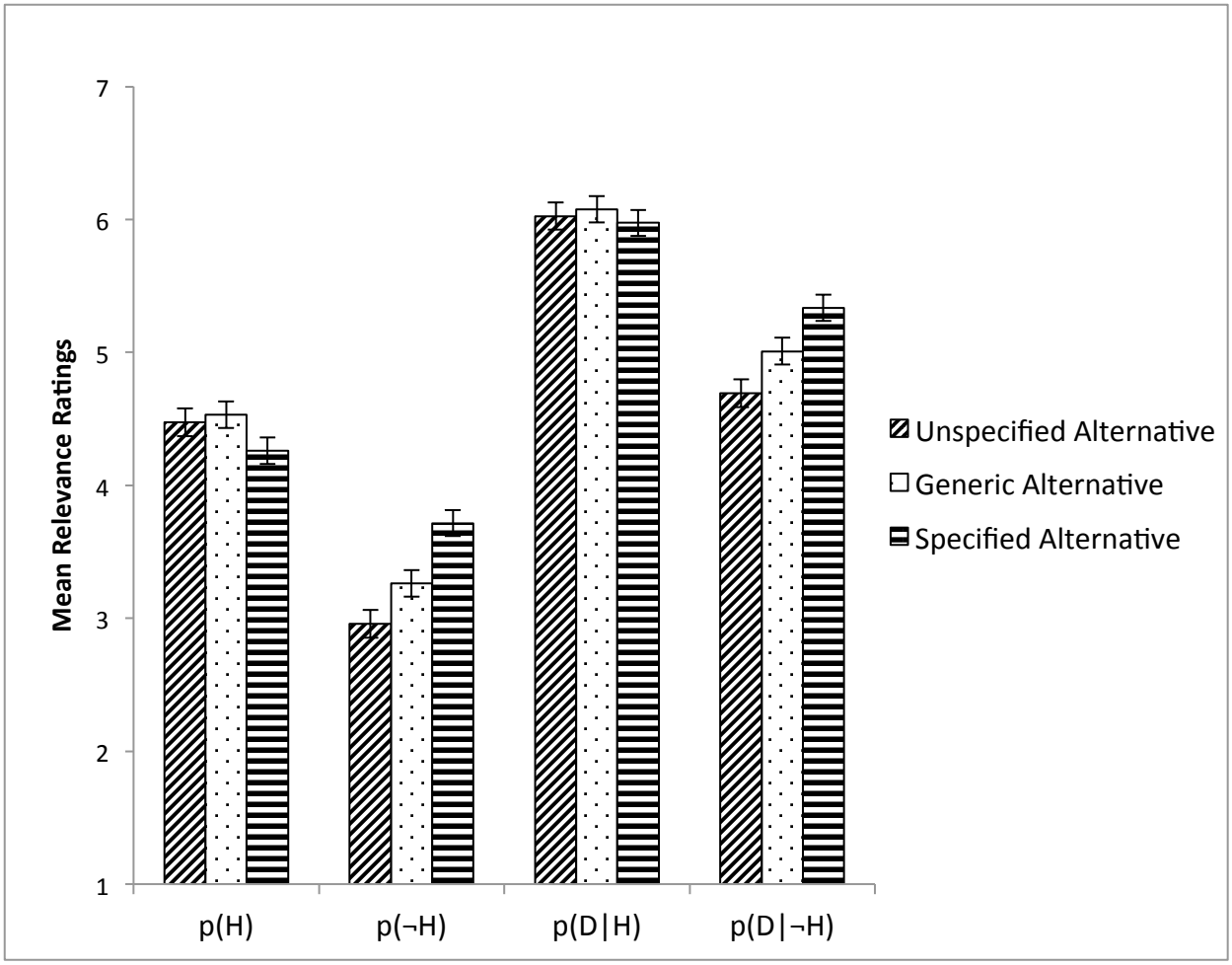

Figure 2. Experiment 2a. Mean ratings of the relevance of each response option (with standard error bars) (1 = not at all relevant; 7 = highly relevant).

