

Bayesian mind: Making sense of probability-based decision making strategies

Kunjoon Byun

Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea

Bachelor of Arts, University of Wisconsin - Madison, 2011

A Thesis presented to the Graduate Faculty
of the College of William and Mary in Candidacy for the Degree of
Master of Arts

Department of Psychology

The College of William and Mary
August, 2016

APPROVAL PAGE

This Thesis is submitted in partial fulfillment of
the requirements for the degree of

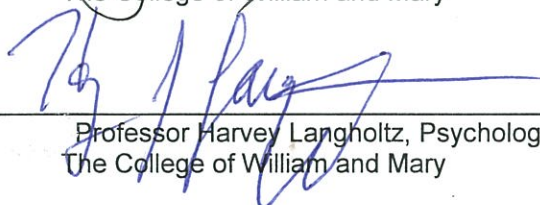
Master of Arts


Kunjoon Byun

Approved by the Committee, June, 2016


Committee Chair
Associate Professor Christopher Ball, Psychology
The College of William and Mary


Assistant Professor Joanna Schug, Psychology
The College of William and Mary


Professor Harvey Langholtz, Psychology
The College of William and Mary

COMPLIANCE PAGE

Research approved by

Protection of Human Subjects Committee

Protocol number(s): PHSC-2014-11-19-9972-ctball

PHSC-2015-03-07-10222-ctball

Date(s) of approval: 12/01/2014

03/22/2015

ABSTRACT

Bayesian probability problems are notoriously difficult for people to solve accurately. Base rate neglect refers to the hypothesis that people ignore base rate information in preference for individuating information when making these probability judgments (Kahneman & Tversky, 1973). Correct answers to base rate neglect problems often require complex Bayesian calculations involving probability information embedded within realistic event descriptions. The past research emphasis on base rate neglect responses for such problems has overlooked the fact that responses can actually vary widely across participants and within participants from one problem to the next. The verbal protocol analyses of participants' decision making processes found in the current study revealed that participants use a variety of cognitive strategies to solve such problems. The between-participant and within-participant variability can result from simple miscalculations to difficulties with translating probability statements into numerical calculations. This translation process can be further complicated by idiosyncratic subjective interpretations of the event descriptions and a basic misunderstanding of objective probability information. The results of the current study highlight that improving an individual's Bayesian reasoning requires the instructional procedure to be tailored to their sources of difficulties, and that individual protocol analyses could help define these instructional procedures.

TABLE OF CONTENTS

Acknowledgements	ii
Dedications	iii
List of Tables	iv
List of Figures	v
Chapter 1. Introduction	1
Chapter 2. Method	18
Chapter 3. Results	20
Chapter 4. Discussion	25
Tables	32
Figures	33
References	46

ACKNOWLEDGEMENTS

This writer wishes to express his appreciation to Professor Christopher Ball, under whose guidance this investigation was conducted, for his patience, guidance and criticism throughout the investigation. The author is also indebted to Professors Joanna Schug and Harvey Langholtz for their careful reading and criticism of the manuscript.

The author also would like to thank the rest of the faculty, staff, and students at the College of William and Mary for sharing valuable experiences and memories.

This thesis is dedicated to my sweet family. Thank you for always having faith in me and supporting me from thousands of miles away. Your love and encouragement is faster than light.

LIST OF TABLES

1. Bayesian computations, success rates, and correlations with numeracy and cognitive reasoning measures found by Byun and Ball (2015).

32

LIST OF FIGURES

1. Total percentage of “other or non-explained” responses that resulted for four past studies that examined base rate neglect.	33
2. Frequency distribution of answers for a sample of three problems used by Byun and Ball (2015).	34
3. Examples of participant responses that highlight consistent probability based responses and inconsistent response patterns.	35
4. Instructions for each of the three Bayesian problems.	36
5. A sample transcript of verbal protocol.	37
6. Frequency distribution of answers for three problems used in the current study.	38
7. A coded example of a correct Bayesian reasoning strategy.	39
8. A coded example of an incorrect probability-based reasoning strategy.	40
9. A coded example of a participant who reduced the four outcomes of the two combined events to just two outcomes.	41
10. A coded example of a participant who uses some of the probability information, but not all of the information.	42
11. A coded example of a strategy involving two sources of information based on personal intuitive judgments.	43
12. A coded example of the intuition-based reasoning approach.	44
13. A descriptive model of how different Bayesian reasoning answers occur.	45

Bayesian mind: Making sense of probability-based decision making strategies

In 1956, Herbert Simon challenged the traditional view of rational decision making which assumed that decision makers have access to the complete information and can recognize consequences of each alternative course of action, as well as, choose to maximize utility or gain. Simon proposed an alternative “bounded rationality” view that described human beings with limited time, knowledge, and computational capabilities to make complex decisions. Simon viewed human decision makers as seeking satisfactory decisions rather than optimal ones. As a result, Simon’s theoretical and research approach focused on the cognitive processes by which people actually make their decisions. This approach was to dominate decision making research for the next 60 years and is integral to the present thesis.

Bayesian Reasoning and Base Rate Neglect

Around the same time as Simon was establishing his research program on general decision making and problem solving, Meehl and Rosen (1955) had come to a similar conclusion, and reported that clinical psychologists tended to make clinical predictions based on subjective intuitions rather than mathematical-supported data. Meehl and Rosen emphasized the importance of base rate data for evaluating the accuracy of psychometric devices, but also noted, that unfortunately base rate information was rarely reported in the literature. Meehl and Rosen made the strong argument that a lack of ability to figure out base rates is the main reason for people disregarding them, and consequently, they urged researchers to collect base rate information, as well as, false negative and false positive rates for psychological tests. They believed that decisions based on this additional objective information would lead to improved clinical decision making. They also

highlighted the use of Bayesian probability formulae to calculate correct judgments based on these data. Their approach highlighted a prescriptive approach for overcoming the bounded rationality inherent to human decision making. Unfortunately, this long standing prescription is still ignored by most people as they rely heavily on their own intuitive decision making skills for the most part, do not involve objective sources of data or mathematical calculations.

Many real-world decisions involve evidence-based judgments regarding the predicted likelihood of an event after incorporating new evidence (e.g., Bayes' Theorem). The mathematical presentation of Bayes' Theorem is as follows:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|\neg A) \times P(\neg A) + P(B|A) \times P(A)}. \quad (1)$$

$P(A)$ and $P(\neg A)$ refer to the base rate probabilities of event A occurring (e.g., have cancer) and event A not occurring (e.g., don't have cancer). $P(B|A)$ is the conditional probability of a related event B (e.g., positive mammogram test result) occurring when event A is true (e.g., person has cancer), whereas $P(B|\neg A)$ is the conditional probability of event B (e.g., positive mammogram test result) occurring when event A is not true (e.g., person does not have cancer). These values provide new evidence regarding the likelihood of event A occurring. Finally, $P(A|B)$ is the updated probability of event A occurring (e.g., person has cancer) if event B is true (e.g., positive mammogram test result). This value results from the combined sources of base rate information and individuating test information. Unfortunately, when people are faced with these decisions in daily life they rarely follow these normative steps for making their final judgments (Azjen, 1977; Bar-Hillel, 1980; Edwards, 1968; Gigerenzer & Hoffrage,

1995; Hammerton, 1973; Kahneman & Tversky, 1972; Krynski & Tenenbaum, 2007; Lyon & Slovic; 1976; McNair & Feeney, 2014; Tversky & Kahneman, 1982).

Edwards (1968) was the first to report that people do not make correct Bayesian decisions, and that they act very conservatively when adjusting their prior belief (base rate) after given new evidence. Edwards tested this claim by giving his participants the following Bayesian reasoning problem: *“There are two book bags, one containing 700 red and 300 blue chips, the other containing 300 red and 700 blue. Take one of the bags. Now, you sample, randomly, with replacement after each chip. In 12 samples, you get 8 reds and 4 blues. What is the probability that this is the predominantly red bag?”* The correct Bayesian answer is 0.97, but most of Edward’s participants chose 0.70 as their answer. This conservatism correction highlighted the difficulties people have solving Bayesian reasoning problems.

But it was the very influential series of papers by Kahneman and Tversky in the 1970s that not only expanded on Simon’s bounded rationality work but also further highlighted the importance of examining how people really make their decisions. They proposed that people used cognitive “heuristics” to make satisfactory judgments that overcame their limited cognitive capacity for fully processing the complex amounts of decision information inherent to many real world decisions. They argued that intuitive judgment processes are not quantitatively simpler than normative models, but rather that they fall into a different qualitative category of processing. They also speculated that a heuristic approach to making decisions would also lead to predictable biases in the final decisions. This early work of Kahneman and Tversky focused on three general heuristics: (1) availability, (2) representativeness, and (3) adjustment and anchoring.

The representative heuristic was proposed by Kahneman and Tversky (1972) to explain why base rate information was often neglected when making decisions (base rate neglect). Their initial definition of the representative heuristic was that people assess probability information for an uncertain event by “the degree to which an event is similar in essential characteristics to its parent population and reflects the salient features of the process by which that probability information is generated” (Kahneman & Tversky, 1972, p. 431). Kahneman and Tversky (1973) illustrated the use of this heuristic with a decision problem that is now referred to as the ‘Engineer and Lawyer Problem’. They gave their participants a brief personality descriptions of several individuals who were randomly drawn from either a group made up of 70 engineers and 30 lawyers or a group made up of 30 engineers and 70 lawyers. Participants were required to determine the probability for each description that it belonged to an engineer rather than to a lawyer. For example, “*Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles. What is the probability that Jack is one of the 30 engineers in the sample?*” (Kahneman & Tversky, 1973, p. 241). Kahneman and Tversky found that participants gave the same probability judgments for both group conditions, and that a specific description of an engineer lead participants to ignore prior probabilities (i.e., base rate information). In fact, participants even ignored prior probabilities when they were given specific but irrelevant evidence (e.g., marital status and reputation at work). However, when participants were not presented with the specific

personality descriptions, participants correctly used the prior probabilities (i.e. base rates) to make their judgments.

