IMPROVING USE OF STATISTICAL INFORMATION BY JURORS BY REDUCING CONFUSION OF THE INVERSE

by

JOHN DAVID RAACKE

B.A., Christian Brothers University, 2000
M.S., Kansas State University, 2003

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the

requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2005

ABSTRACT

In many situations, people are called on to make judgments about the likelihood of an event. Research has shown that when people make these judgments, they frequently equate or confuse conditional probabilities with other conditional probabilities. This equating or confusing of conditional probabilities is known as the *confusion of the inverse.* Research investigating this problem typically focuses on clinical and medical decision-making and the use of statistical evidence to make diagnoses. However, one area in which the *confusion of the inverse* has not been studied is in juror decision-making. Thus, the purpose of this dissertation was to (1) determine if the *confusion of the inverse* influences juror decision-making, (2) interpret reasons why this confusion occurs, and (3) attempt to eliminate it from juror decision-making.

Jurors were presented with four court cases gathered from local and federal courthouses in a small Mid-western city. In each of the four cases, a single piece of evidence was presented (statistical only) which was to be used when rendering verdicts. Finally, each case contained juror instructions for the specific case type: murder, kidnapping, arson, sexual assault.

Overall, jurors fell prey to the *confusion of the inverse*, equating the probability of the data given the hypothesis [P(D|H)] with the probability of the hypothesis given the data [P(H|D)]. However, the research was unable to reduce the effect, much less eliminate it from the task. Interestingly, jurors tended to ignore the statistical evidence (i.e., estimations about probability of a match) in favor of their own personal believe in the strength of the evidence.

Although the original intent of reducing/eliminating the *confusion of the inverse* was not accomplished, the dissertation did accomplish three things. First, researchers have hypothesized three reasons why people engage in incorrect probabilistic reasoning, and the dissertation affirmed that it is indeed a function of the confusion of conditional probabilities – the *confusion of the inverse*. Second, it seems that the use of statistical evidence in a trial is ignored by most jurors in favor of their own personal belief in the evidence's strength. Finally, the criteria needed for "beyond a reasonable doubt" may be too stringent.

IMPROVING USE OF STATISTICAL INFORMATION BY JURORS BY REDUCING CONFUSION OF THE INVERSE

by

JOHN DAVID RAACKE

B.A., Christian Brothers University, 2000
M.S., Kansas State University, 2003

A DISSERTATION

submitted in partial fulfillment of the

requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2005

Approved by

Major Professor
James Shanteau

ABSTRACT

In many situations, people are called on to make judgments about the likelihood of an event. Research has shown that when people make these judgments, they frequently equate or confuse conditional probabilities with other conditional probabilities. This equating or confusing of conditional probabilities is known as the *confusion of the inverse.* Research investigating this problem typically focuses on clinical and medical decision-making and the use of statistical evidence to make diagnoses. However, one area in which the *confusion of the inverse* has not been studied is in juror decision-making. Thus, the purpose of this dissertation was to (1) determine if the *confusion of the inverse* influences juror decision-making, (2) interpret reasons why this confusion occurs, and (3) attempt to eliminate it from juror decision-making.

Jurors were presented with four court cases gathered from local and federal courthouses in a small Mid-western city. In each of the four cases, a single piece of evidence was presented (statistical only) which was to be used when rendering verdicts. Finally, each case contained juror instructions for the specific case type: murder, kidnapping, arson, sexual assault.

Overall, jurors fell prey to the *confusion of the inverse*, equating the probability of the data given the hypothesis [P(D|H)] with the probability of the hypothesis given the data [P(H|D)]. However, the research was unable to reduce the effect, much less eliminate it from the task. Interestingly, jurors tended to ignore the statistical evidence (i.e., estimations about probability of a match) in favor of their own personal believe in the strength of the evidence.

Although the original intent of reducing/eliminating the *confusion of the inverse* was not accomplished, the dissertation did accomplish three things. First, researchers have hypothesized three reasons why people engage in incorrect probabilistic reasoning, and the dissertation affirmed that it is indeed a function of the confusion of conditional probabilities – the *confusion of the inverse*. Second, it seems that the use of statistical evidence in a trial is ignored by most jurors in favor of their own personal belief in the evidence's strength. Finally, the criteria needed for "beyond a reasonable doubt" may be too stringent.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

As I sit here and write this final acknowledgement to all those who helped me along the way, I can’t help but think of a particular professor who said I wasn’t smart enough to work with him and another student was (she has since dropped out)…wow, kind of awkward!

Anyway, I would first like to thank the members of my committee, starting with Dr. Ki-Joon Back for being my outside chair on my committee. Dr. Back was a great person to work with and very flexible with my schedule since I now live in Sioux City, IA. I would also like to thank Dr. Fullagar and Dr. Harris each for their help and guidance over the last 4+ years. Each of you have made things easier and at time difficult, but have always had my best interests at heart, thanks!

I would also like to thank my major professor, Dr. James Shanteau. It is not too often that a person comes along in this world and offers out a hand of support. I came into this program under a backdoor loophole (that many would like to close). Dr. Shanteau told me that if I worked with him and did well in my classes that we would see if I could stay in the program full time. Needless to say I did not to that well in my classes, but the man still went to bat for me and gave me a chance to move from my status as Non-Degree Special student to a Full-time Graduate Student. I will never be able to thank him enough for all of the stuff that he has shown me (grant writing, the wonderful world of Los Angeles) or the opportunities he has afforded me (going to Disney World in ’01 for a week, introducing me to a great friend—Dr. Rickey, and helping me to succeed. Dr. Shanteau, you have reluctantly at times acted like a father figure to me over the past several years, always lending your ear when I needed advice about what GRA to take after your grant was up and what jobs to take when it was time, for that I am truly thankful.

Finally, I would like to thank my family, Jennifer and Callie. This last year on us has been hard, new city, new house, new jobs, and me…completing my dissertation. I know at times it has been hard, but without the two of you in my life, I would not be here today. Both of you have supported me in whatever way possible by pushing me to go to graduate school, pushing me to succeed, by being therefore me through it all to just making me laugh at the end of the day by saying “Daddy, shake your booty!” I love you both very much, without you there could be no tomorrow!

DEDICATION

I would like to dedicate this dissertation to my family – Jennifer my adoring and wonderful wife & Callie my beautiful energetic child. Who knows what else may come…wink, wink!

Often times, people are asked to make judgments regarding the likelihood that an event will occur given imperfect evidence about the event. These judgments are typically made with people having some prior knowledge that is updated with new information. Frequently, these same judgments will involve some level of uncertainty.

For instance, imagine that you are on a jury. As part of the defense's case, you are to hear testimony from a notable expert on forensic evidence. The expert witness plans to show that there is less than a 5% chance that the defendant committed the crime. However, under cross examination, the prosecutor makes it clear that the expert witness is correct about 75% of the time when predicting someone does not match evidence and correct about 95% of the time when predicting someone does match evidence. As a juror, with this information what is the probability the defendant matches the evidence?

The judgment as to the likelihood of a match is known as a *posterior probability judgment*. To estimate, this posterior probability judgment involves assessing the *likelihood of an event* by updating a *prior probability judgment* with new evidence. The most commonly used normative model for calculating posterior probabilities is Bayes Theorem[1]. Bayes Theorem was named after Thomas Bayes (1702[?] – 1761), an 18th century English minister, who wrote a simple essay about conditional probabilities (Edwards, 1954). Bayes Theorem states that *P(H|D)*, the posterior probability (*P*) that hypothesis (H) is true given datum (D), can be calculated as:

[1] Bayes Theorem is not the only model used with these problems. It has also been proposed to treat this problem as a Signal Detection Issue where there are hits, misses, false alarms, and correct rejections. And, subject's sensitivity or criteria is used to make judgments about information as is done by Wason and others (1983).

$$P(\mathrm{H}|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{H})*P(\mathrm{H})}{P(\mathrm{D}|\mathrm{H})*P(\mathrm{H}) + P(\mathrm{D}|{\sim}\mathrm{H})*P({\sim}\mathrm{H})}$$

where,

> **_P_(H|D)** is the probability that the defendant does not match the evidence, given that the expert says he won't (posterior probability),
>
> **_P_(D|H)** is the probability that , if it the defendant does not match, the expert will correctly forecast it (sensitivity or true-positive rate),
>
> **_P_(H)** is the initial probability that the expert believes the defendant does not match the evidence (prior probability or baserate),
>
> **_P_(~H)** is the prior probability that it is the defendant [1 – *P*(H)],
>
> **_P_(D|~H)** is the probability that, if the evidence does match, the expert will incorrectly forecast the evidence as not matching (false-positive rate).

Thus, Bayes Theorem states that the probability of the hypothesis given the data is equal to the probability of the data times the initial probability, divided by the probability of the data given the hypothesis times the initial probability multiplied by the prior probability times the false positive rate.

Using Bayes Theorem, the correct answer to the question posed above can be calculated as follows. Using 5% as the expert's estimate of the prior probability (or baserate) that it was not the defendant and taking into account the new information provided by the prosecutor about the expert, we obtain:

$$P(\mathrm{H}|\mathrm{D}) = \frac{(.75)(.05)}{(.75)(.05) + (.05)(.95)} = .44$$

Thus as a juror, you should estimate there is approximately a 44% chance that the defendant does not match the evidence.

Incorrect Probabilistic Reasoning

Most participants when presented with similar problems do not follow Bayes Theorem and arrive at an incorrect judgment. Specifically research has shown that people often tend to equate (or confuse) the posterior probability, *P(H|D)*, with the sensitivity or true-positive rate, *P(D|H)* (Eddy, 1982; Dawes, 1986; Dawes, Mirels, Gold, & Donahue, 1993; Villejoubert & Mandel, 2002). Thus, most people when asked for *P(H|D)* incorrectly give answers similar to the *P(D|H)*. In the above example, that means people would say the chance of rain was around 75% (= the sensitivity = *P(D|H)*).

Researchers have proposed three reasons why people when presented with probabilistic reasoning problems have such difficulties arriving at correct judgments (Hamm, 1993). The three reasons are (1) the subjects consider the baserate (prior probability) irrelevant, (2) subjects perform an inappropriate integration of the baserate and case information/evidence, and (3) subjects confuse conditional probabilities with each other. Each of these reasons will be considered in turn.

Baserate Neglect. The first hypothesis states that people neglect baserate information relative to the amount of attention that should be given to them (Cohen, 1981; Niniluotto, 1981). In other words, people do not take into account the prior probability judgment when faced with a probability problem.

To illustrate this point, Niniluotto (1981), used the well-known Blue/Green cab problem from Kahneman and Tversky (1974). In the Blue/Green cab problem, subjects are presented with a multi-paragraph description of an automobile accident. After each of four paragraphs, they are asked to make a judgment about the probability that either a Blue or a Green Cab was involved in the accident. The first paragraph contains the introduction, the second contains the baserate information [*P*(H)], the third paragraph contains the evidence[*P*(D/H)], and the fourth contains the reliability of the evidence [*P*(D/~H)]. Thus, the problem reads as:

> [Introduction.] The next word problem is about two taxi cab companies. A cab from one of the companies was involved in a hit and run accident at night. It is hard to know which company it was from. You will be asked to estimate how likely it is that the cab involved in the accident belonged to each of the two cab companies.
>
> In this city there are only two cab companies, the Blue Cab Company and the Green Cab Company.
>
> With what you know now, what is the probability that the cab involved in the hit and run accident was from the Blue Cab Company? ____________.
>
> [Evidence paragraph] There was only one witness to the hit and run accident. The witness identified the cab as blue.
>
> With what you know now, what is the probability that it was a Blue Cab?______.
>
> [Baserate paragraph] The Green Cab Company is larger, with 85% of the cabs in the city.
>
> With what you know now, what do you think is the probability that a cab from the Blue Cab Company was the one involved in the accident? ________.
>
> [Reliability paragraph.] The police were concerned about the accuracy of the witness who saw the accident. They tested the witness's reliability under the same circumstances that existed on the night of the accident and concluded that the witness could correctly identify cabs of each one of the two colors 80% of the time and misidentified them 20% of the time.

With what you know now, what is the probability that the cab was a Blue Cab?__________.

Using Bayes Theorem,

$$P(\mathrm{H}|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{H})*P(\mathrm{H})}{P(\mathrm{D}|\mathrm{H})*P(\mathrm{H}) + P(\mathrm{D}|{\sim}\mathrm{H})*P({\sim}\mathrm{H})}$$

where,

***P*(H|D)** is the probability that the cab identified is a Blue Cab (posterior probability),

***P*(D|H)** is the probability that , if it they see a Blue cab, the witness will correctly identify it as Blue (= 80%),

***P*(H)** is the initial probability that it is a Blue Cab, baserate (= 15% of cabs are Blue),

***P*(~H)** is the prior probability, the amount of time the witness is incorrect in judgments about the cabs (= 1 – *P*(H) = 85%),

***P*(D|~H)** is the probability that, if it is not the Blue Cab, the witness will incorrectly identify it as a Blue Cab (= 20%),

$$P(\mathrm{H}|\mathrm{D}) = \frac{(.80)(.15)}{(.80)(.15) + (.20)(.85)} = .41$$

Thus, the correct answer to this problem is .41. However, Niniluotto (1981) believed that the subjects were misinterpreting the relative frequency of Blue and Green cabs in the city as useless (baserate neglect), leading them to use the reliability as their answer. Their reasoning would go as follows:

1. ignore baserate information as being irrelevant to the prior probability [P(H)]
2. use a prior probability of P(H) = .50 [2 cab companies in the city]
3. Bayes Theorem, yielding P(H/D) = P(D/H)

$$P(H|D) = \frac{P(D|H)*P(H)}{P(D|H)*P(H) + P(D|{\sim}H)*P({\sim}H)}$$

$$= \frac{P(D/H) * .50}{P(D/H)*.50 + (1-P(D/H))*.50}$$

$$= P(D/H)$$

Therefore, using this reasoning, people would arrive at *P(D|H)* as their answer for *P(H|D)*, which is .80 in the Blue/Green Cab problem (Hamm, 1993).