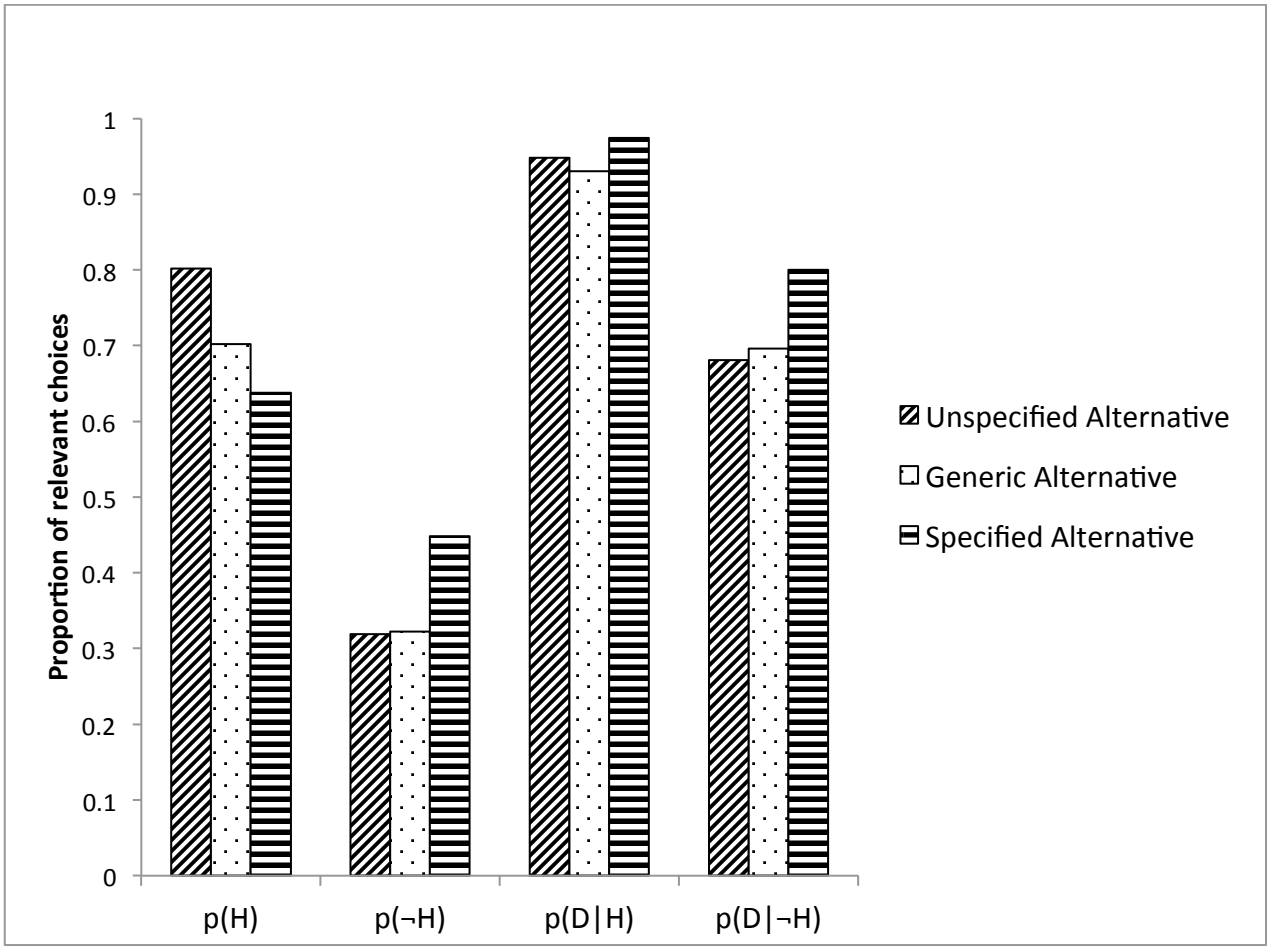

Figure 3. Experiment 2a. Proportion of relevant (vs. irrelevant) choices for each response option.