Tversky and Kahneman (1982) also presented participants with a more complex Bayesian problem that described a common real world problem where prior evidence (e.g., base rate information) is updated with new evidence (e.g., individuating information) to provide an accurate probability judgment of an outcome. In their study, participants were given the following scenario to read (now known as the ‘Taxi Cab Problem’): *“A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green knowing that this witness identified it as Blue?”* (p. 156-157). Tversky and Kahneman (1982) reported that the modal answer obtained from participants was 80% for the Taxi Cab problem, and this result again highlighted the common practice of neglecting base rate information. Other researchers quickly replicated this finding (Bar-Hillel, 1980; Lyon & Slovic, 1976), and Bar-Hillel (1980) even suggested that base rate neglect should be considered a “matter of established fact” (Bar-Hillel, 1980, p. 215). However, Bar-Hillel also noted that the representative heuristic explanation Kahneman and Tversky provided for their Engineer and Lawyer problem did not apply as well to the more complex Taxi Cab problem.

Consequently, Tversky and Kahneman (1977, 1980) proposed an alternative causal schema explanation for the neglect of base rate information with the taxi cab problem, which has its origins in the causal interpretations proposed by Ajzen (1977). Ajzen argued that people look for causal links for events or behaviors, and therefore participants' predictions will be influenced by whether the given information suggest casual factors or not. He proposed that even statistically relevant normative information will be neglected if it is not causally significant, and that this information is mainly used when no causal model can be represented. Ajzen explained the engineer and lawyer problem findings as resulting from a population base rate for each occupation that did not cause any member of the sample to be related to either job whereas the specific descriptions (e.g., traits, ability, and others) of the individuals did provide casually significant information.

In response to Ajzen's argument regarding the causal relevance of the information provided in such problems, Tversky and Kahneman (1977) revised the role of the representative heuristic to be overridden by a casual interpretation, if such an interpretation was readily available. For example, when Tversky and Kahneman (1977) changed the base rate information for the Taxi Cab problem to now describe the frequency of previous accidents "*Although the two companies are roughly equal in size, 85% of cab accidents in the city involve Green cabs, and 15% involve Blue cabs*" (p. 31), the proportion of base rate neglect answers was significantly decreased. They also found such causal details were directional in that this information was more influential than diagnostic data, and that causal details were also specific in their influence and only considered when they fit the causal schema.

More recently, Krynski and Tenenbaum (2007) found further support for a causal explanation for base rate neglect in Bayesian reasoning. They argued that people make use of valuable causal knowledge to make accurate judgments in daily life, and that it is hard to highlight this in traditional laboratory experiments. They often observed base rate information was neglected by participants when given a typical Bayesian problem (e.g., mammogram test that detects breast cancers), but when they provided a reason for why this test might produce an erroneous false positive result (e.g., due to harmless benign cyst), the proportion of base rate neglect answers decreased sharply. They suggested that the addition of this information helped people to set up a causal model to assign appropriate values for making correct Bayesian judgments.

In fact, Hammerton (1973) had already proposed a similar experiential-based explanation for the general neglect of probability information by participants when given such problems to solve. Hammerton presented the following Bayesian problem to his participants:

- 1. A device has been invented for screening a population for a disease known as psylicrapitis.*
- 2. The device is a very good one, but not perfect.*
- 3. If someone is a sufferer, there is a 90% chance that he will be recorded positively.*
- 4. If he is not a sufferer, there is still a 1% chance that he will be recorded positively.*
- 5. Roughly 1% of the population has the disease.*
- 6. Mr. Smith has been tested and the result is positive. (p. 252).*

His participants consistently estimated a much higher probability than the actual answer of around 48%. Changing the order in which the information was presented and taking out some of the probability information did not remove this response bias. However, when Hammerton changed the description of the problem to a mechanical device which screens engine parts for internal cracks, his participants now estimated much lower probabilities (65%). Hammerton believed that participants had rigid prior expectations regarding the infallibility of medical diagnostic tests, and that their previous experiences could influence the degree to which case-specific information could affect their probability judgments. Hammerton's explanation for base rate neglect is not based on a lacking or faulty causal schema as proposed by Kahneman and Tversky, but rather the influence of personal experiences and beliefs for subjectively accepting or rejecting objective sources of probability information.

However, the prior expectations explanation proposed by Hammerton was criticized by Lyon and Slovic (1976) because there are surely no rigid prior evidence from a typical participant's experiences about witness identification infallibility that would be expected for the classic Taxicab problem. Lyon and Slovic (1976) compared the Taxicab problem with a similar structure involving a different context (a machine to detect light bulb failure). They did not find a problem context effect, but support from a null hypothesis prediction is hardly a strong case for rejecting Hammerton's explanation. However, as the current thesis will highlight, making broad theoretical conclusions based on the answers alone provided by participants in these kind of tasks is also fraught with possible confounds. The current thesis will highlight the theoretical benefit of allowing

multiple explanations to play a role in accounting for the variety of responses possible for complex Bayesian decision problems.

Bar-Hillel (1980) pointed out that the causality argument is an incomplete explanation, because the causality argument accounts for when base rates will be ignored but not why base rates will be ignored. She argued that the relevance of the base rate information determines whether people would use this information, and that causal relevance was a special case of relevance. Bar-Hillel proposed that relevance is the main factor why people ignore base rate information, because they view base rate information as irrelevant. She conducted a number of experiments using multiple Bayesian problems in different contexts (e.g., Taxicab accident, suicide, and dreaming), and she found that participants only integrated the two sources of information (base rate and individuating) when both sources of information are equally relevant. She also suggested that providing participants with multiple Bayesian problems will help highlight the relevance of both sources of information. This notion of relevance was never clearly defined by Bar-Hillel, but it does seem to relate to both prior expectations and causal schemas. However, as you will see later when the current study examines within-participant variability in Bayesian reasoning responses, relevance of the two sources of information does not appear to occur naturally with practice.

The previous research suggests that emphasizing the relevance (especially causal relevance) of the two sources of information when solving Bayesian reasoning problems will overcome the base rate neglect bias. However Gigerenzer and colleagues suggest that additional cognitive difficulties still need to be overcome when making these decisions. They argue that the main reason why Bayesian reasoning is so difficult for

human participants to perform is because the human mind was not designed to process information in a probability format (Gigerenzer, 1996; Gigerenzer & Hoffrage, 1995). Gigerenzer (1996) suggested that natural frequencies are easier to process because of our evolutionary history for dealing with naturally occurring frequency data. Gigerenzer also argued that presenting information in natural frequencies gives people more information about the size of a certain reference class, and therefore, helps participants make inferences that more closely resemble the normative Bayesian approach.

Gigerenzer and Hoffrage (1995) gave their participants multiple Bayesian problems to solve that were presented in either frequency or standard probability formats. The following Bayesian problem highlights how frequency information differs from probability information: *“10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who received a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ____ out of ____.”* (p. 688). They found that the number of participants who gave correct (or close to correct) Bayesian answers was 16% for problems presented to participants using the probability format, whereas 46% of the participants provided correct answers to problems presented using a natural frequency format. This is a substantial improvement in Bayesian reasoning, but we should highlight that even after this change in format, over 50% of these Bayesian problems are still being calculated incorrectly. Again highlighting that more than one explanation is needed to account for the variability in participants’ responses with Bayesian reasoning tasks.

Lewis and Keren (1999) also pointed out that the frequency argument posed by Gigerenzer has a major confound when comparing probability and frequency formats; the use of joint versus conditional data presentation in given sampling information. They argued data presented in frequency format (e.g., 10 out of every 1000 women have cancer, 8 of every 10 women with cancer get positive test results, and 95 out of every 990 women without cancer get positive test results) was much easier to comprehend and compute as it described events as jointly occurring, whereas conditional statements were used to describing problems in probability format (e.g., 10 out of 1000 women have cancer, 800 out of 1000 women with cancer get positive test results, and 96 out of 1000 women without cancer get positive test results).

Ayal and Beyth-Maron (2014) suggested that the main difference between these arguments derived from the different number of mental steps individuals needed to compute in each case. They manipulated the number of computations needed to solve a typical Bayesian problem (1 to 4) and the type of data presentation used for each problem (frequency versus probability). For example, some participants only needed to calculate the final computation in the Bayesian calculations whereas other participants needed to calculate three additional earlier computations in the sequence of Bayesian calculations. When participants only needed one calculation, the correct answer was provided 55.2% of the time when the problem's information was presented in a frequency format, whereas only 32.3% of the problems were calculated successfully when the probability format was given. This finding supports Gigerenzer's natural frequencies argument. However, this advantage in performance was diminished with each additional mental computation required. When participants needed to correctly make four computations, the

difference in the success rate between the two formats was less than 4% (10% for frequency format vs. 13.8% for probability format). These findings are a problem for Gigerenzer's frequency explanation because this proposition assumes that the frequency advantage should be greater when more computations are required, but this was not found to be the case. These results suggest that there are more challenges for participants when performing these computations than just the format of the data representation.

One Explanation or Many

When Bar-Hillel (1980) replicated the classic 'Taxicab' problem, she found that only about 10% of her participants gave the normative Bayesian probability answer while another 36% of participants ignored base rate information (i.e., base rate neglect). The other 54% of her participants made non-normative answers which could not be explained by base rate neglect. Many participants gave judgments that defied any clear logical or mathematical computations. This wide range of non-normative judgments even results for highly educated individuals and individuals with strong numeracy skills (Lipkus, Samsa, & Rimer, 2001; McNair & Feeney, 2014), and is typical for all previous research involving Bayesian reasoning (refer to Figure 1 for additional examples). But past research on Bayesian reasoning has typically focused on a single explanation for the non-normative judgments made by participants, and researchers have ignored other non-explained responses, even when these responses constituted the majority of cases. This point is further illustrated by the attempts of researchers to improve Bayesian reasoning, such as by providing causal information (Kahneman & Tversky, 1980; Krynski and Tenenbaum, 2007) or by providing information in a frequency format (Gigerenzer & Hoffrage, 1995). Although these attempts have been successful at improving the success

rates for Bayesian calculations, the majority of participants in these studies still do not show any benefits from these experimental manipulations. Research is needed that does a better job of describing the many different ways that participants approach Bayesian reasoning problems, and with this knowledge, researchers will be better able to target specific interventions to combat biases that result from these differences.