Tversky (1981) provided evidence against this hypothesis, showing that when subjects are given a problem where only the baserate is known, they use it as their response (Kahneman & Tversky, 1972; Lyon & Slovic, 1976; Hamm, 1993). Thus, this reason has been shown to be incomplete in accounting for inaccurate probabilistic inferences.

Mis-Integration of Baserate and Case Information/Evidence. The second reason used to account for why people arrive at incorrect answers an incorrect integration rule.

It has been proposed that people inappropriately combine the baserate and case-specific information when making their judgment (Bar-Hillel, 1980; Tversky & Kanheman, 1982). Tversky and Kanheman (1982) argued that people integrate these two kinds of statistical information such that their answer is in between the two numbers. However, when the two kinds of statistical information are combined, the subject's answer tends to be closer to the case information/evidence than to Bayes Theorem (Fischoff & Bar-Hillel, 1984; Hamm, 1987). Yet, when the baserate information is made more relevant (i.e., it is highlighted in some way), the subject's answers tend to shift in the direction of the baserate (Bar-Hillel, 1980). This hypothesis can account for why people respond to probabilistic problems using the sensitivity, $P(D|H)$. By focusing all attention on the baserate, the research is ignoring the equation of $P(H|D)$ with $P(D|H)$. Unfortunately, this reason does not support the data as well as the third hypothesis.

Confusion of Conditional Probabilities. The final reason is that people have difficulty distinguishing between conditional probabilities. Specifically, people tend to confuse the conditional probability $P(H|D)$ with the conditional probability $P(D|H)$. This confusion of the two apparently occurs because people are unable to understand that the two concepts are different. This hypothesis is known as the *confusion of the inverse* and will be at the focal point of the current study.

Confusion of the Inverse

The term *confusion of the inverse* has also been labeled the *conversion error* (Wolfe, 1995), the *confusion hypothesis* (Macchi,1995), the *Fisherian Algorithm*

(Gigerenzer & Hoffrage, 1995), and the *inverse fallacy* (Koehler, 1996). These terms have been used by researchers to describe people's inability to correctly interpret probabilistic reasoning problems. When presented with a probabilistic judgment people tend to confuse (equate) a conditional probability with its inverse probability. This phenomenon was first described by Meehl and Rosen (1955), "… who contrasted their informal observations of their colleagues' psychiatric diagnoses with normative diagnostic procedures based on the relationship between inverse probabilities inferred from Bayes' Theorem" (Dawes, Mirels, Gold, & Donahue, 1993, p. 396). Thus, these colleagues tended to equate the probability of a hypothesis given some data with the probability of some data given a hypothesis, i.e., the inverse.

One of the first examples in the literature was cited by Eddy (1982). He presented the following scenario to over 100 physicians specializing in cancer (Utts, 2003):

> One of your patients has a lump in her breast. You are almost certain that it is benign, in fact you would say there is only a 1% chance that it is malignant. But just to be sure, you have the patient undergo a mammogram, a breast X-ray designed to detect cancer.
>
> You know from the medical literature that mammograms are 80% accurate for malignant lumps and 90% accurate for benign lumps. In other words, if the lump is truly malignant 80% of the time and will falsely say it is benign 20% of the time. If the lump is benign, the test results will say so 90% of the time and will falsely declare that it is malignant only 10% of the time.
>
> Sadly, the mammogram for your patient is returned with the news that the lump is malignant. What are the chances that it is truly malignant?

Results of Eddy's study showed that close to 95% of physicians responded with an answer near 75%. However, using Bayes theorem, the correct answer is much smaller. Using 1% as the physician's estimate of the prior probability that the mass is malignant and taking into account the new information, we obtain:

$$P(\mathrm{H}|\mathrm{D}) = \frac{(.80)(.01)}{(.80)(.01) + (.10)(.99)} = .075$$

The physicians should have estimated approximately a 7.5% chance that the patient actually has cancer. Eddy (1982) recounted: “When asked about this, the erring physicians usually reported that they assumed that the probability of cancer given that the patient has a positive X-ray was approximately equal to the probability of a positive X-ray in a patient with cancer” (p.254). In other words, the physicians confused the probability of a positive test given cancer [*P(D|H)*] with the probability of cancer given a positive test [*P(H|D)*].

Following Eddy’s (1982) work on *confusion of the inverse*, Dawes has written several papers on this problem. His most recent work has centered on the individuals’ beliefs inherent in their implicit personality theory. Dawes, Mirels, Gold and Donahue (1993) showed that subject’s failure to distinguish between inverse conditional probabilities extended to the kinds of generalized beliefs or propositions that constitute an individual’s implicit personality theory.

For example, in a study released by the Automobile Association Foundation for Traffic Safety (Stutts et al., 2001) researchers found that 1.5% of drivers in accidents reported they were using a cell phone, whereas 10.9% of drivers in accidents reported they were talking to a fellow passenger. Most of the national media quickly made the assumption that talking on a cell phone was much less likely to cause an accident than other distractions, like talking to another in the car or tuning your stereo. This result was touted by the media for several weeks.

The problem with the national media's conclusion is they are confusing two conditional probabilities. The researchers reported a portion of accidents (1.5% or .015) for which the driver was using a cell phone. This proportion represents the probability the driver was using a cell phone *given* that they were in an accident [$P(H|D)$]. But, the value of interest is the probability of an accident *given* the driver was using a cell phone [$P(D|H)$], the inverse. This probability can not be calculated since the researchers did not report the prevalence of cell phone use. Yet, it is likely that more drivers are talking with other passengers than on cell phones at any given time. Thus, the probability of an accident given the driver was using a cell phone is quite possibly higher than when a driver is talking to a fellow passenger.

Studying the Confusion of the Inverse

Much of the work done by researchers on the *confusion of the inverse* has been focused on medical judgments. Specifically, this research has centered on clinical psychologists, physicians and medical personnel interpreting statistical information and communicating that information to their patients.

A review of the medical literature indicates that the results reported by Eddy (1982) are not unusual. For instance, there is a strong tendency in the medical profession to equate the predictive accuracy of a positive report of cancer with the retrospective accuracy of a mammography, i.e., to equate the two conditional probabilities $P(H|D)$ with

P(D|H) (Lusted, 1968; Casscells, Schoenberger, & Grayboys, 1978). For instance, Eddy (1982, p.254) reports:

> …a 1964 article in *Radiology* stated ‘ the total correctness of the X-ray diagnosis was 674 out of 759 or 89%’ (p.254). A contributor to *Obstetrics and Gynecology* in 1966 said, ‘Asch found a 90 percent correlation of mammography with the pathologic findings in 500 patients’ (p.217). ‘The agreement in radiologic and pathologic diagnosis was 91.6 percent’ (Egan, 1972, p.379).

In a variety of medical situations, he reported statements indicate that if a patient has a positive test, they will have cancer 90% of the time, which is not correct.

Dawes (1986) studied this problem by examining clinicians’ judgments. One of his most striking examples revolves around a newspaper article about a clinician named Charles Rogers, M.D.

> Charles S. Rogers, M.D., is removing “high risk” breasts before cancer has developed. The risk factor is determined by mammogram “patterns” of mild ducts and lobules, which show that just over half of the women in the highest risk group are likely to develop cancer between the ages of 40 and 59. The mammogram patterns are the work of Detroit radiologist John N. Wolfe, M.D. He [Dr. Rogers] has performed the surgical procedure on 90 women in 2 years. (McGee, 1979).

The procedural rationale is found in Dr. Rogers’ interpretation of studies done by Wolfe.

> In his research, Wolfe found that 1 in 13 women in the general population will develop breast cancer, but that 1 in 2 or 3 BY (highest risk) women will develop it between the ages of 40 and 59. The lowest risk women (NI) account for 42% of the population, but only 7.5% of the carcinomas. By examining the DY women and those in the next lower risk groups, Pl and P2, Wolfe felt that 93% of the breast cancers could be found in 57 % of the population. (McGee, 1979).

Thus, Rogers concluded that roughly “1 in 2 or 3” women in the high risk group will develop cancer, thereby justifying an operation. However, applying Bayes Theorem to this problem yields an entirely different answer. Using Rogers' number taken from Wolfe, Bayes Theorem states:

$$P(\text{H}|\text{D}) = \frac{P(\text{D}|\text{H})*P(\text{H})}{P(\text{D}|\text{H})*P(\text{H}) + P(\text{D}|\sim\text{H})*P(\sim\text{H})}$$

where,

***P*(H|D)** is the probability that there will be cancer, given a woman is in the high risk group,

***P*(D|H)** is the probability of being in the high risk group given they have cancer **(= .93)**,

***P*(H)** is the probability of cancer **(1 out of 13 = .077)**

***P*(~H)** is the prior probability of no cancer **(1 – *P*(H) = .923)**

***P*(D|~H)** is the probability that, if it is not cancer, the clinician will incorrectly diagnose it as cancer **(= .57)**.

The correct answer to the question posed in the above example can be calculated:

$$P(\text{H}|\text{D}) = \frac{(.93)(.077)}{(.93)(.077) + (.57)(.923)} = .12$$

Thus, the probability that someone will have cancer given they are in the high risk group is approximately 12%, which is nowhere close to Rogers' estimate of "1 in 2 or 3"[2]. Hence, the operations that Rogers was performing in his patients, "90 in 2 years" was unnecessary and reckless due to his misinterpretation of conditional probabilities (Dawes,

[2] 1 in 2 is closer to the inverse probability

1986). The medical literature is replete with examples in which incorrect judgment about probabilistic reasoning problems involve the *confusion of the inverse* and have dire effects on people. The question is why does this confusion happen?

Explaining the Confusion of the Inverse

There are three explanations that have been postulated to account for why people confuse conditional probabilities and fall prey to the *confusion of the inverse.* They are (1) the sample-space framework, (2) the formulation of diagnostic information, and (3) the use of point probabilities as opposed to frequencies.

Sample-Space Framework. The first explanation proposed to account for the *confusion of the inverse* is known as the *sample-space framework* (Gavanski & Hui, 1992; Hanita, Gavanski & Fazio, 1997; Sherman, McMullen, & Gavanski, 1992). This framework proposes that the *confusion of the inverse* is linked to the way in which people partition their memory. Specifically, people access sets of information from memory when assessing probabilities. So, when asked to assess *P(H|D)*, people should base their judgments on the sample space as defined by the feature set D. It has been hypothesized that this sample space of D is a feature and that people are more likely to partition their memories based on categories such as H. Thus, when faced with assessing probabilities people replace an unnatural partition (D) with a more accessible one, particularly a sample space of a category (H) (Villejoubert & Mandel, 2002).

This process by which people replace the originally asked for feature (D) with a more accessible category (H) results in the *confusion of the inverse*. However, this explanation is difficult to test experimentally since research must rely on hypothetical sample spaces of categories and features that an individual may or may not have. Thus, testing the validity of this explanation is not easily accomplished.

Formulation of Diagnostic Information. The second proposed explanation of the *confusion of the inverse* was proposed by Macchi (1995, 2000). She argued that the formulation of diagnostic information is the key factor when a person interprets statistical data. Consider the following formulations of the same statement:

> (1) The percentage of elements presenting the feature D is three times higher among H elements than among ~H elements.
> (2) In the group of elements presenting the feature D, the percentage of H elements is three times higher than the percentage of ~H elements.
> (3) The feature D is presented in x% of H elements, the feature D is presented in y% of ~H elements, and x is three times higher than y (Villejoubert & Mandel, 2002).

Macchi (1995) argued that a formulation of diagnostic information like #(1) would be interpreted as #(2) by people, as opposed to what it logically is supposed to imply, #(3). It is this misinterpretation by people that leads to the *confusion of the inverse*. Simply put, people have trouble interpreting information. More often than not, the formulation of information does not match the way people interpret it, thus producing confusion.

To explain the confusion, Macchi (1995) argues it is due to a lack of clarity of the independence of baserate *P*(H) with *P*(D|H). Koehler and Macchi (in press) argue that whether the information is formulated as a single or multiple target may have an effect. The target of a probabilistic statement identifies a problem-relevant sample space or reference class. Single targets (e.g., your patient has cancer) offers the smallest reference class (n = 1, your patient) by directing attention to a singular occurrence. This deters exemplar production because there are no other patients to consider. Consequently, people who receive single targets are unlikely to think about other people having a positive test. In contrast, multiple targets (e.g., number of positive test given cancer) encourage exemplar production because they offer a larger reference class (n > 1, the number of tests) within which to consider other patients with positive tests.

Research has shown that this proposed explanation has two shortcomings. First, though the explanation is based on the ambiguity of *P*(H) and *P*(~H), it is unclear why this lack of clarity of the independence would lead people to confuse the *P(H|D)* and *P(D|H)*, i.e., the explanation is circular. The second shortcoming is more of a discrepancy between this explanation and the first explanation. In Macchi's explanation, she suggests that people's interpretations (i.e., formulation #(2)) rely on the sample space defined by feature (D). However, the *sample-space framework* indicates that this is a deviant and dubious basis for probability judgment because people do not base their judgments on feature (D), but rather on category (H).

Probability and Frequency. The final explanation of the *confusion of the inverse* is the frequency hypothesis (Gigerenzer, 2000, 2002; Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2001; Cosmides & Tooby, 1996). Gigerenzer and Hoffrage (1995) have demonstrated that near- Bayesian answers often can be elicited using natural frequencies instead of probabilities. Thompson and Schumann (1987) showed that the use of frequency formats reduced the number of times subjects committed the *confusion of the inverse* in a court trial using "fallacious" statistical argument. However, Eagan (1972) used frequencies and participants still fell prey to the *confusion of the inverse.*

This explanation has been shown to be promising in explaining why people confuse conditional probabilities. However, it does have several short comings. Firstly, it has been shown only in a single study to reduce the number of *confusion of the inverse* errors that subjects commit (Thompson & Schumann, 1987). However, Eagan (1972) previously showed that physicians still fell prey to the *confusion of the inverse* when presented with information about breast exams in a frequency format. Secondly, there has been much debate as to the validity of Gigerenzer's argument about the use of natural frequencies versus Kahneman and Tversky's use of point probabilities when exploring probabilistic judgment (Kahneman & Tversky, 1996; Gigerenzer, 1996).