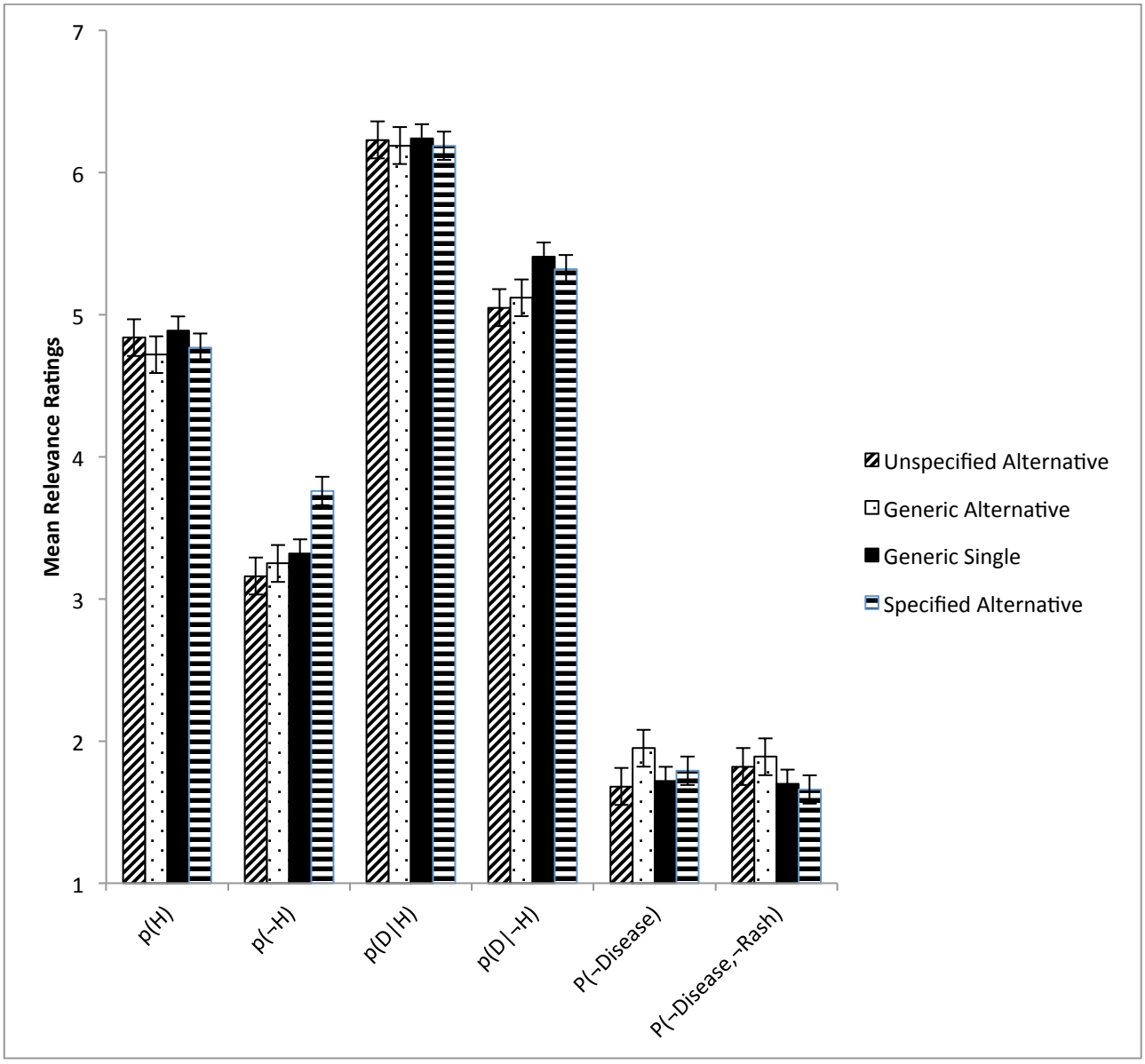

Figure 4. Experiment 2b. Mean ratings of the relevance of each response option (with standard error bars) (1= not at all relevant; 7 = highly relevant).