After synthesizing the results from previous research on Bayesian reasoning, it is evident that the range of responses that result reflect the interaction between an individual's subjective, intuitive reasoning and the individual's ability to perform the necessary computations. Gigerenzer and Hoffrage (1995) alluded to this interaction when they stated that "*Maybe the mind does a little of both Bayesian computation and quick-and-dirty inference.*" (p. 685). Unfortunately, they also concluded that incorporating both perspectives would make very little theoretical progress because they believed this would just compromise or weaken the two different viewpoints. Gigerenzer and Hoffrage also argued those viewpoints were incomplete because they focused on cognitive processes not cognitive algorithm (frequency or probability) and information format. The findings of the current study will reveal that incorporating both perspectives is actually the best way forward for understanding the theoretical underpinnings of Bayesian reasoning, as well as prescribing methods for improving Bayesian reasoning skills in the general population.

Between-participant and Within-participant Variability

The literature review presented so far highlights the considerable variability in judgments made by participants when given a single Bayesian decision problem (between-participant variability). However, this is not the only source of variability when

examining Bayesian reasoning. When participants are given multiple Bayesian problems there is also evidence of within-participant variability in their responses.

Byun and Ball (2015) presented 110 participants with nine Bayesian reasoning problems to solve. Each Bayesian problem described different real world events that a person could experience, but the probability information provided for each problem was designed to be numerically quite similar. As expected from previous research with such problems, a wide range of answers were provided for each and every problem (between-participant variability). The wide distribution in responses are highlighted for a sample of problems in Figure 2. In addition, Byun and Ball found that the overall distribution of responses could differ from one problem to the next even though the probability information was very similar in each case. This finding suggested that perhaps there was some degree of within-participant variability in the responses provided by the same participant to different Bayesian problems.

To highlight this within-participant variability in responses, Byun and Ball (2015) mathematically calculated measures of response inconsistency across the nine Bayesian problems using known response types for comparisons (e.g., correct answer, base rate neglect answer). They found that some participants use a consistent probability based strategy (i.e., very low response inconsistency) even if their answers were generally incorrect. Other participants provided such large response inconsistency scores that a common probability-based strategy was unlikely to have been used by these participants (refer to Figure 3). In fact, more than half of the participants showed these inconsistent response patterns, suggesting that within-participant variability can also account for the wide range of answers typically found with Bayesian reasoning tasks. Unfortunately, it is

very difficult to determine a participant's specific judgment strategy when researchers only have the participants' answers to determine these strategies.

One possible explanation for these two sources of variability could still be that the mathematical computations undertaken by the participants are so incorrect that they could produce substantial variability in responses (even across problems). To test this explanation, Byun and Ball (2015) also provided their participants with the classic Taxicab problem and asked their participants to specifically calculate each computational step of this Bayesian problem by isolating each verbal description for that probability calculation. The percentage of correct answers decreased as participants moved from one computation to the next, and the number of answers given by participants for each computation quickly multiplied (refer to Table 1). These results suggest that many participants have considerable trouble translating verbal probability descriptions into numerical calculations, and that early miscalculations could combine to compound the variability in the final answers obtained. If this is the case, however, one would expect that the participant's numeracy ability (Lipkus, Samsa, & Rimer, 2001; Lipkus & Peters, 2009) or cognitive reasoning ability (Frederick, 2005) would successfully predict each participant's ability to successfully complete each computational step.

The Numeracy scale was developed by Lipkus, Samsa, and Rimer (2001) to measure the numerical capabilities necessary for performing the mathematical operations inherent to probability calculations. The three item Cognitive Reflection Test (CRT) was developed by Frederick (2005) to assess the ability of an individual to suppress an intuitive, heuristic (but wrong) response in favor of a reflective, analytical (correct) response. Researchers have found the CRT to correlate with an individual's knowledge

of statistical concepts (Obrecht, Chapman, & Gelman, 2007) and tendency to produced biased responses, such as the gambler's fallacy and outcome bias (Toplak, West, & Stanovich, 2011).

Byun and Ball (2015) incorporated the Numeracy scale and a two-item version of the CRT in their study. They found significant correlations for these measures with the performance of participants on the early computation steps of the Taxi Cab problem (refer to Table 1). However, as the computation steps became more complex, these correlations became smaller and non-significant predictors of computational success (refer to Table 1). These findings highlight that computational calculations (or more appropriately – miscalculations) do play a role in explaining the variability of participant responses, but these findings also suggest that other factors could play a role in the unaccounted response variability.

Verbal Protocol Analysis

The goal of the current study was to fully explain the sources of variability in Bayesian reasoning that is evident in typical base rate neglect problems. Previous explanations are based solely on the answers provided by participants to usually one Bayesian problem. However, there are different ways of coming up with the same incorrect answers, and these answers cannot clearly depict a misstep or miscalculation in the computation process. Our goal was to examine the thought processes that participant followed when attempting to solve complex Bayesian decision problems. Verbal or think-aloud protocols have been used successfully in the past to examine the cognitive processes and strategies used by participants when making decisions (e.g., Ball, Langholtz, Auble, & Sopchak, 1998). The verbal protocol procedure requires the

participant to speak out aloud their thought processes as they make their decisions or solve problems (concurrent verbal protocol), or speak out aloud their thought processes after they have provided their decision or answer (retrospective verbal protocol).

However, researchers have raised concerns that the verbal protocol methodology could affect the thought processes of the participant or inadequately elicit all of the thoughts that went into making a decision or solving a problem (Nisbett & Wilson, 1977).

Regardless of these concerns, the verbal protocol procedure still allowed us to conduct the first detailed analysis of the cognitive processes involved in Bayesian reasoning.

Consequently, a much richer source of individual data will be provided in this study than has been achieved in past research that has relied on the participant's final answers as the sole source of their research conclusions.

Pilot research was conducted to establish the best verbal protocol procedure when examining Bayesian reasoning. The concurrent verbal protocol procedure was rejected by the experimenters because it interfered too much when solving these complex, information laden problems. The retrospective protocol procedure did not interfere with the thought processes needed to solve such problems, and also provided good data when the protocol was undertaken as soon as possible after the participant provided their answer. We also found that additional verbal prompts provided by the experimenter after the retrospective protocol data was collected, could also help pinpoint the use or non-use of probability information when solving each problem. We hypothesized that the strategies participants use to solve Bayesian reasoning problems will vary in the type of (probability and story) information they use and how they use this information.

Furthermore, we speculated that the reasons for these differences (between-participant

and within-participant) will be a function of their subjective interpretation of the Bayesian problem's story context and their ability to perform the necessary mathematical computations. Participants will vary in the relative influence of these factors, and we hope to document where these factors will influence the reasoning processes during each participant's computations.

Method

Participants

Thirty-four undergraduate students (17 male and 17 female) at the College of William & Mary participated in this experiment for course credit. We omitted data from three participants who were not native English speakers. The final sample of 31 participants consisted of 15 men and 16 women with a mean age of 18.61 years ($SD = 1.02$).

Apparatus

The experimenter sat in front of a computer and controlled the experiment using Microsoft PowerPoint. Participants sat in front of a second monitor connected to the experimenter's computer. A standing screen separated the experimenter from the participant so no visual contact between the two people was possible. A microphone was placed next to the monitor and in front of the participant, and the participant's verbal responses were recorded using the Audacity software. Participants were provided with a pen and scratch paper if they needed to write down any calculations.

Procedure

After the participant provided Informed Consent, the think-aloud Bayesian reasoning task was presented by the computer to each participant. The computer

presented the instructions for performing the task: *“In this experiment, you are going to read three scenarios that describe real world decisions people are faced with making. Many decisions like these, require us to incorporate probability information in our decision making. Please treat each decision as seriously as you would if each decision scenario applied to you or a close family member or friend.”*

Participants were asked to solve three Bayesian reasoning problems which were adapted from the problems used by McNair and Feeney (2014). The problem descriptions were edited by the author to make them more realistic to the participants and appear less like mathematical problems. The descriptions and images presented for each problem are displayed in Figure 4.

After the participant gave their answer to each of the three problems, the following retrospective verbal protocol instructions were provided: *“Now we would like you to go back through each decision problem and explain your thought processes in coming up with each answer. Please tell the experimenter in as much as detail as possible the thoughts that came to mind during this decision making process, no matter how relevant or important they may sound to you now. Please tell us what you were thinking about when you read the specific items of information (numerical and non-numerical) that make up each decision problem. Please explain the thoughts that went into interpreting each problem and how you used the probability information provided to solve the problem. Please try to provide as much detail as you can regarding the thinking you followed from the start to the end of the mental decision making steps you undertook.”* To facilitate this recall, the participant was presented with the same problem descriptions again (refer to Figure 4) with an additional instruction displayed at the

bottom of each problem: “*Please describe all of your thought processes.*” If the participant did not say anything about their use or non-use of the specific probability information provided for each problem, the experimenter prompted them with specific questions about whether they had used that information and why they did or did not use this information. After the participant provided their verbal protocol for each problem, the experimenter asked the participant to rate the credibility of the probability information provided for each problem (0 = not credible to 7 = very credible) and to rate how realistic they found the description of the situation (0 = not realistic to 7 = very realistic). The order of the three Bayesian problems was counterbalanced across the participants. The experiment took approximately 30 minutes to complete.