Probabilities versus Frequencies

There has been a long standing disagreement among statisticians and philosophers about the interpretation of probability. This argument pits two-sides: the Bayesian School and the Frequentist School. The Bayesian School interprets probability as a subjective measure of belief. This allows for the assignment of probability to unique events and requires these assignments obey probability parameters (Kahneman & Tversky, 1996). The counter argument is presented by the Frequentist School. They interpret probability as a "long-run relative frequency" and do not assign probability to unique events (Gigerenzer, 1996).

This argument led to a debate in the literature about the use of heuristics and biases. Kahneman and Tversky have used probabilities to assess human judgment and have shown that people are often biased when faced with probabilistic reasoning. This bias has been attributed to the use of heuristics, simple rules of thumb, which people utilize when trying to think probabilistically. Heuristics have been used as a descriptive basis for evaluating human judgment under uncertainty, with results replicated across numerous empirical studies. Kahneman and Tversky concluded that people are not natural Bayesians; people are not Bayesian at all (Kahneman & Tversky, 1972).

However, Gigerenzer (2000) argued that there is no normative basis for diagnosing judgments of uncertainty as wrong or biased. Rather, the heuristics and biases demonstrated by Kahneman and Tversky (1984) can be eliminated by the use of natural frequencies. In other words, heuristics and biases occur due to the use of point probabilities. When frequencies are used, these heuristics and biases often disappear. Specifically, Gigerenzer (2000) argues that people can indeed be Bayesian; however, our Bayesian systems are adapted for natural frequencies rather than point probabilities as asserted by Kahneman and Tversky. Thus, when confronted with natural frequencies (out of 100 trials, 50 were correct) and point probabilities (50% of the time they were correct), our inherent Bayesian system functions better with natural frequencies.

Using this approach, Gigerenzer has been able to show that under the right conditions, overconfidence disappears (Gigerenzer et al., 1991); the conjunction fallacy in the Linda problem can be minimized (Fiedler, 1988; Hertwig & Gigerenzer,1995); and people's answers are close to Bayes in other settings (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996). However, his arguments still can not account for all of the errors that people make when using statistical information. Thus, while each side seems to have its arguments, it is still unclear which statistical format is more appropriate.

Past Research to Eliminate the Confusion of the Inverse

Despite all of the research trying to explain the *confusion of the inverse*, very little research has tried to eliminate its effect on people. Birnbaum and Mellers (1983) indirectly discovered one way to minimize the effect of the *confusion of the inverse*. They conducted a study in which respondents saw 80 versions of the same probabilistic problem in a 2-hour period. For each of the problems, they varied the baserate and the reliability of the evidence. Due to the repeated instances of the same problem, participants' tendency to respond with the conditional inverse all but disappeared. This is believed to occur because when presented with repeated instances of the same problem, subjects develop a different strategy from the one they used when answering just one such problem (Slovic, Lichtenstein, & Edwards, 1965).

Though this result occurred due to repeated exposure, many real-world probabilistic problems do not occur in such a fashion. Rather many real-world problems are singular (Fischoff, Slovic & Lichtenstein, 1979), as with cases presented to jurors. Thus, the elimination of the *confusion of the inverse* in the above study may not be generalizable.

In work by Christensen-Szalanski and Beach (1982), Lichtenstein and McGregor (1984), and Pollatsek, Well, Konold, Hardiman, and Kobb (1987) subjects were given a 2 by 2 Bayesian table defining all possible combinations of the hypothesis (true or false) with the evidence (supporting or non-supporting). The use of the Bayesian tables had an effect on reducing errors associated with *confusion of the inverse*. Reexamining the previous example by Eddy (1982), the use of a Bayesian table can be illustrated as follows:

> One of your patients has a lump in her breast. You are almost certain that it is benign, in fact you would say there is only a 1% chance that it is malignant. But just to be sure, you have the patient undergo a mammogram, a breast X-ray designed to detect cancer.
>
> You know from the medical literature that mammograms are 80% accurate for malignant lumps and 90% accurate for benign lumps. In other words, if the lump is truly malignant 80% of the time and will falsely say it is benign 20% of the time. If the lump is benign, the test results will say so 90% of the time and will falsely declare that it is malignant only 10% of the time.
>
> Sadly, the mammogram for your patient is returned with the news that the lump is malignant. What are the chances that it is truly malignant?

A table for this example would resemble:

Results of X-Ray	*Malignant Lesion*	*Benign Lesion*
Positive	.80	.10
Negative	.20	.90

The hypotheses are either a malignant or benign lesion, while the evidence are either a positive or negative X-rays. The entries in the cells are conditional probabilities (e.g., $P(D|H)$ = .80). Such a table reduced the effects of the *confusion of the inverse* and helped people to better interpret conditional probability statements. However, the tables alone were only able to reduce the effect, not to eliminate it entirely.

Additionally, work done by Nisbett (1993; Nisbett, Krantz, Jepson, & Kunda, 1984, 2002) and others has shown that graphical information can aid in the interpretation of statistical information. However, no research has been done to investigate the effects that graphical information such as a pie chart or Venn diagram has on the *confusion of the* *inverse.* Thus, additional research is needed on how to achieve an elimination of the *confusion of the inverse.*

Modeling Juror Decision-Making

Over the past several decades, cognitive psychologists have spent a large amount of time investigating decision-making in jurors. The question is why? The answer has to do with the task that jurors face. A juror's task, although complicated at times requiring higher-order mental processing, is relatively well-defined with set rules and guidelines. Additionally, a juror's task is usually a singular experience shielded (in most cases) by many socially mediated external influences (such as outside parties, detailed discussion with many people, etc.). Finally, a juror's task has provided many in cognitive psychology a *playground* to develop and test different types of decision-making models. According to Hastie (1993), the study of juror decision-making has evolved from four types of descriptive models; (1) *Cognitive Algebra,* (2) *Stochastic Process*, (3) *Information Processing*, and (4) *Probability Theory* models.

Each of these four models has its own inherit advantages and disadvantages. Cognitive algebra models have been shown to clearly represent a juror decision task accurately as well as provide evidence as to the weighting of information presented during a trial. However, the disadvantage of this model is that it tends to only work well with simplified, abbreviated judgment tasks (Hastie, 1993).

Stochastic process models provide a unitary formulation of one's decision threshold as well as the inclusion of evidence evaluation. This allows the researcher to evaluated individual adjustments in judgment based on changing information during a trial. However, stochastic models tend to separate evidence evaluation and threshold comparison into two separate processes and do an inadequate job of evaluating the combination of these two processes by a juror when arriving at a decision.

The foremost information processing model used in juror decision-making is the cognitive story model developed by Pennington and Hastie (1981). This model involves three component processes by which juror construct a story of the events based on the evidence leading to a decision. The first component involves the actual construction of a story based on evaluation of the evidence. Once a story is constructed, the juror then represents all of the possible decision-making alternative (i.e., verdicts) by learning about their attributes and elements for each. Finally, once the evidence evaluation and decision alternatives are decided upon, the juror reaches a verdict through the classification of the constructed story into the best fitting verdict (or decision alternative) (Pennington & Hastie, 2000). The biggest advantage to this model is it's ability to combine both the evidence evaluation and the threshold determination for decision alternatives into one decision process. Despite this, the cognitive story model still has many ill-defined processes not yet fully understood.

The probability theory models are based on Bayes Theorem. As discussed, Bayesian probability is formal, rational and is by far the best at showing how to improve juror decision-making. However, as has been discussed in other tasks (such as the medical profession), people are imprecise in their use of probabilities. Additionally, some research has shown that it does not provide an adequate description of the juror decision-making process. Yet, probability theory based on work using Bayes theorem does provide a good description as to how information should be used/combine to make an appropriate judgment.

Juror use of Statistical Evidence. Until now, the research on the use of statistical information has primarily focused on work with physicians and clinicians. However, one arena in which the *confusion of the inverse* has been studied little is legal decision-making. Yet, the legal system's use of statistical evidence has grown considerably in recent years. Whether or not jurors are equipped to interpret statistical evidence remains very much in question.

Over the past half-century, the use of statistical evidence in the courtroom by jurors has been the topic of much debate. Many researchers have questioned the use of statistical evidence in trials (Finkelstein & Farley, 1970; Weinstein, Mansfield, Abrams, & Berger, 1983; Tribe, 1971; Saks & Kidd, 1981; Wells, 1992; Wells & Luus, 1990; Wright, MacEacher, Stoffer, & MacDonald, 1996; Niedermeier, Kerr & Messe, 1999). Most of this debate centers on a juror's ability to use and comprehend statistical evidence appropriately (Faigman & Baglioni, 1988).

Starting, in the early 1970s, a landmark case brought the use of statistical evidence to the forefront (Zabell, 1993). The trial, *Griggs v. Duke Power Co.*, established that if,

> …the plaintiff could show that an employment practice had a disparate impact on a protected class of citizens (e.g., women, minority group members), then the burden of proof (to demonstrate 'non-discrimination') shifts to the defendant (Zabell, 1993, p.268).

Therefore in such a case, the evidence makes the employer justify hiring and promotion decisions by creating a purely statistical category of discrimination. Additionally, other court rulings (e.g., *Castaneda v. Partida*, 1977; *Hazelwood School District v. US*, 1977) increased the use of statistical evidence in the courtroom. The reliance on statistical evidence includes the Supreme Court, which has used statistical formulas in to explain and justify their decisions.

Recently, jurors have faced statistical issues such as the determination of confidence intervals for multiple comparisons and the interpretation of Simpson's Paradox type patterns in cross-classified frequency data (Zabell, 1993). Still, examples exist which shows the misuse of statistical evidence in jury trials (Feinberg, 1989).

The most common occurs with DNA testing. DNA testing, which is often framed in probabilistic terms and has been claimed to be incontrovertible, has become wide spread. However, whether or not jurors use this information correctly has been called into question. For example, during a trial in Canada, a child support application was rejected despite a DNA match on a blood test revealed only a 99.8% chance that the man being sued was the biological father (Niedermeier, Kerr and Messe, 1999).

More famously, jurors acquitted O.J. Simpson of murder despite blood evidence which indicated that only 1 out of 170 million people had the same genetic markers as the defendant (Linedecker, 1995). Researchers have hypothesized that these results occurred due to people being reluctant to make pro-plaintiff decisions that are based on solely probabilistic evidence (Wells, 1992).

One of the biggest opponents to the use of statistical evidence in trials is Tribe (1971). Tribe's biggest criticism revolves around the assumption that jurors are capable of using statistical evidence appropriately. Rather, he argues that jurors ignore statistical evidence and rely on less concrete evidence when making judgments (e.g., eyewitness testimony which is not always entirely accurate). Therefore, much of the work conducted on juror decision-making and statistical evidence has centered on juror's ability to use this type of evidence (Wells, 1992).

The overwhelming consensus is that jurors tend to ignore statistical evidence. This especially holds true when "naked statistics", or probabilities that are not case specific in that they are present prior or independently to the particular case being tried, are used (Wright, MacEachern, Stoffer, & MacDonald, 1996).

Specifically, "naked statistics" are generic probabilities that are given and exist regardless of a court case, but are used as part of the evidence in a court case. Therefore, in the Blue/Green Cab problem, the knowledge that 85% of all cabs in the city are Green is a "naked statistic". This is because the ratio of Blue and Green cabs exists without the cases occurring.

What makes this type of statistic interesting is that though it might have a bearing on a case, people tend to dismiss it altogether unless there is some other type of evidence attached to it (i.e., eyewitness testimony, etc.) (Neidermeier, Kerr, & Messe, 1999). Hence, most researchers have explored why other evidence is used when making judgments as opposed to statistical.

For example, Faigaman and Baglioni (1988) showed that jurors ignore statistical evidence. This occurred even when an expert statistician explained how to use Bayes to understand what the statistical evidence was demonstrating. Even with the interpretation by a statistician, it is unclear if jurors can even interpret, let alone use, statistical information appropriately.

Practically no research has looked at whether jurors fall prey to *the confusion of the inverse* when faced with probabilistic reasoning. If jurors are subject to the *confusion of the inverse*, they will of course interpret statistical evidence incorrectly. And, this incorrect interpretation of statistical evidence may lead them to very well ignore statistical evidence in favor of other evidence. Therefore, there is a need to study the *confusion of the inverse* in the context of juror decision-making to better understand juror's use of statistical evidence.

Research Question

There are three purposes to the dissertation. (1) Currently there has not been any research conducted to determine whether jurors fall prey to the *confusion of the inverse*. Though research has shown physicians and clinicians are susceptible to this problem, jurors may or may not succumb to the same extent. Therefore, the research will determine to what extent jurors are susceptible to the *confusion of the inverse*.

Additionally, it is important to investigate further why $P(H|D)$ is so often confused with $P(D|H)$. Some researchers have implied that the *confusion of the inverse* may be occurring because people simply think the two conditional probabilities sound the same (Villejoubert & Mandel, 2002). Presently, no research has explicitly asked subjects if one conditional probability sounds like the other. Therefore, the research will determine whether or not subjects equate $P(H|D)$ with $P(D|H)$ due to "sounding" similar.

(2) Assuming jurors are subject to the *confusion of the inverse*, what is the reason? Two possible reasons involving formatting that have been postulated in the literature. First, there is considerable debate among researchers as to how statistical

information should be presented. Although point probabilities are standard, others argue for natural frequencies. Thus, how should the legal system present statistical evidence to juror's to better improve understanding -- probabilities or frequencies?

Second, Macchi (1995, 2000) and Koehler and Macchi (in press) argue that the formulation of diagnostic target information is a key factor, i.e., whether the probabilistic statement is formulated as a single or multiple target. How should the legal system formulate diagnostic information so that statistical evidence is understood? As a single or multiple target?