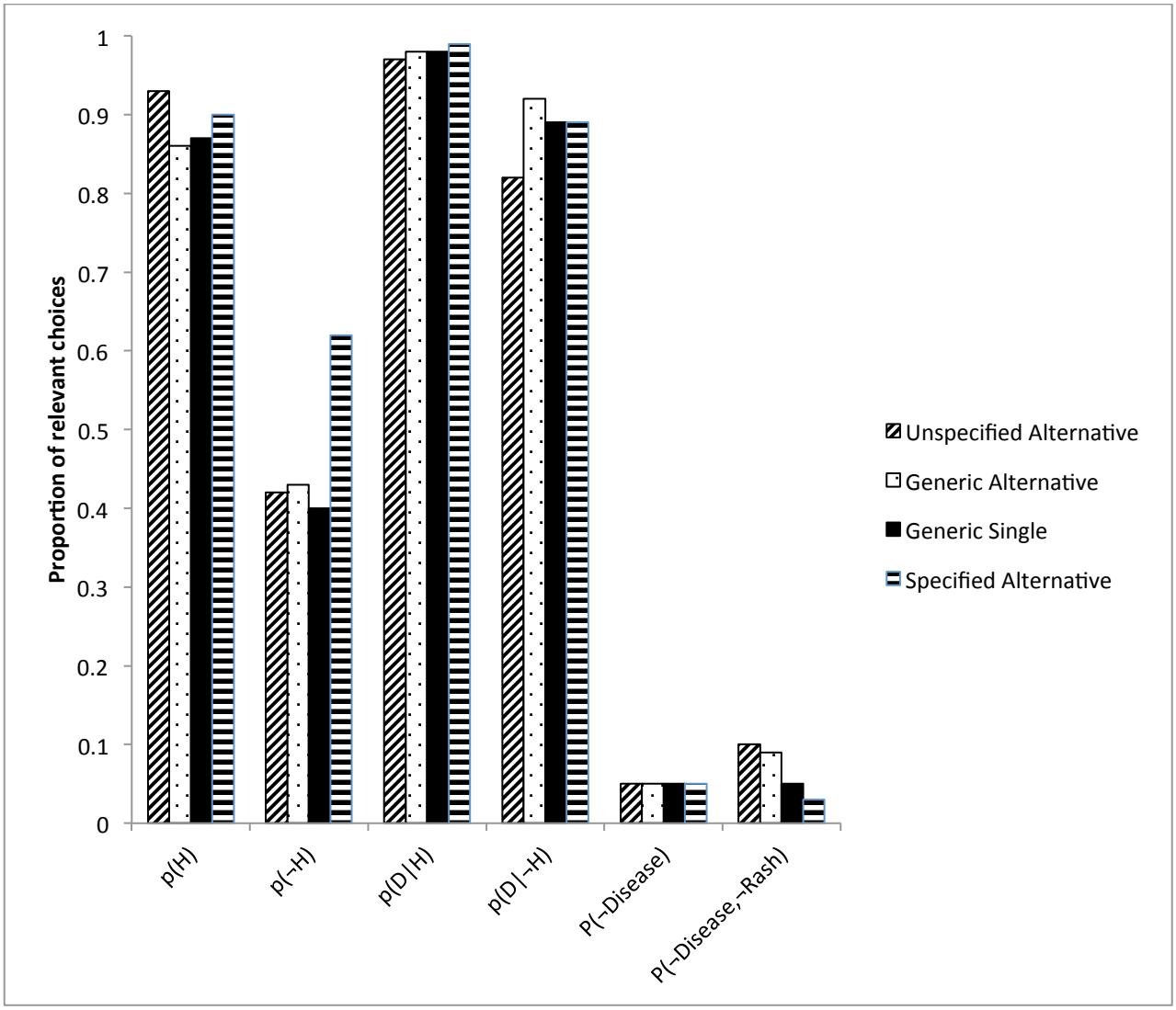

Figure 5. Experiment 2b. Proportion of relevant (vs. irrelevant) choices for each response option.