Data Coding

The recorded verbal descriptions (including answers to prompts) provided by the participants were then transcribed. The transcriptions were segmented to highlight statements that related to the thought processes involved in making their judgments. These statements were coded for content that related to subjective interpretations of the information provided and for content that related to objective use of the probability information provided. An example of a segmented transcript from a participant and the coding that resulted for this transcript is provided in Figure 5. The reliability of the content analysis was tested by using two coders to code the same random sample of transcripts, and the inter-coder agreement approached 100% agreement.

Results

Answers

The current study provided a similar patterns of responses as reported previously by Byun and Ball (2015) when giving participants multiple Bayesian probability problems to answer (refer to Figure 6). The large number of answers and the wide range of values provided for the Pregnancy Test problem (number of answers = 15; range of answers = .16 – 90%), the Sniffer Dog problem (number of answers = 16; range of answers = 1.4 – 90%), and the Spam E-mail problem (number of answers = 20; range of answers = .08 – 100%) are consistent with the findings of Byun and Ball (2015). The base rate neglect answer for each problem was not the modal answer provided by participants as is often found in previous research (Bar-Hillel, 1980; Byun & Ball, 2015; Cohen & Staub, 2015; Krynski & Tenenbaum, 2007; McNair & Feeney, 2014, 2015).

Realism and Credibility of Problem Descriptions

The realism ratings of the event descriptions for the Pregnancy Test problem ($M = 5.00$, $SD = 1.77$), Sniffer Dog problem ($M = 4.97$, $SD = 1.78$), and Spam E-mail problem ($M = 4.52$, $SD = 1.63$) were all above the mid-point for this scale (3.5), suggesting that the majority of the participants believed these events were similar to those experienced in real life. In addition, the participants' ratings of the credibility of the probability information provided for the Pregnancy Test problem ($M = 5.34$, $SD = 1.70$), the Spam E-mail problem ($M = 5.06$, $SD = 1.32$), and the Sniffer Dog problem ($M = 4.39$, $SD = 1.50$) were also above the mid-point for this scale (3.5), suggesting that most participants also found the probability values to be consistent with what they would expect for real-life events of these types.

Correct Bayesian Reasoning

Two participants (6%) used the correct Bayesian reasoning steps to calculate all three of their answers. These participants did not provide subjective interpretations of event information, and provided the necessary probability calculations given by Bayes Theorem (refer to Figure 7). One participant used the standard probability format in making these calculations while the other participant converted the probability values to frequencies before performing their calculations.

There were two other participants who used correct Bayesian calculations for one problem but not for the other two problems. One of these participants was particularly interesting, because his data highlights the within-participant variability issues that can result with complex real-world decision making problems like those used in the current study. This participant used a very simple intuition-based reasoning approach to give non-normative answers for the first two problems, but then changed his strategy for the final problem and used the correct Bayesian calculations for this problem. This participant clearly had the mathematical knowledge to utilize the normative Bayesian reasoning procedure, but chose not to do so for the first two problems.

In total, only eight (9%) out of the 93 total problems completed by all participants were correctly solved by the participants in this study. This result again highlights the need for a better understanding of the many possible reasons people are not using the normative Bayesian computations when making these types of decisions.

Incorrect Probability-based Reasoning – Missing or Miscalculated Step(s)

The next group of participants can be categorized into two distinctive sub-groups who made use of some or all of the probability information provided for each problem, but then made errors in how they used this information. In the first sub-group, six

participants (19%) incorporated all of the probability information provided in one or more of the problems, but they combined this information incorrectly when calculating their final answers (refer to Figure 8). Some examples of miscalculations used by these participants are: using $(P(H) + P(\neg H))P(D|H)$ instead of $P(\neg H)P(D|H)$, and using $(P(H)P(H|D)) / (P(\neg H)P(H|\neg D))$ instead of $(P(H)P(H|D)) / (P(H)P(H|D) + P(\neg H)P(H|\neg D))$.

The other group of three participants (10%) correctly calculated $P(\neg H)$, $P(H)P(D|H)$, and $P(\neg H)P(D|\neg H)$ in at least one of the problems, but they then selected one of the conditional probability values as their final answer (refer to Figure 9). These participants often had a problem seeing that four outcomes are possible when combining two events (e.g., event 1 = drugs in suitcase and event 2 = dog detects drugs) and involve ‘hits’ (e.g., correct detection of drugs in suitcase), ‘false positives’ (e.g., incorrect detection of drugs in suitcase), ‘misses’ (e.g., missed detection of drugs in suitcase), and ‘correct rejections’ (e.g., did not detect drugs in suitcase when they were not present). These participants incorrectly perceived that there were only two outcomes for combinations of two events (e.g., drugs were present or drugs were not present), and the two outcomes they chose related to the event they perceived as more important to the question.

Incorrect Probability-based Reasoning – Failure to Utilize All Information

This group of participants attempted to use some of the probability information to calculate their answers, but they did not use all of the probability information necessary to do so correctly (refer to Figure 10). Some stopped too early in their calculations (e.g., $P(H)P(D|H)$) and/or combined probability information in incorrect ways (e.g.,

$P(H)P(\neg H)$, $(P(D|H) - P(D|H)P(D|\neg H))$). Some of these participants even failed to calculate $P(\neg H)$, that is a vital early step in the correct sequence of Bayesian computations. Eight participants (26%) solved one or more of the problems in this way.

Intuition-based Reasoning Combined with Probability-based Reasoning

Participants in this group only utilized one or two pieces of probability information as simple guides for making their intuitive judgments. They often incorrectly rejected sources of probability information as irrelevant (e.g., ‘it involved other people’, ‘it was what happened in the past’, ‘the accuracy of dogs is not relevant’, ‘because they have made their detection already’). These participants also often incorporated their own personal beliefs or experiences in making these decisions of relevance (refer to Figure 11). Twenty-one (68%) participants used this approach for at least one of the problems.

Intuition-based Reasoning

The last group of participants often focused on one source of probability information and then subjectively adjusted it up or down depending on the other source of probability information. These participants often appeared to infer that the probability information is only given to simply help them interpret the story (e.g., dogs are accurate – 90% is a good level of accuracy). A couple of participants did not even bother utilizing any of the probability information; preferring to rely on their own intuitive-judgments of what this probability should be. They sometimes related the problem to their own personal experiences without a need to incorporate any of the objective probability information provided in the problem description (refer to Figure 12). Thirteen participants (42%) used this very simple approach to determining their answers for at least one of the problems.

Discussion

Since Kahneman and Tversky labeled base rate neglect as a cognitive bias in the 1970s, the bias has been regarded as a strong and consistent phenomenon for decades. The definition of base rate neglect as ‘the neglecting of base rate information in favor of individuating information’ describes the phenomenon but does not help us understand why such a bias occurs. In addition, researchers who have investigated this bias have ignored the wide range of possible answers provided by participants that go beyond the one described in the previous definition. Base rate neglect answers are often not the most frequent answer provided by participants in previous research (refer to Figure 1), and some participants even show the opposite bias by relying solely on base rate information and neglected the individuating information (refer to Figure 6 for an example from the current study).

The variability of answers provided across participants when solving these Bayesian reasoning problems is further compounded by the within-participant variability in answers when multiple problems are given to participants (Bar-Hillel, 1980; Byun & Ball, 2015; Cohen & Staub, 2015; McNair & Feeney, 2014, 2015). The base rate probability information and individuating probability information (hits and false alarms) are usually varied across problems in these research studies. The values given still satisfy the general experimental requirement of a low base rate probability for the primary outcome, [e.g., having cancer $P(H)$] in comparison to a much larger complementary probability for the alternative outcome [e.g., not having cancer $P(-H)$]. The individuating information generally describes a relatively high hit rate for the second source of information [e.g., positive mammogram test result for women with cancer

$P(D|H)$] with a corresponding low false positive rate [e.g., positive mammogram test for women without cancer $P(D|\neg H)$]. The event descriptions will usually change from one problem to the next (e.g., judging cancer likelihood, judging the success of a spam filter) and can reflect very different real world decision making contexts. Unfortunately, past research is unclear whether the different answers provided by participants across the different problems result from consistent or inconsistent use of the same decision making strategy (Bar-Hillel, 1980; Byun & Ball, 2015; Cohen & Staub, 2015; McNair & Feeney, 2014, 2015).

The current study revealed that both uses are possible. The same decision making strategy can produce considerable variability in the answers provided by a participant, while conversely different decision making strategies can produce very little variability in the answers provided by a participant. For example, one participant gave an answer of 5% for the Sniffer Dog problem by calculating a value that fell between $P(H)$ and $P(D|H)$ (i.e., did not use base rate information for non-occurrence of H). However, the same participant used a different calculation strategy for the Spam E-mail problem [$P(D|\neg H)$] (i.e., only used individuating information for non-occurrence of H) and obtained an answer of similar magnitude (10%). This participant used another different calculation for the Pregnancy Test problem [$P(H)$] (only used base rate information for occurrence of H) and again provided an answer of similar magnitude (5%). The answers provided by participants are not a reliable source for researchers to use when distinguishing participants' decision making strategies. What researchers need to know is how participants actually came up with these answers.

The between- and within-participant variability in answers found in previous research has been explained by individual differences in numeracy and reasoning abilities (Lipkus, Samsa, & Rimer, 2001; Lipkus & Peters, 2009; Frederick 2005). Byun and Ball (2015) also found that measures of these abilities did correlate weakly with participant performance at solving early steps of the Bayesian probability calculations. However, one would expect these relationships to get stronger as the numeracy and reasoning required for latter steps increased, however Byun and Ball did not find this to be the case (refer to Table 1). Their findings supplement the results of the current study and suggest a numeracy-reasoning explanation is not comprehensive enough to explain the between- and within-participant variability found in Bayesian reasoning responses. In addition, this explanation does not explain why some participants are more likely to subjectively interpret the story or relate the story description to their own personal experiences rather than use the objective probability information provided for each problem.