(3) Once the factors behind the *confusion of the inverse* are understood, the research will attempt to reduce the effect. Studies have shown that the use of graphical information reduces the effect of the *confusion of the inverse.* Therefore, the research is intended to determine if the legal system should present statistical evidence graphically to better improve understanding. Also, does a Bayesian table format or some type of graphical representation provide a better illustration of the evidence?

Pilot Study

To begin, the researcher needed to assess the criteria needed by potential jurors to reach a verdict of not guilty or guilty beyond a reasonable doubt, i.e., potential jurors' level of sensitivity for guilt needed to be assessed. To accomplish this, a pilot study was undertaken in which guilt levels for several types of crimes were assessed.

Method – Pilot Study

Prior to any study being conducted Internal Review Board approval was requested and granted from Kansas State University and Briar Cliff University.

Participants

Thirty-six undergraduate students from a small Midwestern liberal arts university completed the pilot study. Each student was told that participation was completely voluntary, and those who complete the study were offered extra course credit for their participation. Of the 36 participants, 11 were male and 25 were female. The mean age of the participants was 20.7 years and over 85% had no experience with a legal trial.

Materials

A survey was composed to assess the criterion needed for proof beyond a reasonable doubt. The survey contained two definitions of beyond a reasonable doubt. Definition A was taken from a law text and definition B was taken from a court case in Florida. The first two questions asked participants to rate the likelihood (0 = not at all sure, 100 = completely sure) of finding someone either guilty or not guilty of an unspecified crime. The remaining questions asked participants to rate the likelihood of finding someone either guilty or not guilty of (1) murder, (2) sexual assault, (3) burglary, (4) assault, (5) manslaughter, (6) kidnapping, and (7) arson. Finally, some demographic questions were asked (See Appendix A: Pilot Study Survey).

Procedure

Participants were told the purpose of the present study was to assess “reasonable doubt.” They were told to answer as truthfully and honestly as possible, and that all answers would remain anonymous. Following the collection of surveys, the participants were debriefed and told their answers would aid in the development of future stimuli.

Results – Pilot Study

Overall results (for all crimes and results, see Table 1) showed that participants had clear thresholds for guilty beyond a reasonable doubt, 90.3% and a clear threshold for what they would consider to be not guilty beyond a reasonable doubt 85.4% respectively. Although, the concept of not guilty beyond a reasonable doubt is not a true legal concept, the assessment of it led to the lower level of criteria in which subjects felt confident enough that a person did not commit a crime. Additionally, participants had clear thresholds for guilty and not guilty beyond a reasonable doubt for each of the previously mentioned crimes. Specifically, the crimes of murder (93.5% - guilty; 91.2% - not guilty) and manslaughter (93.7% - guilty; 90.4% - not guilty) had the highest criteria for guilty and not guilty, respectively whereas the crime of burglary (81.9% - guilty; 77.6% - not guilty) and assault (86.0% - guilty; 81.5% - not guilty) had the lowest criteria for guilty and not guilty.

Interestingly, each case had a definite criterion for guilty beyond a reasonable doubt and not guilty beyond a reasonable doubt. And for each case, there seemed to be a notable *gray area* in which reasonable doubt was not a sure concept. This fluctuating range indicates that the criteria of reasonable doubt is not a static concept as previously believed, rather, reasonable doubt is a fluid and dynamic concept that appears to be case specific.

Discussion – Pilot Study

The purpose of the pilot study was to provide information needed to design future studies. Specifically, the study provided estimates of the threshold levels for specific crimes. Based on the results, 4 crimes were selected for future research: murder, sexual assault, arson, and kidnapping. Of the crimes selected, murder had a high threshold level; the other 3 were relatively equivalent, having thresholds in the range of the mid-eighties. Therefore, future studies were developed using the crimes as the stimuli.

Legal Case Search Study

In order to begin the first study, the stimuli needed to be gathered. In order to maintain ecological validity, real court cases were chosen to use in future studies. However, the availability of real court cases for use is wanting. This is most likely due to the difficulty in obtaining court cases that are real. Most court cases are open to the public. Still, much of the cases themselves are confidential and much of the relevant information is lost due to the court restrictions. However, to achieve ecological validity several steps were taken, leading the researcher to spend time gathering information for court cases across a 4 month period.

To maintain ecological validity for the jurors, the researcher needed to determine some of the nuances that exist when being selected to participate in a jury trial. Thus, the researcher examined court information about the types of procedures used in a real trial. Once that information was obtained, the retrieval of the court cases proceeded.

Over a four month period, the researcher visited a local and county courthouse in a small mid-western city several times. The specific purpose was to find court cases that were to be used for future studies. The four cases types had already been selected from the pilot study above, therefore only the cases themselves were needed. In order for a case to be considered applicable, it had to meet several criteria. First, the cases had to match one of those chosen from the pilot study. Second, the case had to be in a case that when to Appeals Court. An Appeals Court case was selected because; this type of case was most likely to have resulted in a verdict that may have been ambiguous and not cut and dry. Finally, the Appeals Court case had to have with it, most of the relevant information. This meant that the cases needed to have a section entitled, *Facts of the*

Trial, and the complete juror instructions presented at the trial. The *Facts of the Trial* section contained a 2-5 page summary of the original trial proceedings including evidence and outcome. The jury instructions were important to have because they were to be given to the future subjects just as they were presented in the original trial.

Although the selected court cases appear to be simple to find, matching all of the criteria necessary was not quite as easy. The cases that were finally decided upon were taken from hundreds of potential cases in the two court houses. Each case met the criteria above and the information in each was parsed down to be included in future studies.

Study One

Most research has focused on the *confusion of the inverse* in medical decision-making. However as stated previously, research has yet to evaluate the *confusion of the inverse* in legal decision-making. Jurors, unlike physicians or clinicians, are not professionals in a specific field, thereby increasing the risk for using statistical information incorrectly (Finkelstein & Farley, 1970; Weinstein, Mansfield, Abrams, & Berger, 1983; Tribe, 1971; Saks & Kidd, 1981; Wells, 1992; Wells & Luus, 1990; Wright, MacEacher, Stoffer, & MacDonald, 1996; Niedermeier, Kerr & Messe, 1999). Thus, the primary intent of this study was to evaluate whether or not jurors indeed commit the *confusion of the inverse*.

The secondary intent of the study was to evaluate if participants believe that the probability the defendant matches the evidence and the probability the evidence matches the defendant are the same statement. Whereas the previous research question assess the *confusion of the inverse* within a specific task, this research question will be done by asking subjects about these two conditional probabilities with a direct question not tied to a context. This is being done because research has yet to evaluate whether or not people believe these two separate statements mean the same thing (Villejoubert & Mandel, 2002).

Method – Study One

Prior to any study being conducted, Internal Review Board approval was requested and granted from Kansas State University and Briar Cliff University. The statistical evidence was framed (similarly to problems in medical and clinical research) in terms of probabilities with a single target class. Additionally, past research had indicated that repeated exposure to statistical inference problems leads to a reduction of the *confusion of the inverse* (Birnbaum & Mellers, 1983). However, many real-world problems are singular (Fischoff, Slovic & Lichtenstein, 1979), as with cases presented to jurors and the presentation of many cases might lead to false results due to repeated exposure. Therefore, jurors were only exposed to a small number of cases (4) to avoid this issue

Participants

One hundred and five undergraduate students from a small Midwestern liberal arts university completed the study. Each student was told that participation was completely voluntary, and those who completed the study were offered extra course credit for their participation. Of the 105 participants, 40 were male and 65 were female with a mean age of 19.7 years. Most of the participants had never been called for jury duty (91%) and over three-quarters had no prior experience with a legal trial.

Materials

Based on information from the pilot study, four crimes were selected – murder, sexual assault, kidnapping and arson. For each of the crimes, real court cases were identified from local county and federal court houses[3]. Each of the cases was abbreviated

[3] The original intent of the project was to find cases in both criminal and civil trials. However, the researcher spent over 4 months searching for cases and was able to only find 4 criminal cases to meet the necessary criteria. Additionally, while real court cases were obtained for use in this study, the descriptions were limited since much of the material was considered confidential by the local and federal governments. Thus, cases were chosen based on the availability of the information needed. Since most information about the civil court cases was suppressed, they were excluded from this dissertation.

to eliminate irrelevant information. Additionally, all physical evidence was removed from the case descriptions, except for a solitary piece of statistical evidence. The case description consisted of facts from the trial section (similar to those seen in an appeals proceedings) (See Appendix B: Study One Court Descriptions).

Besides the court case descriptions, juror instructions were provided for each case. The instructions, attached to the backside of the cases, contained the information needed by the jurors to achieve a verdict for criminal cases (See Appendix C: Juror Instructions for Each Case). In addition to the juror instructions, a sample juror "swearing in" form was obtained (See Appendix D: Participant Instructions and Swearing-In Procedure)[4] .

At the end of each case, subjects were asked four questions. The first question asked the jurors to rate the strength of the evidence on a 7-point Likert scale, where 1 was "not at all strong" and 7 was "extremely strong." The second question asked participants to estimate the probability that the evidence matched the suspect by indicating a number between 0 – 100, where 0 was "not at all" and 100 was "complete." The third question asked participants to estimate the probability the suspect committed the specific crime by indicating a number between 0 – 100, where 0 was "not at all" and 100 was "complete." The final question asked participants to render a verdict based on the evidence.

Following completion of all cases, several demographic questions were asked. Finally, the last two question on the packet asked subjects to decide if the following two statements were the same: The probability that the "defendant matches the evidence" and

[4] In order to maintain as much external validity as possible, real court cases, juror instructions and juror swearing in procedures were used. It was believed that real cases, instructions, and swearing in procedures would add an inherent realism to the study, which was necessary to maintain a certain level of external validity.

the probability that the "evidence matches the defendant." Once participants had made a choice, they were asked to rewrite these statements in their own words. Specifically, participants were asked to explain what they believed the two statements meant (See Appendix C: Juror Instructions for Each Case & Appendix E: Survey Questions at the End of Packet).

Procedure

Participants were told the purpose of the present study was to aid in the consideration of appeals cases going to a Federal Court. They were told that they would be reviewing four cases and that they were to be honest as possible remembering that all answers would remain anonymous. Finally, they were to be sworn-in as jurors for a federal court case.

Once the preliminary proceedings were finished, each juror received a packet containing the four case descriptions, juror instructions, and questions for each case. Jurors were instructed to evaluate the cases with the evidence given and that, upon completion of the packet, to turn it over and remain seated until all had finished. Once the packets were completed and collected, the participants were debriefed and told about the research. From start to finish, the study lasted approximately 60 minutes.

Results – Study One

Calculation of Bayesian Answers for Cases

In order to determine if subjects committed the *confusion of the inverse*, the Bayesian solutions for each of the cases were calculated. By applying Bayes Theorem, the probabilities that the suspect matches the evidence based on the description for kidnapping, arson, sexual assault, and murder are .94, .77, .90, and .61 respectively (See Appendix F: Calculations for Bayesian Solutions to Each Court Case).

Analytical Results – Study One

Overall results showed that most people did fall prey to the confusion on the inverse(See Table 2). Specifically, a one-sampled t-test was performed to determine if subjects' estimates of a match between the suspects and the evidence were significantly different from the correct response. Results showed that for three of the four cases, kidnapping [*t* (104) = -5.89, p < .001], murder [*t* (104) = -6.73, p < .001] and sexual assault [*t* (104) = 17.89, p < .001], the subjects were significantly different from the correct Bayesian solutions. Surprisingly, subjects were extremely accurate for the case on arson [*t* (104) = -.432, p > .001][5] (For a list of means see Table 2).

When comparing subjects' answers with the inverse probability, *P(D|H)*, results showed that two of the four were not significantly different: murder [*t* (104) = -1.12, p > .001] and kidnapping [*t* (104) = -3.261, p > .001], and the other two were significantly different: arson [*t* (104) = 3.53, p < .001] and sexual assault [*t* (104) = 4.51, p < .001] (See Table 3). For sexual assault, subjects' answers were far closer to *P(D|H)* than they were to *P(H|D)*. Finally all answers for each case were significantly different from given baserates.

[5] Do to the number of analyses performed, a Bonferonni correction was implemented to reduce the likelihood of a Type I error. Therefore a criterion of p < .001 was used to determine significance.

Additionally, results showed that half (53) of the 105 participants believed that the probability that the defendant matches the evidence and the probability that the evidence matches the defendant were the same statement. This result is basically a chance result. Therefore, it was not surprising a MANOVA indicated there was no significant difference in subjects' who said they were the same and subjects who indicated they were not the same. Thus, subjects committed the *confusion of the inverse* regardless of their belief in whether the two statements meant the same thing.

Finally, regression analyses were conducted to determine which of the three dependent variables (strength of evidence, estimates as to the probability of a match to the subject given the evidence, or estimates as to the probability the subject committed the crime) predicted subjects' assignment of verdicts. Stepwise regression results showed that for each case subjects' belief in the strength of evidence was the best predictor of verdict (See Tables 4 -7). Strength of evidence also accounted for a significant proportion of the variance in assignment of verdict for kidnapping [$R^2 = .23$, $F(1, 103) = 30.73$, $\underline{p} < .001$], arson [$R^2 = .23$, $F(1, 102) = 30.87$, $\underline{p} < .001$], murder [$R^2 = .43$, $F(1, 101) = 75.53$, $\underline{p} < .001$], and sexual assault [$R^2 = .47$, $F(1, 103) = 89.83$, $\underline{p} < .001$].