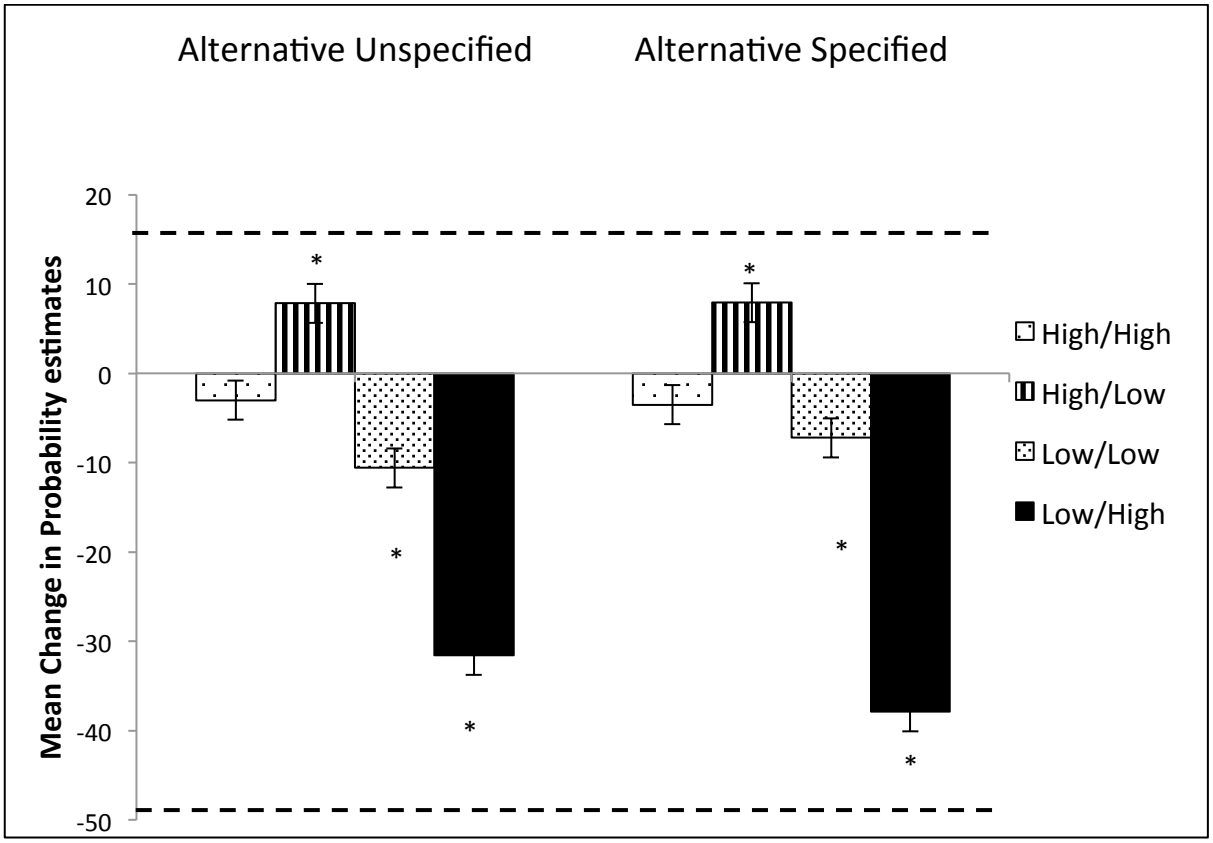

Figure 6. Experiment 3. Mean change in probability estimates from stage 1 to stage 2 (adjusted for covariate). Dashed lines show normative levels of belief revision for the high/low (upper line) and low/high (lower line) conditions.

* Significantly different from zero change, $p<.001$