A more comprehensive model for describing the different points where individual differences can result in complex Bayesian reasoning tasks is illustrated in Figure 13, and this model was developed from the verbal protocol analyses conducted in the current study. This model highlights the importance of distinguishing story interpretation effects from basic probability calculation mistakes. Consequently, this model goes beyond the simple numeracy-reasoning explanations provided by other researchers (Ajzen, 1977; Bar-Hillel, 1980; Edwards, 1968; Gigerenzer & Hoffrage, 1995; Hammerton, 1973; Kahneman & Tversky, 1972, 1973; Krynski & Tenenbaum, 2007; Lyon & Slovic, 1976). The model depicted in Figure 13 also highlights the difficulties in using a ‘one-method-fits-all’ prescriptive approach to help people perform better (more normative) Bayesian

reasoning and overcome the common base rate neglect bias (e.g., Gigerenzer & Hoffrage, 1995; Krynski & Tenenbaum, 2007; Siegrist & Keller, 2011; Tversky & Kahneman, 1977, 1980). The current research suggests that a combination of these methods will work best, and that the methods selected need to be tailored for each participant. The focus of past research on conclusions based only on participants' answers has resulted in researchers overlooking the different needs of different participants when performing complex Bayesian reasoning.

There were some participants in the current study who relied solely on their own subjective interpretations of the event descriptions that were independent of the probability information provided. They either deemed this information to be irrelevant to the problem or just simply ignored it (see also Bar-Hillel, 1980). The most frequent and distinguishable reason for neglecting probability information was a faulty understanding of the spatial and/or temporal relationships between the base rate event and the individuating event. This neglect hides the important causal relationship connecting these two events. Some participants viewed the base rate information as involving other people in different situations from the past, and therefore not relevant to what is happening now and described by the individuating information. For example, some participants stated that the fact other people have hidden drugs in their suitcase in the past is irrelevant to what their friend is going through now. But this is obviously illogical reasoning, because past statistics often predict future event outcomes. For example, intersections that have a history of accidents are probably dangerous intersections – why would you ignore the historical data when you are driving through this intersection? Other participants viewed the probability information as simply probability-based metaphors for the verbal

statements describing the event. Just like common probability metaphors in the English language like “one in a million” is another way of saying an event occurrence is rare or unlikely, and the statement “fifty fifty” suggests that two outcomes have roughly an equal probability of occurrence. One participant even relied on these subjective interpretations even though this participant showed in their last problem calculations that they possess the probability training to perform the normative Bayesian computations. One possible way of helping individuals to see the benefits of objective probability information is to provide them with indicators of the causal relationships between these sources of probability information and the final decision outcome. This was first suggested by Tversky and Kahneman (1977, 1980), and later shown by Krynski and Tenenbaum (2007) to help some participants make better Bayesian judgments. Although, this causal facilitation effect may only result for highly numerate participants (McNair & Feeney, 2015).

Other participants showed a difficulty translating verbal descriptions into probability formulae needed for each step of the probability calculations. Previous research suggests that providing the probability information in terms of natural frequencies can help participants overcome these difficulties. For example, saying one women out of 100 women can have cancer while 99 of these women don't have cancer, is easier for participants to understand than saying there is a one percent chance of having cancer in this population (Gigerenzer & Hoffrage, 1995). Researchers have also found performance benefits after they provide participants with a visual representation of the solution (e.g., tree-diagram) (Beitzel & Staley, 2015; Sirota, Kostovičová, & Vallée-Tourangeau, 2015). These researchers taught participants to perform Bayesian probability

calculations by providing them with these visual aids, and the researchers found these improvements to persist over time. Likewise, providing tools (pencil and paper and sets of cards matching the quantities stated in a Bayesian problem) to participants to help solve these problems can increase participant interactivity when solving Bayesian problems that results in improved performance (Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2015). Future research could involve a verbal protocol analysis of each participant's initial attempt to solve one Bayesian problem, and then tailor an intervention training strategy based on the model depicted in Figure 13 to help this participant. This identification-multi-method approach could be tested against the typical 'one-method-fits-all' approach offered in past research (Gigerenzer & Hoffrage, 1995; Krynski & Tenenbaum, 2007; Siegrist & Keller, 2011; Tversky & Kahneman, 1977, 1980).

The conclusions of the current research were based on retrospective verbal protocols and this methodology has been criticized in the past for not being completely objective when analyzing an individual's thought processes (Nisbett & Wilson, 1977). The participant may not be able to provide a verbal description of their thought processes, or alternatively, they may provide verbal descriptions that they feel match the demand characteristics of the experiment, especially when answering direct questions about the use of specific sources of probability information in the current study. A future study could utilize eye-movement recording to examine what problem information is being examined by the participant and how well this data corresponds with the verbal protocols collected. We would predict that when participants rely for the most part on the story description for their answers, their corresponding eye-movements will show very few fixations on numerical probability values. Likewise when participants stop their

calculation steps too early in the problem solving sequence, we would expect less eye-movement regressions to include prior probability calculations in these later steps.

Base rate neglect is a catchy term for a cognitive bias that can result during Bayesian reasoning, but it has limited the focus of past research to only examining why base rate neglect can occur. However, base rate neglect is simply one of many ways that participants can fail to conduct the normative steps required for solving Bayesian reasoning problems. The current thesis does not limit itself in this way, and this has resulted in a much more comprehensive model of the different ways that faulty Bayesian reasoning can occur. Our model also highlights that prescriptive procedures for overcoming these faults needs to be tailored to each individual and may require more than one method to resolve their difficulties.

Table 1

Bayesian computations, success rates, and correlations with numeracy and cognitive reasoning measures found by Byun and Ball (2015)

Calculation	Correct answer	% of Correct	# of Answer	Performance of Numeracy Scale	Performance of 2-item CRT
$P(H)$	15%	98.2%	2	.02	-.08
$P(\neg H)$	85%	98.2%	2	.02	-.08
$P(d H)$	80%	92.7%	5	.26**	.31**
$P(d \neg H)$	20%	92.7%	5	.17	.27***
$P(H) \times P(d H)$	12%	60.2%	23	.32*	.24*
$P(\neg H) \times P(d \neg H)$	17%	59.3%	20	.49***	.41***
$P(H) \times P(d H) + P(\neg H) \times P(d \neg H)$	29%	31.5%	23	.35***	.33**
$P(H d)$	41.38%	3.6%	25	.14	.16

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

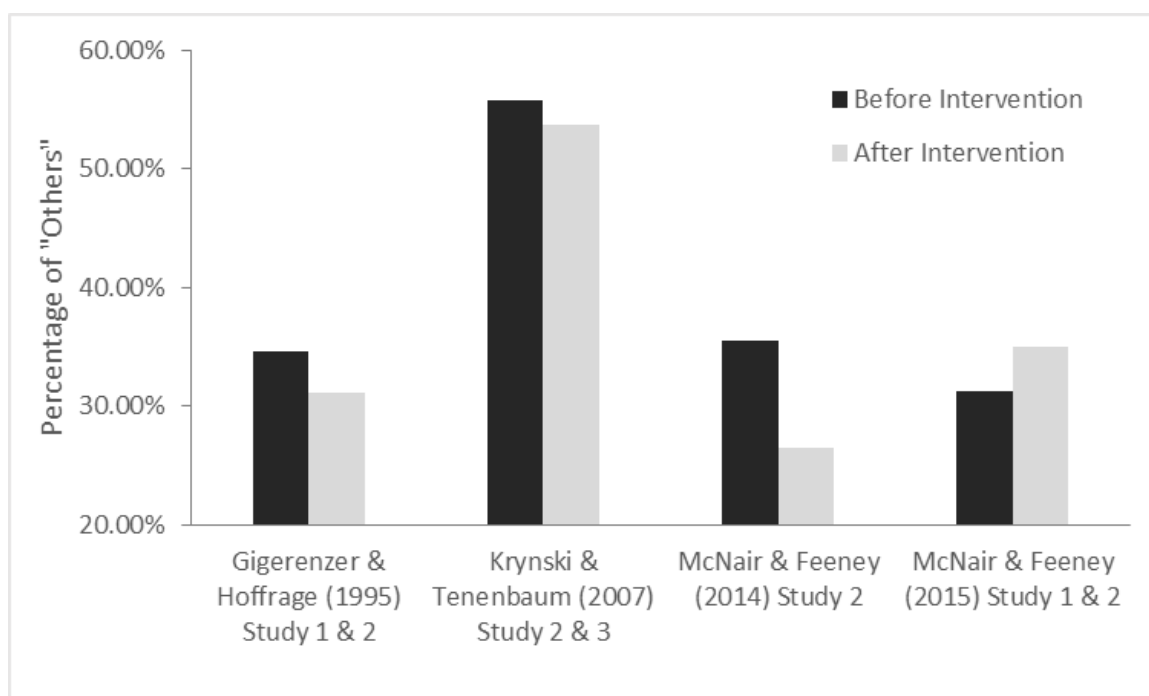


Figure 1. Total percentage of “other or non-explained” responses that resulted for four past studies that examined base rate neglect.