Discussion – Study One

As with previous research using conditional probabilities and other populations, jurors also committed the *confusion of the inverse*. Rather than answering with the correct conditional probability of $P(H|D)$, subjects consistently answered using the inverse probability of $P(D|H)$. In only one case were subjects accurate in their estimation of the probability of a match given the evidence. Thus, the results of the first study indicate that subjects are indeed mistaken in their interpretation of statistical information (as predicted by Finkelstein & Farley, 1970; Weinstein, Mansfield, Abrams, & Berger, 1983; Tribe, 1971; Saks & Kidd, 1981; Wells, 1992; Wells & Luus, 1990; Wright, MacEacher, Stoffer, & MacDonald, 1996; Niedermeier, Kerr & Messe, 1999).

Despite the misuse of the statistical evidence when evaluating the case, nearly half the subjects (right at chance) indicated they knew there was a difference between the probability that the defendant matches the evidence and the probability that the evidence matches the defendant. And, when reading subjects' responses to the question, "what do these two statements mean?", it became evident that while some may have indicated they understood the difference, they actually did not. This qualitative data was supported by analyses that showed there was no difference in responses between subjects who indicated there was a difference and those who did not indicate there was a difference. This lends further evidence that subjects do not interpret conditional probabilities accurately.

Even with incorrect evaluations of the evidence, subjects were still able to consistently respond to each case with a verdict: guilty or not guilty. The results indicated that subjects' belief in the strength of the evidence or their subjective assessment of the evidence was actually the best predictor of their verdicts. In other words, although subjects misinterpreted the evidence, it was their belief in the evidence strength (not their actual understanding of the evidence) that pushed them to a verdict. When informally discussing the study with subjects, many indicated they made assumptions about the case independent of the evidence, which should not occur in a trial setting.

The results of this study clearly indicate subjects are unable to interpret conditional probabilities correctly and have a propensity to commit the *confusion of the inverse*. Therefore, a second study was conducted to determine the reasoning behind subjects committing the *confusion of the inverse*.

Study Two

The previous study indicated that, like clinicians and physicians, jurors fall prey to the *confusion of the inverse*. The question is why? Theorists have suggested that the *confusion of the inverse* may be committed due to formatting issues (Gigerenzer, 2000, 2002; Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2001; Cosmides & Tooby, 1996; Macchi 1995, 2000; Koehler and Macchi, in press). Typically, probabilistic reasoning problems are presented using point probabilities and a single target class. However, researchers have suggested two different views that may aid in better understanding of these problems.

First, some researchers have suggested that point probabilities are not easily understood (Gigerenzer, 2000, 2002; Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2001; Cosmides & Tooby, 1996). Rather natural frequencies are a more natural form of presenting information for people. This Frequentist argument is supported by showing that misuse of point probabilities can be eliminated by using natural frequencies (Gigerenzer et al., 1991; Fiedler, 1988; Hertwig & Gigerenzer,1995; Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996). Thus, the use of point probabilities could be leading people to commit the *confusion of the inverse*, whereas the use of frequencies may reduce this effect.

The second argument has to do with the target that is used, specifically a single or multiple targets. As already indicated, typical probabilistic reasoning problems use single targets. Research has shown that the use of a single target reduces the effect of exemplar cueing, i.e., thinking about other instances occurring for the same event. However, the use of a multiple target has been hypothesized to be a better format.

Specifically, a multiple target encourages exemplar production because they offer a larger reference class, thereby increasing the likelihood of looking at other potential people.

In order to evaluate the reason behind people's potential to commit the *confusion of the inverse*; a study was designed to investigate the impact of formatting. In particular, the stimuli from the previous study were augmented and used for this study. Although jurors in the first study did not commit the *confusion of the inverse* in the arson case, it was included in this study in order to evaluate whether the previous result was an anomaly.

Method – Study Two

Participants

Two hundred and forty three undergraduate students from a small Midwestern liberal arts university complete the study. Each student was told that participation was completely voluntary, and those who complete the study were offered extra course credit for their participation. Of the 243 participants, 87 were male and 156 were female with a mean age of 20.8 years. Most of the participants had not been called for jury duty (90%) and over three-quarters had no prior experience with a legal trial.

Materials

As with the first study, 4 criminal case descriptions were used - kidnapping, arson, murder, and sexual assault. Each of the cases was abstracted to eliminate all extraneous information. Additionally, all evidence was removed from the case descriptions, except for a solitary piece of evidence that was statistical in nature. The case description consisted of the trial section (similar to those seen in an appeals proceedings), containing the only evidence (statistical in nature) the jurors had to base their judgments on.

A 2 (Format Type: Probabilities or Frequencies) x 2 (Target Class: Single or Multiple) x 2 (Reference Class: Small or Large) between subjects design was used. The evidence was either presented as point probabilities or frequencies. An example of evidence presented as a probability would be ".15," whereas that same probability in frequencies would be "15 out of 100." Additionally, evidence was presented as a single or multiple target. An example of a single target is the chance the suspect would match the blood found on the shirt if he were NOT the source is 15 in 100, whereas a multiple target is 15 in 100 people in a town who are not the source would nonetheless match the blood found on the shirt. An example of a small reference class would be a city the size of 500 people and a large reference class would be a city the size of 5,000,000 (See Appendix G: All Possible Combinations of the 2x2x3 Design for Each Case).

Besides the court case descriptions, juror instructions were provided for each case. The juror instructions, attached to the backside of the cases, contained the information needed by the jurors to arrive at a verdict beyond a reasonable doubt (See Appendix C: Juror Instructions for Each Case). In addition to the juror instructions, a sample juror "swearing in" form was provided (See Appendix D: Participant Instructions and Swearing-In Procedure)[6]. Finally, at the end of each case, subjects were asked the same four questions, as well as the questions about demographic information from Study One.

[6] The swearing-in procedure was used for the same reasons indicated in Study One.

Procedure

Participants were told the purpose of the present study was to aid in the consideration of appeals cases going to a Federal Court. They were told that they would be reviewing four cases and that they were to be honest as possible remembering that all answers would remain anonymous. Finally, as part of being jurors for a Federal Court, they were to be sworn-in as jurors.

Once the preliminary proceedings were finished, each juror received a packet containing the four case descriptions, juror instructions and questions for each case. Jurors were instructed to try the cases with the evidence given and that upon completion of the packet to turn it over and remain seated until all had finished. Once all the packets were completed and collected, the participants were debriefed and told about the study. From start to finish, the study lasted approximately 60 minutes.

Results – Study Two

Two different difference scores for each case were calculated[7]. Specifically, difference scores were calculated (1) between jurors' estimates of a match between the suspects and the evidence and the correct Bayesian solutions and (2) between estimates of a match between the suspects and the evidence and the inverse probability (See Table 8). These scores were used to determine if differences existed between the different formulations thought to reduce the effects of the *confusion of the inverse*.

[7] Difference scores are recommended when individual differences in change are appreciable (Rogosa & Willett, 1983).

Initial Raw Estimates of a Match

Before examining the difference scores, an analysis of raw scores was conducted. A three-way Multivariate Analysis of Variance (MANOVA) was conducted to determine if any differences existed between formulation types. Surprisingly, the MANOVA and preceding ANOVAs revealed no significant differences between the formulation types for any of the cases.

Differences Scores from Bayesian Solutions

A three way MANOVA was conducted to determine if any differences exist between the types of formulations for each case. Means can be found in Table 9. For each of the cases, results indicated that a significant interaction between target and reference class was present for the combined dependent variables of the difference between estimates of a match and the correct Bayesian solutions [Wilks' $\Lambda = .951$, $F(4, 231) = 2.99$, $\underline{p} < .05$, $\eta^2 = .049$, power = .79]. Additionally, a marginally significant three-way interaction was seen between format type, target, and reference class on the combined dependent variables of estimates [Wilks' $\Lambda = .963$, $F(4, 231) = 2.24$, $\underline{p} = .066$, $\eta^2 = .037$, power = .65].

Follow-up univariate ANOVAs indicated a significant interaction for target by reference class for the case of arson [$F(1, 234) = 4.49$, $\underline{p} < .05$, $\eta^2 = .020$, power = .56]. Specifically, there is a crossover interaction between target and reference class, where a single target in a large reference class and a multiple target in a small reference class produce responses that are closer to Bayesian solutions (See Figure 1).

Follow-up analyses also revealed a significant three-way interaction for the case of murder [F (1, 234) = 4.58, $p < .05$, $\eta^2 = .020$, power = .57]. In a large reference class, frequency judgments led to consistently larger deviations from the Bayesian solutions for both targets (i.e., single and multiple). However, in a small reference class, frequency judgments using a single target and probability judgments using a multiple target led to greater deviation from the Bayesian solution (See Figure 2). No other differences existed between the variables.

Furthermore, one-sampled t-tests were performed to determine if subjects' estimates of a match between the suspects and the evidence were significantly different from the correct Bayesian responses. Results indicated that for three of the four cases, kidnapping [t (244) = -9.29, $p < .001$], murder [t (242) = -10.05, $p < .001$] and sexual assault [t (244) = 21.19, $p < .001$], the subjects were significantly different from correct Bayesian solutions. As with the first study, subjects were accurate for the case on arson [t (243) = 1.35, $p = .177$] (See Table 10:)[8] One-sampled t-test for all possible combinations of the 2 (Format Type: Probabilities or Frequencies) x 2 (Target Class: Single or Multiple) x 2 (Reference Class: Small or Large) can be seen for each case in Tables 11.

Difference Scores from Reliability [P(D|H)]

A three way MANOVA was conducted to determine if any differences exist between the types of formulations for each case (Means can be found in Table 12). Results revealed a significant interaction between target and reference class for the

[8] Again, due to the number of analyses performed a Bonferonni correction was implemented to reduce the likelihood of a Type I error. Therefore a criterion of $p < .001$ was used to determine significance.

combined dependent variables (of the difference between estimates of a match for each of the four cases and the inverse probability) [Wilks' $\Lambda = .960$, $F(4, 231) = 2.40$, $\underline{p} < .05$, $\eta^2 = .040$, power = .68]. Additionally, a significant interaction between format and reference class was present for the combined dependent variables of the difference between estimates of a match for each of the four cases and the inverse probability [Wilks' $\Lambda = .950$, $F(4, 231) = 3.11$, $\underline{p} < .05$, $\eta^2 = .051$, power = .81].

Follow-up univariate ANOVAs indicated a significant three-way interaction for the case of murder [$F(1, 234) = 5.03$, $\underline{p} < .05$, $\eta^2 = .021$, power = .61]. Again, in a large reference class, frequency judgments led to consistently larger deviations from the Bayesian solutions for both targets (i.e., single and multiple). Yet, in a small reference class, frequency judgments did not differ among target (i.e., single and multiple) but probability judgments resulted in lower deviations from the reliability in a single target and greater deviations in a multiple target (See Figure 3). No other differences existed between the variables.

When comparing subjects' answers with the inverse probability, *P(D|H)*, results showed that only one was significantly different, murder [$t(242) = -1.87$, $\underline{p} = .062$]. While murder was significantly different, the other three were not: kidnapping [$t(244) = -6.28$, $\underline{p} < .001$], arson [$t(243) = 7.06$, $\underline{p} < .001$] and sexual assault [$t(244) = 5.32$, $\underline{p} < .001$] (See Table 13). However, for kidnapping and sexual assault, subjects' answers were closer to *P(D|H)* than to *P(H|D)*. One sampled- t-test results for all possible combinations of the 2 (Format Type: Probabilities or Frequencies) x 2 (Target Class: Single or Multiple) x 2 (Reference Class: Small or Large) can be seen for each case in Tables 14.

Additional Analyses

Just as in the first study, this study also showed that over half of participants indicated they understood the difference (52.2%) between the probability that the defendant matches the evidence and the probability that the evidence matches the defendant. This result is again due to chance. Therefore, it was not surprising that another MANOVA indicated there was no significant difference in subjects' who said they were the same and subjects who indicated they were not the same. In other words, subjects committed the *confusion of the inverse* regardless of their belief in whether the two statements meant the same thing. Thus, subject who believe they understand the difference between the two statements are just as poor as those who do not understand the difference in making probabilistic inferences.

Finally, regression analyses were conducted to determine which of the three dependent variables (strength of the evidence, estimates as to the probability of a match to the defendant given the evidence, or estimates as to the probability the defendant committed the crime) predicted subjects' assignment of verdicts. Stepwise regression results showed that for the cases of kidnapping and murder, subjects' belief in the strength of the evidence was the best predictor of verdict (See Tables 15-16). Strength of evidence also explained a significant proportion of the variance in assignment of verdict for kidnapping [$R^2 = .29$, $F(1, 240) = 97.87$, $\underline{p} < .001$] and arson [$R^2 = .29$, $F(1, 239) = 98.62$, $\underline{p} < .001$]. For the case of murder, subjects' estimates of the probability the defendant committed murder and the subjects' belief in the evidence were the best predictors of verdict, explaining a significant portion of the variance [$R^2 = .32$, $F(2, 237)$

= 55.49, p < .001] (See Table 17). However, the subjects' estimates of the probability the defendant committed sexual assault was the best predictor of verdict, explaining a significant portion of the variance [$R^2 = .19$, $F(1, 239) = 56.57$, $p < .001$] (See Table 18).

Discussion – Study Two

The primary result of study two was subjects' lack of differentiation between the manipulated variables. Specifically, subjects, regardless of which group they were in, did a poor job at using the information presented. Subjects routinely made the mistake of confusing $P(H|D)$ with the $P(D|H)$ in the cases of kidnapping, murder and sexual assault. In contrast, the case of arson consistently produced results similar to the correct Bayesian solution (as it did in study one), regardless of the format, target or reference class.

Within the case of murder, there was a reduction of the *confusion of the inverse* when probabilities were used in a large reference class (no matter target) or with a small reference class and a single target. Thus, there does appear to be some support for probabilities having an effect on the reduction of the *confusion of the inverse*, but only for a single case.

More interesting are subject's accurate responses in the case of arson regardless of format, target and reference class. This can be explained by examining the mean responses for the estimates of a probability of a match for each case. For the 4 cases, the mean responses where positioned between 77 – 87: 81.5, 78.7, 77.9, 86.6 for kidnapping, arson, murder, and sexual assault, respectively. It appears as though subjects consistently responded within this range regardless of case type. And, only in the case of arson did the Bayesian solution fall within this range - all other Bayesian solutions fell well outside. This indicates that subjects may be reluctant to make judgments about a match outside of the typical range, i.e., the previously indicated criterion for guilty or not guilty beyond a reasonable doubt (from the pilot study) might be an unattainable level for subjects to reach.

When it came time to decide upon a verdict for each case, subjects had no problems indicating guilty or not guilty. However, as in the first study, subjects ignored the validity of the evidence in favor of their own subjective assessment of the evidence. This was supported by regression analyses showing subjects' subjective assessment of the evidence was the best predictor of verdict. In other words, subjects appear to base their judgments of verdict on something beyond actual fact. If subjects had behaved according to the jury instruction and tried the cases with the information presented, they should have arrived at a verdict using only the factual information presented. However, due to their lack of understanding of the evidence, they relied on their own subjective assessment of the evidence. This idea was supported by subjects' indicating they assumed the defendant was guilty because he was on trial. To subjects' this meant that once the state had sufficient evidence to go to trial, that subject is likely guilty regardless of the evidence's validity.

Based on the inconsistent (or lack of) results from this study and the apparent discrepancy in subjects' belief in the evidence and the actual validity of the evidence, the proposed third study of eliminating the *confusion of the inverse* was abandoned. Though it is clear jurors commit the *confusion of the inverse*, the reason as to why is still not apparent. Therefore, a third study eliminating something not fully understood is impractical because, jurors must be making their inferences on something else: their belief in evidence strength. Yet, the possibility of studying the subjects and their use of personal belief in evidence strength existed. However, this research direction, although interesting, would not aid in overall intent of the dissertation, understanding the *confusion of the inverse* (which apparently exist regardless of associated task). Additionally, the lack of understanding of the evidence statistically, could be the variable leading to subjects using their belief in the evidence as the basis for rendering a verdict.

General Discussion

As expected, the dissertation showed that the *confusion of the inverse* is not only a problem with expert clinicians and medical professionals, it is also a problem for the novice faced with probabilistic mental reasoning – jurors in this case. However, the project did not discover why jurors committed the *confusion of the inverse*. Therefore, it appears that no minor manipulation in the literature is capable of removing the *confusion of the inverse.*

However, the study did lead to (1) the affirmation of the *confusion of the inverse* as the reason for incorrect probabilistic reasoning in juror decision-making, (2) discovery of the basis for juror judgment – subjective assessment of evidence strength, not actual evidence strength, and (3) the support of previous research indicating that the notion of "beyond a reasonable doubt" is an unrealistic legal criterion. Finally, additional considerations are discussed.

Affirmation of the Confusion of the Inverse

As mentioned before, people engage in incorrect probabilistic reasoning. However, why this is occurs has been debated in the literature, resulting in three possible explanations. The consensus is that the *confusion of the inverse* is the probable reason for this incorrect reasoning. Yet base-rate neglect and mis-integration of base-rate and case information has also been theorized.

One major finding of this project is the strong evidence that people commit the *confusion of the inverse* and not base-rate neglect or a mis-integration of base-rate and case information when faced with a legal probabilistic reasoning problem.

Base-Rate Neglect. The first hypothesis states that people neglect baserate information relative to the amount of attention that should be given to them (Cohen, 1981; Niniluotto, 1981). In other words, people do not take into account the prior probability judgment when faced with a probability problem. However, when examining participants' responses, a small number used the base-rate information as their response to an estimate of the probability of a match. This result is very similar to Tversky's (1981) argument that subjects who are given a problem where only the baserate is known use it as their response (Kahneman & Tversky, 1972; Lyon & Slovic, 1976; Hamm, 1993). Granted, in this study the base-rate was not the only piece of information, but it did appear to be used. For example, more than one participant left calculations on their survey packets. Nearly all of the calculations on packets involved an equation similar to:

$$P(H) \text{ x } P(D|H) = X$$

$$X \text{ x } 100 = \text{Probability of a Match}$$

Thus, subjects are not neglecting the base-rate, rather they are using it in an algorithm which is incomplete. This leads to the second hypothesis.

Mis-Integration of Baserate and Case Information/Evidence. The second hypothesis to account for why people arrive at incorrect answers is that they are following an incorrect integration rule. It is thought that people inappropriately combine the baserate and case-specific information when making their judgments (Bar-Hillel, 1980; Tversky & Kanheman, 1982). Tversky and Kanheman (1982) argued that people integrate these two kinds of statistical information such that their answer falls in between the two numbers, when the two kinds of statistical information are combined, the subject's answer tends to be closer to the case information/evidence than to Bayes Theorem (Fischoff & Bar-Hillel, 1984; Hamm, 1987). This hypothesis can account for why people respond to probabilistic problems with the answers closer to the sensitivity, *P(D|H)* than Bayesian solutions. Although a few subjects showed such calculations, the number was small. Instead, most people in the study indicated a probability of a match using the sensitivity for each case.

Confusion of the Inverse. The project has been centered on the argument that incorrect probabilistic reasoning is a result of subjects confusing the conditional probabilities *P(H|D)* with *P(D|H)*. This assumption, shared with most researchers, has been based on a thorough review of the literature. However, as many have pointed out, consensus is not always the best way to reach the correct solution (Weiss & Shanteau, 2003). Therefore, the results of study one and two have led to the conclusion that the *confusion of the inverse* is the reason people engage in probabilistic mental reasoning.

In studies one and two, many people (over 20% for each case) confused *P(H|D)* with *P(D|H)*. Specifically in study two, 42%, 26.7%, 25.5% and 31.4% of people responded with the sensitivity for kidnapping, murder, arson, and sexual assault, respectively. Neither of the other two possible hypotheses had percentages of this order.

Additionally, although many people believed they knew the difference between *P(H|D)* and *P(D|H)*, their qualitative data showed this was not the case. Respondents indicated they knew the difference, but would indicate the two statements meant "the evidence that points to the direction that ties the defendant to the evidence [is] not the same as the evidence pointing to the defendant." Here, the subject is saying the same thing twice. Although they are attempting to make a distinction with "points to the direction that ties", they are still just reiterating their view twice that the evidence points to the defendant, which is the data given (points to) the hypothesis, $P(D|H)$. They clearly did not understand the difference between the two and not too surprisingly used the sensitivity in their response.

Subjective Assessment of the Evidence

Although the reason that jurors commit the *confusion of the inverse* was not explained, the information they use to make their decisions in these situations was uncovered. Specifically, the evidence itself seemed to have very little to do with the actual judgment that jurors made – guilty or not guilty. In Study One and Study Two, regression analyses indicated that the best predictor of verdict was subjects' belief in the evidence or their subjective assessment of the evidence, not the actual evidence. In other words, subject's failed to understand the evidence and appeared to use their own personal opinions about the strength in evidence as the reasoning behind the verdicts they rendered. This was supported both by anecdotal evidence by subjects, by the analyses showing they misinterpreted the conditional probabilities, and regressions analyses showing they used strength of evidence when ruling on the case.

While talking with some subjects after the study, many indicated they ignored the information presented and made assumptions about the defendants. Often, this blind eye to the evidence was unintentional and was done without conscious effort. However, many subjects indicated that they intentionally ignored the evidence due to its statistical nature indicating, “I did not understand what was going on…so I made some assumptions!” These purposeful actions reaffirm the view, held by many, that jurors should not be exposed to statistical information (Tribe, 1971; Faigman & Baglioni, 1988).

Some research has hypothesized that jurors tend to dismiss statistical evidence altogether unless some direct evidence is attached to it (i.e., eyewitness testimony, etc.) (Neidermeier, Kerr, & Messe, 1999). This occurs whether the statistical information is case specific or not. This would be consistent with statements made by many of the subjects in this study. Thus, the subjective assessment of the evidence is independent of the actual evidence.

So, should we use evidence in the courtroom at all if jurors are not going to understand how to use it? Yes! The use of statistics in the courtroom helps to disseminate information about the case. Rather than abandoning the use of statistics, we should look to improve their understanding.

Reasonable Doubt – An Unrealistic Psychological Criterion

The final piece of information obtained from the study is that subjects have difficulty making judgments using the standard legal criterion for guilt in criminal cases - reasonable doubt. Pilot data indicated that subjects needed to be sure of the evidence in order to convict someone beyond a reasonable doubt (estimates over 87-90). However, in the main study, most subjects’ answers fell within the range of 77-85 regardless of case type or manipulation. It seems subjects have a hard time judging information with the strict criteria indicated for reasonable doubt. Rather, jurors have lowered criteria when making judgments in relation to evidence in a trial.

Some researchers have hypothesized that “reasonable doubt” is too stringent of a criteria for assessing evidence in a court trial (Arkes & Mellers, 2002). Based on this study, it appears that this may be true. Jurors, regardless of case type, assess guilt within a narrow range of possibilities. Specifically, all jurors rated the evidence to be with a range of 77-85. Thus, it seems that jurors are reluctant to make estimates regarding evidence for a court cases using the levels that were previously indicated (from the pilot study) for guilt “beyond a reasonable doubt.” Although reasonable doubt is intended to be (ideally) a strict legal tool, it seems to be more of an inaccurate subjective measure in which jurors are ill-equipped to use. Rather, jurors appear to base their judgments on a lower level than that which reasonable doubt was originally established. This lower level has many legal ramification, such as increasing the number of wrongly convicted individuals (those innocent, yet still found guilty), but also the possibility for the reducing of wrongly acquitted individuals (those guilty, yet still found innocent).

Additional Considerations

Thus, does this mean that one will always fall prey to the *confusion of the inverse*? No. As was the situation with the case of arson, jurors do not always commit the *confusion of the inverse*. None of the manipulations worked and yet, the case of arson was overcome. One explanation for this occurrence could be due to the numbers used in the court case as was mentioned in the discussion of study two. However, an alternate explanation exists.

Of the four cases, only arson was different and only arson is different[9]. Of the cases, murder, sexual assault, and kidnapping are all personal crimes; whereas arson is a property crime. Thus, in this case the jurors may need a certain context in which to eliminate the confusion of the inverse. Therefore, an unintentional manipulation did work, context.

Additionally, some would argue the use of Bayes to achieve the correct answer is not the best way. However, whether one uses Bayes, Signal Detection Theory or any other method, the correct answer is not the focus. Rather, the focus should be on the fact that people equate conditional probabilities. Irregardless of the method to determine the *confusion of the inverse*, one should never commit the *confusion of the inverse*. By confusing conditional probabilities, a person can never get to higher order effects of a problem and will therefore always make a mistake.

Conclusions

The main question behind the dissertation - do jurors commit the *confusion of the inverse*? – led to an unequivocal, yes! However, though the confusion of the conditional probabilities $P(H|D)$ and $P(D|H)$ occurs, subject tend to ignore all statistical information when making their judgments. Rather, subjects use their subjective assessment of the evidence strength when making a verdict.

However, verdicts consistent with Bayes did occur for those who made correct inferences about the conditional probabilities. Therefore, the reduction/elimination of the

[9] The author would like to thank Anne Pigenot, fellow graduate student, for this insightful explanation.

confusion of the inverse should lead to jurors using evidence correctly rather than using their subjective assessment in a trial.

Limitations and Future Research

Limitations

There are several possible limitations to the current project. The first has to do with the type of trials used. As previously indicated, only criminal trials were used. However, the use of civil trials may be of some benefit. The criteria for judgment in criminal trials is reasonable doubt (a very strict criterion), but in civil trials it is preponderance of evidence (a much lower criterion). These lowered criteria for civil trials may lead to different judgments and more success using different formats, targets, and reference classes.

Second, although ecological validity was maintained through the use of court cases, the use of students does not correspond with ecological validity. However, the student subjects used in the small Midwestern College had never participated in a study before this one, so they were not exposed to the research environment that occurs at many large state schools. Thus, the procedure used (i.e., the swearing-in and cover story) was far more believable to these students there by increasing the likelihood they took the study seriously and increasing ecological validity.

The third limitation has to do with a design issue. The cases where presented in the same order in both Study One and Two. The case presentation could have been counter-balanced to ensure that no one case was leading subjects to respond in a particular manner. However, this was unfeasible for the second study due to the number of groups being examined (8). Counter-balancing would have only led to many more groups and larger numbers of subjects, which would have been difficult to obtain.

The fourth limitation surrounds the cases used. The four court cases where chosen after months of looking through court cases at the local and federal court houses in Sioux City, IA. The crime types where chosen because of the Pilot data results and due to convenience. Other crimes, such as one which are not sensitive to the public (such as the case of sexual assault in which nearly all jurors rated the defendant as guilty, even though the evidence did not support that verdict) could have been better for this type of project, possibly leading to the reduction of personal belief in evidence as the predictor of juror verdict.

Finally, the last limitation has to do with subjects in the study. Subject in the study were young college students in a mid-size Midwestern town. Nearly all where white and had no prior experience with criminal trials. A more representative population (older subjects, those exposed to criminal trials, and those of different ethnic background) could have been used to better understand jurors' use of conditional probabilities and evidence.

Future Research

Future research should continue to eliminate the *confusion of the inverse*. As mentioned previously, not all of the court cases resulted in the confusion being committed. The case of arson brought up an interesting point about the potential for case context having an effect on the *confusion of the inverse*. Therefore, future research should look at the different types of crimes (crimes against people, property, and yourself) that can be committed and determine if the context has an effect in the way jurors interpret information.

In addition, future research should address order effects. The limitations section has already eluded to the order effects that might have occurred due to the same cases being presented in the same order (kidnapping, arson, murder and sexual assault). However, additional research should also examine the order effects in the way the evidence is presented. Typical studies on the *confusion of the inverse* present information in a particular order: baserate information followed by a reliability statement. This approach assumes people process information in this linear form. The question arises, what if this form is changed, will the *confusion of the inverse* dissipate?