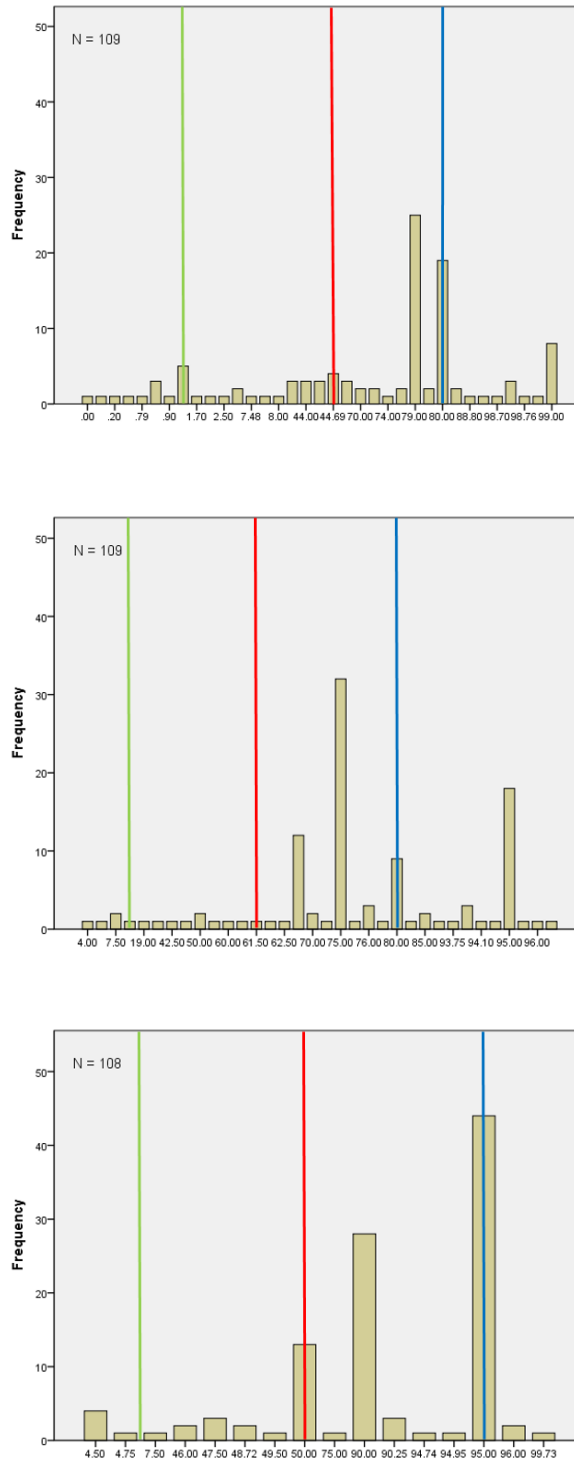


Figure 2. Frequency distribution of answers for a sample of three problems used by Byun and Ball (2015). Left line indicates base rate probability $P(H)$, middle line indicates correct Bayesian answer $P(H|D)$, and right line indicates hit rate probability $P(D|H)$.

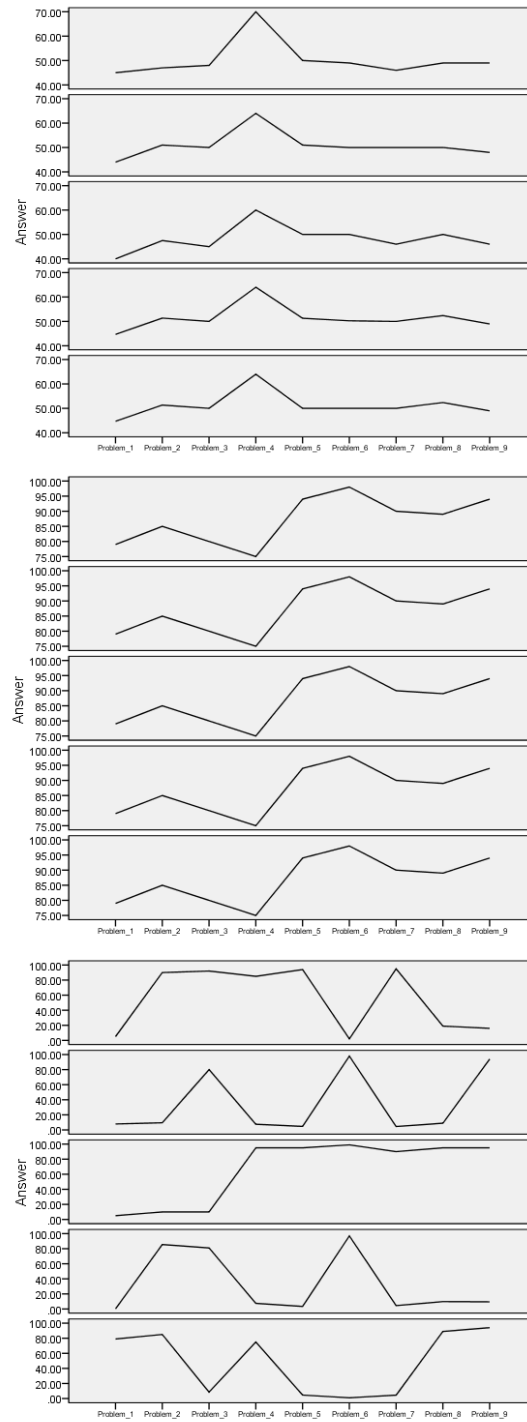


Figure 3. Examples of participant responses that highlight consistent probability (i.e., smallest inconsistency measure values) based responses (top and center panels) and inconsistent response patterns (bottom panel) (i.e., highest inconsistency measures).

A friend is traveling with you on a study abroad program. The Department of Customs and Excise reports that 2% of luggage at the airport you are travelling through contain drugs. Sniffer dogs at this airport can correctly detect drugs in 90% of cases when drugs are present. However, sniffer dogs will also incorrectly detect the presence of drugs in 20% of luggage at this airport that do not contain drugs.

The sniffer dogs at this airport detect drugs in your friend's suitcase. What is the probability your friend's suitcase actually contains drugs? What would you advise your friend to do next?



Your sexually active friend is 5 days late with her period. In 5% of cases when a sexually active woman's period is five days late, the woman will be pregnant. A recently developed pregnancy test offered for free at the college health center correctly identifies a pregnancy in 80% of women who are pregnant. However, the same test also incorrectly detects pregnancy in 10% of women who are not pregnant.

Your friend gets a positive result on this test. What is the probability that she is actually pregnant? What would you recommend she do next?



1% of emails are spam emails. The college's spam-filters successfully detect and block out 80% of spam emails. However, 10% of non-spam emails are mistakenly blocked as spam by these spam filters.

What percentage of your emails that are blocked by the college's spam filters are actually spam emails? You are waiting on important job application emails – should you be worried?

Figure 4. Instructions for each of the three Bayesian problems.

Participant: Um, this one was harder to think about because it kind of depends on the friend. Because I was just thinking about how to make it real life relatable, so I guess the 2% of luggage that contains drugs is a really small probability so that I wouldn't think that my friend had drugs, but if the dog smelled the drugs I would probably be a little skeptical, because dogs are right, like I mean sniffer dogs are right 90% of the time, which is really a lot. And, um, yeah, that's all. Well it's weird how they're right 90% of the time and not right 20% of the time, because that's more than 100 but, so I guess that makes it a little less reliable.

Experimenter: So, you said your answer was 70%. How did you come up with 70% as the answer?

Participant: I just subtracted 90 and 20 because those seemed like the most relevant numbers, 90% right and 20% wrong so a 70% chance. (Respondent #7)

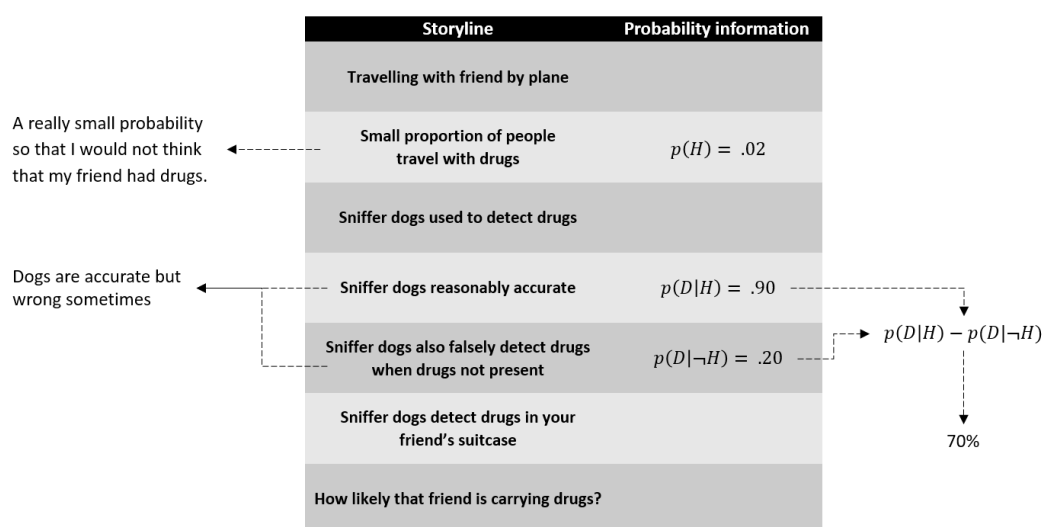


Figure 5. A sample transcript of verbal protocol. Transcripts are segmented (underlined) and then coded using the template provided below for each problem. Calculation processes are depicted on the right side of the coding template and non-calculation processes (e.g., subjective interpretation of value or situation) are depicted on the left side of the coding template.