Moreover, the importance of ecological validity has been well documented throughout. Yet, with the scarcity of usable court cases, the augmentation of these cases may need to occur. Thus, future research should explore changing these cases to meet the needs of future study. However, in order to use augmented cases for future study, they have to be ecologically valid. Therefore, new cases derived from the current ones would have be checked by those in the legal profession to ensure their validity to real court proceedings.

Furthermore, the work being done on jurors and their inability to use statistical evidence should be furthered. A number of reasons have been hypothesized, but none explored deeply. Therefore, research should look at quantitative as well as qualitative data to understand why jurors tend to ignore statistical information and how they reach a decision based on their own personal belief in the strength of the evidence.

Although this study used statistical information to evaluate the *confusion of the inverse*, this does not have to be the case. Some researchers have said the *confusion of the inverse* may not exist if one use qualitative statements such as "there is a strong likelihood" of X occurring versus saying there is a "90% chance" of X occurring. This qualitative way of phrasing information may cause jurors to understand the evidence better and avoid confusing conditional probabilities. All of these future research plans are part of a line of research being developed by the researcher and his undergraduate students at his new institution.

Finally, future research should also address the four previously mentioned limitations. The first step in future research is to understand why people commit the *confusion* of the inverse. The primary focus should be using civil trials in which the criteria is lowered for verdict – preponderance of evidence. The same manipulation used in this study could very well lead to different responses and a clear answer to why people commit the *confusion of the inverse*. Once this is understood, research could be conducted to eliminate it as was mentioned previously in this paper.

Additionally, future research should look at different populations. Specifically, a population more representative of the American public and not such a homogeneous sample as was the case in this study. This could be achieved by going to court houses and requesting recent juror lists. Then, researchers could contact subjects and have them complete similar surveys as was done in this study.

References

Arkes, H. & Mellers, B. (2002). Do juries meet our expectations? *Law and Human Behavior 26,* 625-639.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Birnbaum, M. & Mellers, B. A. (1983). Bayesian inference: combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45*, 792-804.

Casscells, W., Schoenberger, T., & Grayboys, N. (1978). *New England Journal of Medicine, 299*, 999.

Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977).

Christensen-Szalanski, J.J.J. & Beach, L.R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance, 29*, 270-278.

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

Connors, E., Lundregan, T., Miller, N., & McEwen, T. (1996). *Case Studies in Use of DNA Evidence Series: NIJ Research Report*. Technical Report; available online at http://www.psychology.iastate.edu/faculty/gwells/dnanijreport.htm. July, 2004.

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Dawes, R.M. (1986). Representative thinking in clinical judgment. *Clinical Psychology Review, 6*, 425-441.

Dawes, R.M., Mirels, H.L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science, 4*, 396-400.

Egan, R.L. (1972). *Mammography (2nd Edition)*. Springfield, IL: Charles C. Thomas.

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

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

Faigman, D.L. & Baglioni, A.J. (1988). Bayes theorem in the trial process: Instructing jurors on the value of statistical evidence. *Law and Human Behavior, 12*, 1-17.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50*,123-129.

Feinberg, S.E. (1989). *The evolving role of statistical assessment as evidence in the courtroom.* New York: Springer-Verlag.

Finkelstein. M.. & Fairley, W. B. (197Q) A Bayesian approach to identification evidence. *Havard Law Review, 83*, 489-517.

Fischoff, B. & Bar-Hillel, M. (1984). Focusing techniques: A shortcut to improving probability judgments? *Organizational Behavior and Human Performance, 34,* 175-194.

Fischoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23,* 339-359.

Gavanski I. & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology, 63,* 766-780.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506-528.

Gigerenzer. G (1996). The psychology of good judgment: Frequency formats and simple algorithms. *Journal of Medical Decision-Making, 16*, 273-280.

Gigerenzer, G. (2000). *Adaptive thinking.* New York: Oxford University Press.

Gigerenzer, G. (2002). In the year 2054: Innumeracy defeated. In P. Seldmeir & T. Betsch (Eds.) *ETC. Frequency processing and cognition*. New York: Oxford University Press. Pp. 55-66.

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

Griggs v. Duke Power Co., 401 U.S. 424, 91 S. Ct. 849,28 L. Ed. 2nd 158 (1971).

Hamm, R. (1987). Diagnostic inference: People's use of information in incomplete Bayesian word problems. Institute of Cognitive Science Publication No. 87-11, University of Colorado, Boulder.

Hamm, R. (1993). Explanations for common responses to the blue/green cab probabilistic inference word problem. *Psychological Reports, 72*, 219-242.

Hanita, M., Gavanski I., & Fazio, R.H. (1997). Influencing probability judgments by manipulating the accessibility of sample spaces. *Personality and Social Psychology Bulletin, 23,* 801-813.

Hastie, R. (Ed.) (1993). *Inside the juror: The psychology of decision-making.* New York: Cambridge University Press.

Hazelwood School District v. U.S., 433 U.S. 301 (1977).

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision-Making,* 12,275-305.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2001). Communication of statistical information, *Science, 290*, 2261-2262.

Kahneman D. & Tversky, A. (1972). On prediction and judgment. *Oregon Research Institute Research Monograph, 12(4).*

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

Koehler, J.J. (1996). The base-rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19*, 1-53.

Koehler, J.J. & Macchi, L. (in press). Thinking about low probability events: A exemplar cueing theory. *Psychological Science*.

Lichtenstein, S. & McGregor, D. (1984). Structuring as an aid to performance in base-rate problems. Report No. 84-16, Decision Research, Eugene, OR.

Linedecker, C. (1995). *O.J. A to Z: The complete handbook to the trial of the century*. New York: St. Martin's Griffin.

Lusted, L.B. (1968). *Introduction to medical decision-making.* Springfield, IL.: Charles C. Thomas.

Lyon, D. & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica, 40*, 287-298.

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Quarterly Journal of Experimental Psychology, 48A,* 188-207.

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristic and frequency format explanations. *Organizational Behavior and Human Decision Processes, 82*, 217-236.

McGee, G. (1979, February 6). Breast surgery before cancer. Ann Arbor News, Section B, B-1.

Meehl, P. & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs of patterns, or cutting scores. *Psychological Review, 106*, 210-214.

Niedermeier, K.E., Kerr, N.L., & Messe, L.A. (1999). Jurors' use of naked statistical evidence: Exploring bases and implications of the Wells effect. *Journal of Personality and Social Psychology, 76,* 533-542.

Niniluotto, I. (1981). L.J. Cohen versus Bayesianism. *Behavioral and Brain Sciences, 4*, 349.

Nisbett, R. (1993), *Rules for Reasoning*, Mahwah, NJ: Lawrence Erlbaum

Nisbett, R., Krantz, D., Jepson, C, & Kunda, Z. (1984). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90(4),* 339-363.

Nisbett, R., Krantz, D., Jepson, C, & Kunda, Z. (2002). The use of statistical heuristics in everyday reasoning. In T. Gilovich & D. Griffin (Eds.) *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press. Pp.510-533.

Pennington, N. & Hastie, R, (1981). Juror decision-making models: The generalization gap. *Psychological Bulletin, 89,* 246-287.

Pennington, N. & Hastie, R. (2000). Explanation-based decision making. In Connolly, T., Arkes, H., & Hammond, K. (Eds.) *Judgment and Decision Making: An Interdisciplinary Reader* (pp. 212-228). New York: Cambridge University Press.

Pollatsek, A., Well, A.D., Konold, C., Hardiman, P., Kobb G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40*, 255-269.

Rogosa, D. R. & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20(4)*, 335-343.

Saks, M. J. & Kidd. R. (1981). Human information processing and adjudication: Trial by heuristics*.* *Law and Society Review, 5*,123-160.

Shanteau, J. & Weiss, D. (2003). Empirical assessment of expertise. *Human Factors, 45,* 104-114.

Sherman, S.J., McMullen, M.N., & Gavanski, I. (1992). Natural sample spaces and the inversion of conditional judgments. *Journal of Experimental Social Psychology, 28*, 401-421.

Slovic, P., Lichtenstein, S., Edwards, W. (1965). Boredom induced changes in preferences among bets. *American Journal of Psychology, 78*, 208-217.

Stutts, J.C., Reinfurt, D.W., Staplin, L., Rodgman, E.A. (2001). *The Role of Driver Distraction in Traffic Crashes*. Technical Report; available online at www.aaafoundation.org, May 2001.

Thompson, W.C. & Schumann, E.L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior, 11,* 167-187.

Tribe. L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review, 84,* 1328-1393.

Tversky, A. (1981). L.J. Cohen , again: On the evaluation of inductive intuitions. *Behavioral and Brain Sciences, 4*, 354-356.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Tversky, A. & Kahneman, D. (1982). Evidential impact of base-rates. In D. Kahneman, P. Slovic and A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press. Pp. 153-160.

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician, 57(2)*, 74- 79.

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

Wason, P. (1983). Realism and rationality in the selection task. In J. St. B. T. Evans (Ed.). *Thinking and reasoning: Psychological approaches* (pp. 44-75). Boston: Rutledge & Kegan Paul.

Weinstein, J. B., Mansfield, J. H.. Abrams, N.. & Berger, M.A. (1983). *Cases and materials on evidence (7th ed.).* Minneola, New York: The Foundation Press.

Wells. G L (1992) Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology, 62*, 739-752.

Wells G. & Luus, E. (1990). The diagnosticity of a lineup should not be confused with the diagnosticity value of non-lineup evidence. *Journal of Applied Psychology, 75(5),* 511-516.

Wolfe, C.R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy-trace theory account. *Journal of Behavioral Decision-Making, 8*, 85-108.

Wright, E.F., MacEacher, L., Stoffer, E., & MacDonald, N. (1996). Factors affecting the use of naked statistical evidence liability. *The Journal of Social Psychology, 136(6),* 677-688.

Zabell, S. (1993). A mathematician comments on models of juror decision-making. In R. Hastie (Ed.) *Inside the Juror.* New York: Cambridge University Press. Pp. 263-269.

Appendices

Appendix A

Pilot Study Survey

Instructions:
Below you will find several legal definitions. Please read each definition carefully and answer the questions below. All of your answers are anonymous and will remain confidential.

Reasonable Doubt

Definition A:

The level of certainty a juror must have to find a defendant guilty of a crime. A real doubt, based upon reason and common sense after careful and impartial consideration of all the evidence, or lack of evidence, in a case.

Proof beyond a reasonable doubt, therefore, is proof of such a convincing character that you would be willing to rely and act upon it without hesitation in the most important of your own affairs. However, it does not mean an absolute certainty.

Definition B:

An accused person is entitled to acquittal if , in the minds of the jury, his guilt has not been proved beyond a "reasonable doubt" that state of the minds of jurors in which they cannot say they feel and abiding conviction as to the truth of the charge.

1: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of a crime beyond a reasonable doubt? _____________

2: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of a crime beyond a reasonable doubt? ____________

Next, you will be instructed to determine reasonable doubt for specific crimes:

3 On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of murder beyond a reasonable doubt?

____

4: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of murder beyond a reasonable doubt?

____

5 On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of sexual assault beyond a reasonable doubt?

____

6: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of sexual assault beyond a reasonable doubt? _____________

7 On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of burglary beyond a reasonable doubt?

____

8: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of burglary beyond a reasonable doubt?

____

9 On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of assault beyond a reasonable doubt?

____

10: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of assault beyond a reasonable doubt?

____

11 On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of manslaughter beyond a reasonable doubt?

____

12: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of manslaughter beyond a reasonable doubt? ___________

13 On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of kidnapping beyond a reasonable doubt?

____

14: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of kidnapping beyond a reasonable doubt? _____________

15 On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **guilty** of arson beyond a reasonable doubt?

____

16: On a scale of 0-100 (0 = not at all sure, 100 = completely sure), how sure would you have to be to find someone **not guilty** of arson beyond a reasonable doubt?

____

Gender: ____________________

Age: ______________________

Have you ever been called for jury duty, apart of a criminal trial or studied legal issues before? (circle one)

Yes No

Appendix B
Study One Court Descriptions

Case #1

Facts of the trial:
In the early morning hours of August 13, 1994, a woman contacted police and informed then she had been kidnapped by 3 males. The woman testified she was at a bar with a friend when she met 3 men. She danced with a one of the men once and he bought her a beer. The bar was closing and a friend of the woman's announced there was a party and all were invited. The 3 men were standing nearby and the woman's friend spoke with them. The woman assumed the 3 men were going to the party and accepted a ride with them. She climbed into the back seat of the vehicle with the man whom she danced with, while the other two men sat in the front seat. One of the men said, "We are going to a party."

The four did not go to the party. The woman testified that, after driving around, they went to a gas station. She became frightened because one of the men was making sexual advances and she asked to be let out of the car. She testified that the man who she danced with placed her in a head lock and would not let her out. The car left the gas station and the four drove around. She struggled and kicked. One of the men in the front seat grabbed her legs and the three men began yelling at her. During the struggle, the man whom she danced with cut himself and bled on her shirt.

Eventually they parked in a remote and overgrown area. The woman asked to leave the car so she could go to the bathroom. The man whom she danced with escorted her to some weeds, at which time the woman attempted to run away. She ran into a fence and eventually climbed over it and escaped.

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a large city 15% of people will have DNA that matches the blood found on the shirt.

The pathologist is testified that they are right about 90% of the time when indicating someone's DNA will match with evidence of this type and right about 99% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #2

Facts of the trial:
In the late evening hours of February 26, 1993, two men, helped Mr. Smith push his car out of the snow. Mr. Smith and his girlfriend invited the two helpers back to their apartment for drinks. They drank until the early morning hours.

The events next were related to the court by Mr. Smith, who testified at the trial. After Mr. Smith and his girlfriend retired to the bedroom, the woman decided to return to the living room to tell the two men to leave. Mr. Smith followed his girlfriend a few minutes later and found the two men molesting her. The two men beat Mr. Smith and tied his and his girlfriend's hands behind their back with electrical cord. As the two men were leaving, they set the apartment on fire.