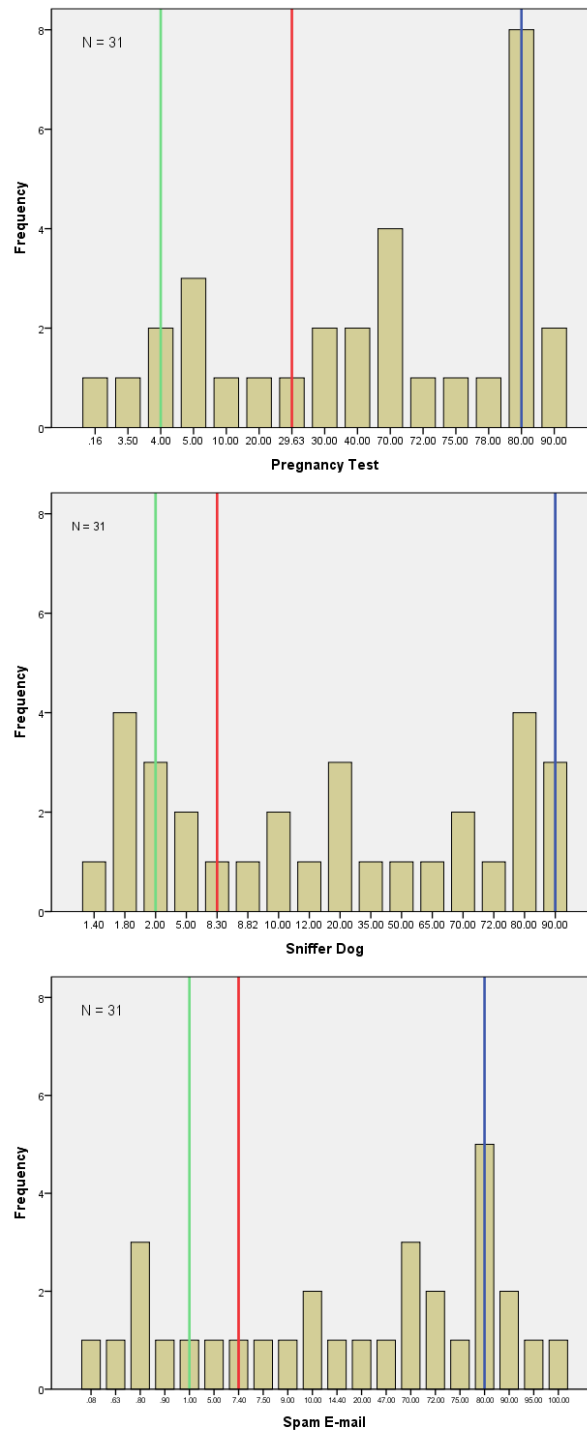


Figure 6. Frequency distribution of answers for three problems used in the current study. Left line indicates base rate probability $P(H)$, middle line indicates correct Bayesian answer $P(H|D)$, and right line indicates hit rate probability $P(D|H)$.

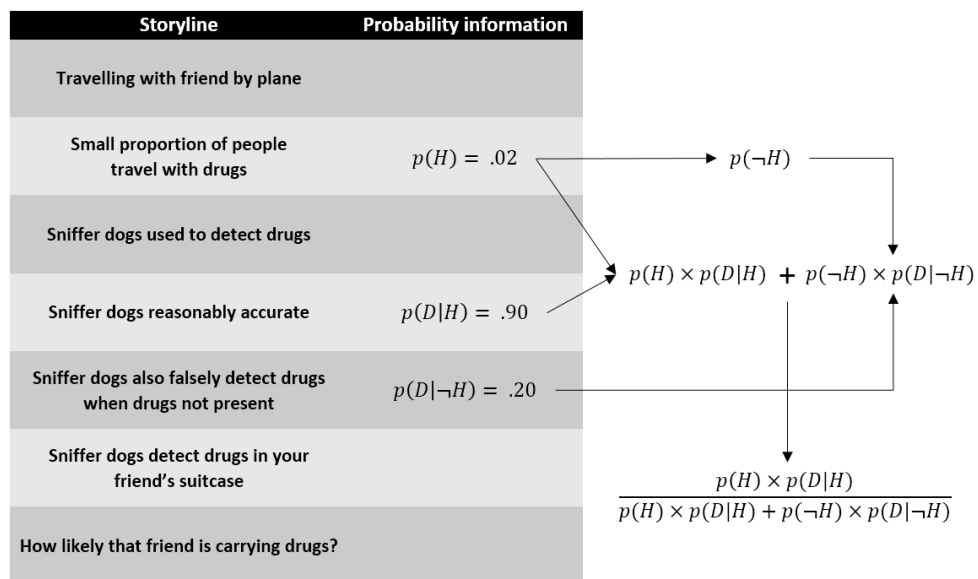


Figure 7. A coded example of a correct Bayesian reasoning strategy.

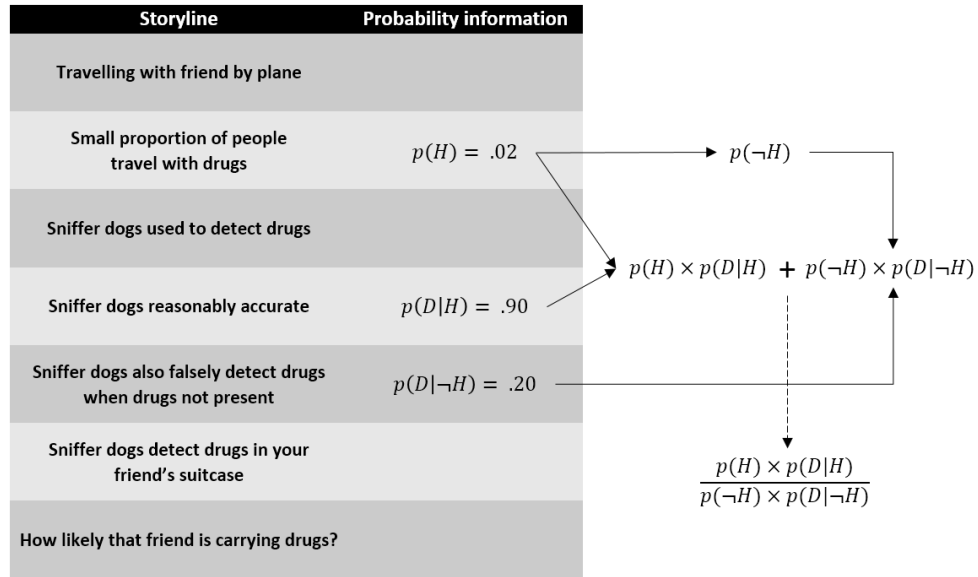


Figure 8. A coded example of an incorrect probability-based reasoning strategy where all the necessary information is utilized, but not correctly. This participant did not add $P(H) + P(H|D)$ to the denominator for the last step of the Bayesian probability calculation.

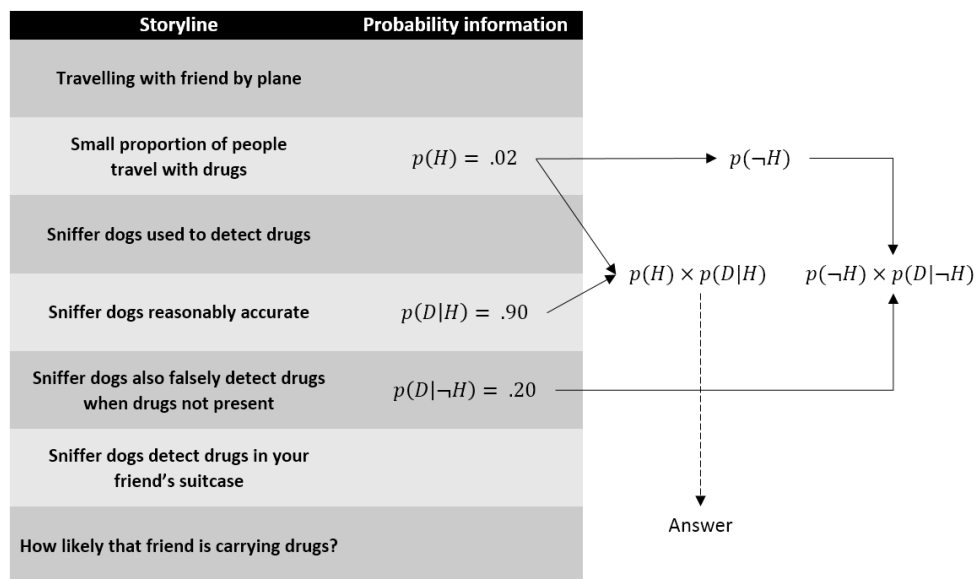


Figure 9. A coded example of a participant who reduced the four outcomes of the two combined events to just two outcomes. The participant stops at the conditional probability calculation as a result of this misperception.

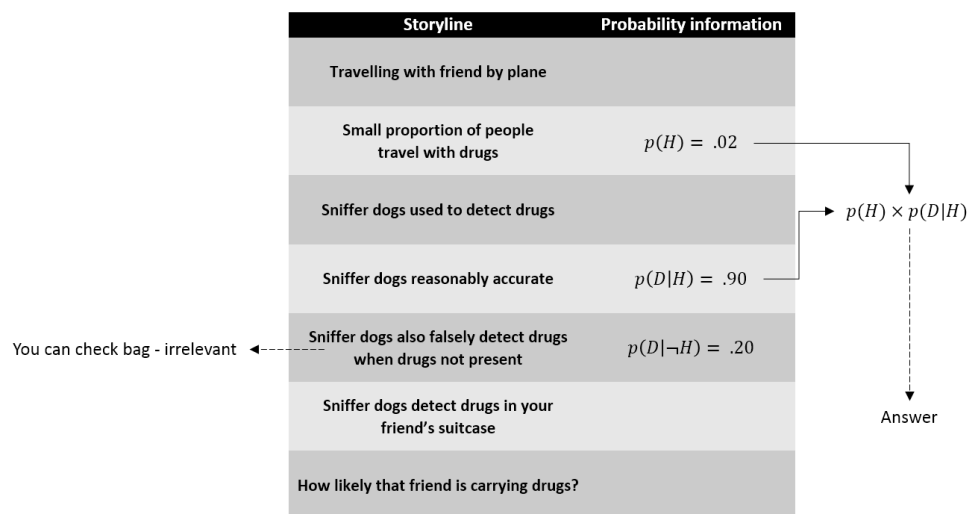


Figure 10. A coded example of a participant who uses some of the probability information, but not all of the information required to provide the correct answer.

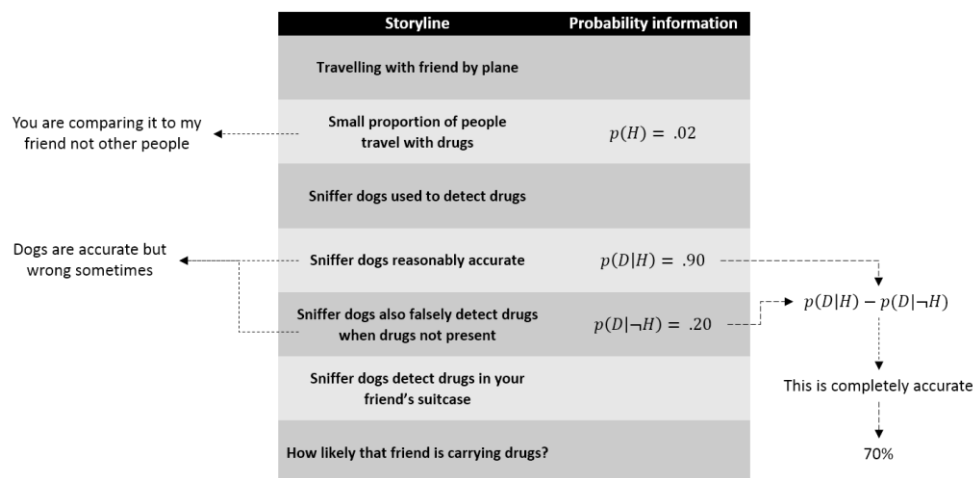


Figure 11. A coded example of a strategy involving two sources of information based on personal intuitive judgments.

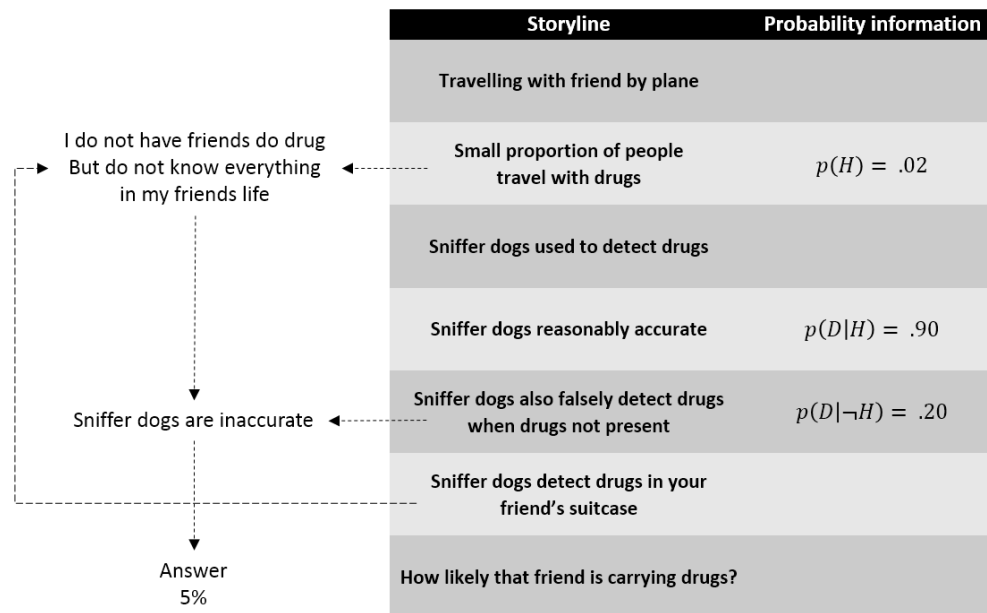


Figure 12. A coded example of the intuition-based reasoning approach. This participant made their final judgment after subjective interpretation of the story and reflecting on their own personal experiences.

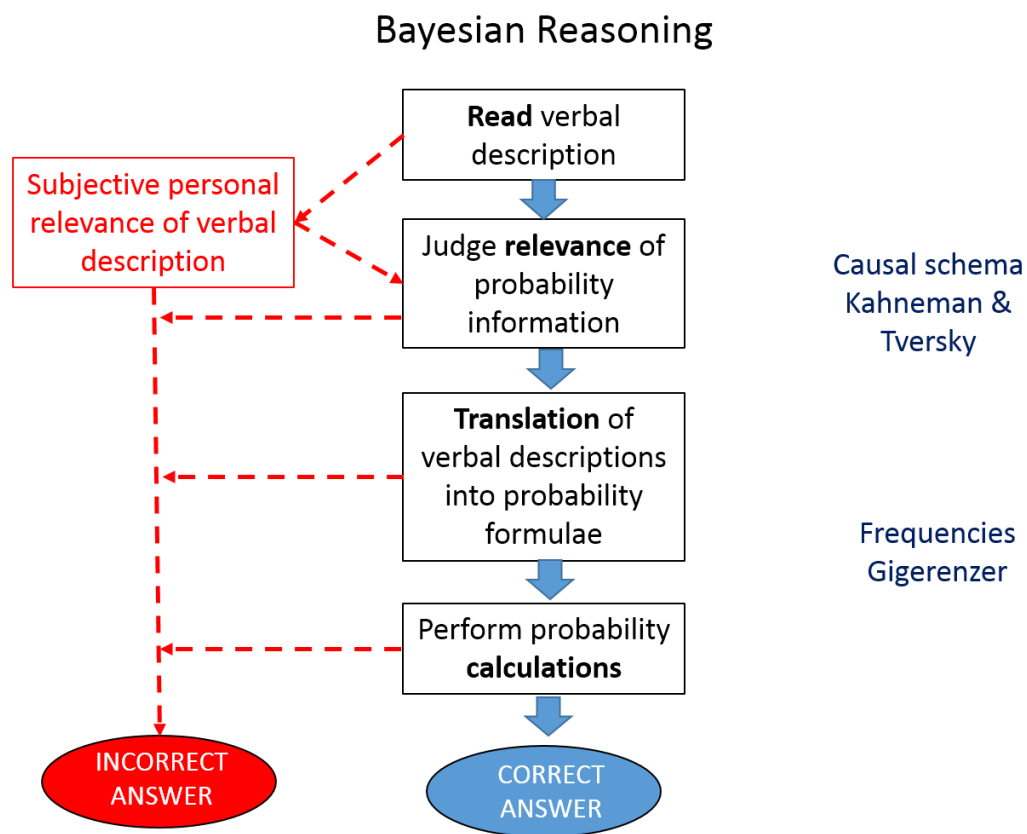


Figure 13. A descriptive model of how different Bayesian reasoning answers occur. Each stage contribute to the unique variability of answers and individuals need different help to facilitate their Bayesian decision making.

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.
<http://doi.org/10.1037/0022-3514.35.5.303>
- Ayal, S., & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision Making*, 9(3), 226.
- Ball, C. T., Langholtz, H. J., Auble, J., & Sopchak, B. (1998). Resource-Allocation Strategies: A Verbal Protocol Analysis. *Organizational Behavior and Human Decision Processes*, 76(1), 70–88. <http://doi.org/10.1006/obhd.1998.2798>
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233. [http://doi.org/10.1016/0001-6918\(80\)90046-3](http://doi.org/10.1016/0001-6918(80)90046-3)
- Beitzel, B. D., & Staley, R. K. (2015). The efficacy of using diagrams when solving probability word problems in college. *The Journal of Experimental Education*, 83(1), 130-145.
- Byun, K., & Ball, C. T. (2015, November). *Lost in Translation: There is more to Base Rate Neglect than Neglect*. Poster session presented at the Annual Meeting of the Society for Judgment and Decision Making, Chicago, IL.
- Cohen, A. L., & Staub, A. (2015). Within-subject consistency and between-subject variability in Bayesian reasoning strategies. *Cognitive Psychology*, 81, 26–47.
<http://doi.org/10.1016/j.cogpsych.2015.08.001>
- Edwards, W. (1968). *Conservatism in human information processing*. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. New York: Wiley, 1968, 17-52.

- Frederick, S. (2005). Cognitive Reflection and Decision Making. *The Journal of Economic Perspectives*, 19(4), 25–42.
- Gigerenzer, G. (1996). The Psychology of Good Judgment Frequency Formats and Simple Algorithms. *Medical Decision Making*, 16(3), 273–280.
<http://doi.org/10.1177/0272989X9601600312>
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704.
<http://doi.org/10.1037/0033-295X.102.4.684>
- Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*, 101(2), 252–254. <http://doi.org/10.1037/h0035224>
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454.
[http://doi.org/10.1016/0010-0285\(72\)90016-3](http://doi.org/10.1016/0010-0285(72)90016-3)
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. <http://doi.org/10.1037/h0034747>
- Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3), 430–450.
<http://doi.org/10.1037/0096-3445.136.3.430>
- Lipkus, I. M., & Peters, E. (2009). Understanding the Role of Numeracy in Health: Proposed Theoretical Framework and Practical Insights. *Health Education & Behavior*, 36(6), 1065–1081. <http://doi.org/10.1177/1090198109341533>
- Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21(1), 37–44.

- Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychological Review*, 106(2), 411-416.
- Lyon, D., & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40(4), 287–298.
[http://doi.org/10.1016/0001-6918\(76\)90032-9](http://doi.org/10.1016/0001-6918(76)90032-9)
- McNair, S., & Feeney, A. (2014). When does information about causal structure improve statistical reasoning? *The Quarterly Journal of Experimental Psychology*, 67(4), 625–645. <http://doi.org/10.1080/17470218.2013.821709>
- McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, 22(1), 258-264.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52(3), 194–216. <http://doi.org/10.1037/h0048070>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.
<http://doi.org/10.1037/0033-295X.84.3.231>
- Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, 14(6), 1147-1152.
- Siegrist, M., & Keller, C. (2011). Natural frequencies and Bayesian reasoning: the impact of formal education and problem context. *Journal of Risk Research*, 14(9), 1039-1055.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63(2), 129-138.

- Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). How to train your Bayesian: a problem-representation transfer rather than a format-representation shift explains training effects. *The Quarterly Journal of Experimental Psychology*, 68(1), 1-9.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7), 1275–1289.
- Tversky, A., & Kahneman, D. (1977). *Causal Schemata in Judgments under Uncertainty*. Decisions and Designs, McLean, Virginia.
- Tversky, A., & Kahneman, D. (1980). Causal schemas in judgment under uncertainty. In M. Fishbein (ed.), *Progress in social psychology*. Hillsdale, N.J.: Erlbaum.
- Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (eds), *Judgment under uncertainty: Heuristics and biases* (pp. 153-163). Cambridge: Cambridge University Press.
- Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology: General*, 144(3), 581-603.