Police arrived and were able to save the two victims. Police arrested two suspects based on the events related to them by Mr. Smith. During their investigation, police recovered a partial fingerprint from a bottle of lighter fluid.

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a large city, 12% of people's fingerprints will match those found at the scene.

The pathologist also testified that they are right about 70% of the time when indicating someone's fingerprints will match with evidence of this type and they are right about 96% of the time when indicating that someone's fingerprints do not match with evidence of this type.

Case #3

Facts of the Trial:
In December of 2003 a man, accompanied by his woman friend and two young children, arrived in a large city homeless and destitute. Through Traveler's Aid they were given shelter in the home of Mr. Smith. The people stayed with Mr. Smith and his wife, for just over a week but were asked to leave when Mr. Smith discovered that the man had drug paraphernalia. However, the man continued to receive aid from the Smiths in the form of money and transportation. Eventually, the Smiths began to feel they were being used and withdrew all support. The man resented the discontinuance of aid.

On February 27th, 2004, Mr. Smith returned home from a night class and found his wife, dead on the living room floor. An autopsy showed that she had been strangled and stabbed repeatedly in the throat. Found in the house was a baseball cap.

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a large city 26% of people will have DNA that matches the hairs found in the cap.

The pathologist is right about 80% of the time when indicating someone's DNA will match with evidence of this type and right about 97% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #4

Facts of the Trial:
A young woman testified that on the evening of July 18th, 1997, she was walking home when a man forced her into a car and threatened to kill her if she "screamed or anything." The woman testified that the man used a choke hold to get her into the car. He then drove her to an empty apartment, told her to undress, engaged in various sex acts, which included oral and anal sex.

The woman promptly reported the incident to her brother after the man drove her from the apartment to a point near her home. Police were immediately called and the woman examined at a hospital. The doctor also determined the woman had sustained a recent laceration of her anus. During the examination, semen from the man was recovered.

Police arrested a suspects based on the events related to them by the woman. During their investigation, police recovered a partial fingerprint from a bottle of liquor.

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a large city, 9% of male's will match the semen found at the scene.

The pathologist also testified that they are right about 80% of the time when indicating someone's semen will match with evidence of this type and they are right about 95% of the time when indicating that someone's semen does not match with evidence of this type.

Appendix C
Juror Instructions for Each Case

Jury Instructions – Case #1

Please Read:
The trial information in this case charges the defendant with the crime of Kidnapping in the First Degree.

The defendant has entered a plea of not guilty. A plea of not guilty is a complete denial of the charges and places the burden on the State to prove guilt beyond a reasonable doubt.

A reasonable doubt is one that fairly and naturally arises from the evidence or lack of evidence produced by the state.

If, after a full and fair consideration of all the evidence, you are firmly convinced of the defendant's guilt, then you have no reasonable doubt and you should find the defendant guilty.

But, if, after a full and fair consideration of all the evidence or lack of evidence produced by the State, you are not firmly convinced of the defendant's guilt, then you have a reasonable doubt and you should find the defendant not guilty.

The State must prove all of the following elements of Kidnapping in the First Degree:

1. On or about the 13th day of August, the defendant removed the woman from one place to another.
2. The defendant knew he did not have the consent of the woman to do so.

If the State has proved all of these elements, the defendant is guilty of Kidnapping in the First Degree. If the State has failed to prove any of these elements, then the defendant is not guilty of Kidnapping in the First Degree.

Juror Questions:

1. How strong do you believe the evidence is against the defendant?
   (1= not at all strong, 7= extremely strong)

   1-----------2------------3-----------4-----------5-----------6-----------7

2. What is the probability that the suspect's DNA is the same as the blood found at the scene? Indicate a number between 0 (not at all) – 100 (completely sure).______.

3. What is the probability that the suspect kidnapped the woman?
   Indicate a number between 0 (not at all) – 100 (completely sure).____________.

4. Based on the evidence, is the person guilty or not guilty beyond a reasonable doubt?
   ____________________.

Jury Instructions – Case #2

Please Read:
The trial information in this case charges the defendant with the crime of Arson in the First Degree.

The defendant has entered a plea of not guilty. A plea of not guilty is a complete denial of the charges and places the burden on the State to prove guilt beyond a reasonable doubt.

A reasonable doubt is one that fairly and naturally arises from the evidence or lack of evidence produced by the state.

If, after a full and fair consideration of all the evidence, you are firmly convinced of the defendant's guilt, then you have no reasonable doubt and you should find the defendant guilty.

But, if, after a full and fair consideration of all the evidence or lack of evidence produced by the State, you are not firmly convinced of the defendant's guilt, then you have a reasonable doubt and you should find the defendant not guilty.

The State must prove all of the following elements of Arson in the First Degree:

1. On or about the 26th day of February, the defendant caused a fire or placed combustible material in or near property.
2. The defendant intended to destroy or damage the property or knew the property would probably be destroyed or damaged
3. The presence of a person in the property could have been reasonably anticipated.

If the State has proved all of these elements, the defendant is guilty of Arson in the First Degree. If the State has failed to prove any of these elements, then the defendant is not guilty of Arson in the First Degree.

Juror Questions:

1. How strong do you believe the evidence is against the defendant?
   (1= not at all strong, 7= extremely strong)

   1-----------2------------3-----------4-----------5-----------6-----------7

2. What is the probability that the suspect's fingerprints is the same as the fingerprint found at the scene? Indicate a number between 0 (not at all) – 100 (completely sure).__________ _______.

3. What is the probability that the suspect committed arson?
   Indicate a number between 0 (not at all) – 100 (completely sure)._______________.

4. Based on the evidence, is the person guilty or not guilty beyond a reasonable doubt?
   _____________.

Jury Instructions – Case #3

Please Read:
The trial information in this case charges the defendant with the crime of Murder in the First Degree.

The defendant has entered a plea of not guilty. A plea of not guilty is a complete denial of the charges and places the burden on the State to prove guilt beyond a reasonable doubt.

A reasonable doubt is one that fairly and naturally arises from the evidence or lack of evidence produced by the state.

If, after a full and fair consideration of all the evidence, you are firmly convinced of the defendant's guilt, then you have no reasonable doubt and you should find the defendant guilty.

But, if, after a full and fair consideration of all the evidence or lack of evidence produced by the State, you are not firmly convinced of the defendant's guilt, then you have a reasonable doubt and you should find the defendant not guilty.

The State must prove all of the following elements of Murder in the First Degree:

1. On or about the 26th day of February, the defendant strangled and stabbed Mrs. Smith.
2. Mrs. Smith died as a result of being strangled and stabbed.
3. The defendant acted with malice forethought.
4. The defendant acted willfully, deliberately, premeditatedly and with the specific intent to kill Mrs. Smith.

If the State has proved all of these elements, the defendant is guilty of Murder in the First Degree. If the State has failed to prove any of these elements, then the defendant is not guilty of Murder in the First Degree.

Juror Questions:

1. How strong do you believe the evidence is against the defendant?
(1= not at all strong, 7= extremely strong)

1----------2-----------3----------4----------5----------6----------7

2. What is the probability that the suspect's hair is the same as the hair found at the scene? Indicate a number between 0 (not at all) – 100 (completely sure).______.

3. What is the probability that the suspect committed murder?
Indicate a number between 0 (not at all) – 100 (completely sure).________________.

4. Based on the evidence, is the person guilty or not guilty beyond a reasonable doubt?
______________.

Jury Instructions – Case #4

Please Read:
The trial information in this case charges the defendant with the crime of Sexual Abuse in the First Degree.

The defendant has entered a plea of not guilty. A plea of not guilty is a complete denial of the charges and places the burden on the State to prove guilt beyond a reasonable doubt.

A reasonable doubt is one that fairly and naturally arises from the evidence or lack of evidence produced by the state.

If, after a full and fair consideration of all the evidence, you are firmly convinced of the defendant's guilt, then you have no reasonable doubt and you should find the defendant guilty.

But, if, after a full and fair consideration of all the evidence or lack of evidence produced by the State, you are not firmly convinced of the defendant's guilt, then you have a reasonable doubt and you should find the defendant not guilty.

The State must prove all of the following elements of Sexual Abuse in the First Degree:

1. On or about the 18^{h} day of July, the defendant performed a sex act with the woman.
2. The defendant performed the sex act by force or against the will of the woman.
3. During the commission of sexual abuse, the defendant caused the woman a serious injury.

If the State has proved all of these elements, the defendant is guilty of Sexual Abuse in the First Degree. If the State has failed to prove any of these elements, then the defendant is not guilty of Sexual Abuse in the First Degree.

Juror Questions:

1. How strong do you believe the evidence is against the defendant?
(1= not at all strong, 7= extremely strong)

1-----------2------------3-----------4-----------5-----------6-----------7

2. What is the probability that the suspect's semen is the same as the semen found at the scene? Indicate a number between 0 (not at all) – 100 (completely sure).______.

3. What is the probability that the suspect committed sexual abuse?
Indicate a number between 0 (not at all) – 100 (completely sure).____________.

4. Based on the evidence, is the person guilty or not guilty beyond a reasonable doubt?
______________.

Appendix D
Participant Instructions and Juror Swearing-In Procedure

Thank you for all participating. Today, you will sworn-in as acting jurors for a Federal Court. You will be reading several court cases with evidence presented at a trial. As a juror, you are instructed to read each case and the corresponding jury instructions for each charge. Once you have completed reading each case, you will be asked 4 questions about the case. Your answers will be used to decide if the cases are suitable enough to go to trial or if they are not suitable. Please answer honestly and within the description of the instructions. Thank you again for participating.

Please respond "I do" to the following question and sign the first sheet of paper.

Do each of you swear or affirm that you will well and truly try he matter in issue between the parties, and give a true verdict according to the law and evidence?

Appendix E
Demographic Questions at the End of the Packet
Demographic Information

1. Age:_______________

2. Gender (circle one):   Male   Female

3. Have you ever been called for jury duty? (circle one)

   Yes   No

4. Have you ever been apart of a criminal trial? (circle one)\
   Yes   No

5. Have you ever studied legal issues before? (circle one)
   Yes   No

6. What is your race? (Check One)

   _______ Caucasian
   _______ Latino/Hispanic
   _______ Black
   _______ Asian
   _______ Native American
   _______ Other

7. Does the statement, the probability that the defendant matches the evidence, mean the same thing as the statement, the probability that the evidence matches the defendant? (circle one)
   Yes   No

8. What do these statements mean in your own words?

____
____
____
____
____
____
____
____
____

Appendix F
Calculations for Bayesian Solutions to Each Court Case

<u>Bayes Formula:</u>

$$P(\mathrm{H}|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{H})*P(\mathrm{H})}{P(\mathrm{D}|\mathrm{H})*P(\mathrm{H}) + P(\mathrm{D}|{\sim}\mathrm{H})*P({\sim}\mathrm{H})}$$

Kidnapping:

$$P(\mathrm{H}|\mathrm{D}) = \frac{(.90)(.15)}{(.90)(.15) + (.01)(.85)} = \underline{\mathbf{.941}}$$

Arson:

$$P(\mathrm{H}|\mathrm{D}) = \frac{(.70)(.12)}{(.70)(.12) + (.03)(.83)} = \underline{\mathbf{.761}}$$

Murder:

$$P(\mathrm{H}|\mathrm{D}) = \frac{(.80)(.26)}{(.80)(.26) + (.03)(.74)} = \underline{\mathbf{.904}}$$

Sexual Assault:

$$P(\mathrm{H}|\mathrm{D}) = \frac{(.80)(.09)}{(.80)(.09) + (.05)(.91)} = \underline{\mathbf{.613}}$$

Appendix G
All Possible Combinations for the 2x2x3 Design for Each Case

Case #1 – **Multiple, Frequency, Large**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a large city of 5,000,000, 15 in 100 people in a town who are not the source would nonetheless match the blood found on the shirt.

The pathologist also testified that they are right about 90 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 99 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #2– **Multiple, Frequency, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a large city of 5,000,000, 12 in 100 people in a town who are not the source would nonetheless match the fingerprints found at the scene.

The pathologist also testified that they are right about 70 out of 100 times when indicating someone's fingerprints will match with evidence of this type and they are right about 96 out of 100 times when indicating that someone's fingerprints do not match with evidence of this type.

Case #3– **Multiple, Frequency, Large**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a large city of 5,000,000, 26 in 100 people in a town who are not the source would nonetheless match hair found in the cap.

The pathologist is right about 80 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 97 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #4– **Multiple, Frequency, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a large city of 5,000,000, 9 in 100 people in a town who are not the source would nonetheless match the semen found at the scene.

The pathologist also testified that they are right about 80 out of 100 times when indicating someone's semen will match with evidence of this type and they are right about 95 out of 100 times when indicating that someone's semen does not match with evidence of this type.

Case #1 – **Multiple, Frequency, Small**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a small city of 500, 15 in 100 people in a town who are not the source would nonetheless match the blood found on the shirt.

The pathologist also testified that they are right about 90 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 99 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #2– **Multiple, Frequency, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a small city of 500, 12 in 100 people in a town who are not the source would nonetheless match the fingerprints found at the scene.

The pathologist also testified that they are right about 70 out of 100 times when indicating someone's fingerprints will match with evidence of this type and they are right about 96 out of 100 times when indicating that someone's fingerprints do not match with evidence of this type.

Case #3– **Multiple, Frequency, Small**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a small city of 500, 26 in 100 people in a town who are not the source would nonetheless match hair found in the cap.

The pathologist is right about 80 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 97 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #4– **Multiple, Frequency, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a small city of 500, 9 in 100 people in a town who are not the source would nonetheless match the semen found at the scene.

The pathologist also testified that they are right about 80 out of 100 times when indicating someone's semen will match with evidence of this type and they are right about 95 out of 100 times when indicating that someone's semen does not match with evidence of this type.

Case #1 – **Multiple, Probability, Large**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a large city of 5,000,000, 15% of the people in a town who are not the source would nonetheless match the blood found on the shirt.

The pathologist is testified that they are right about 90% of the time when indicating someone's DNA will match with evidence of this type and right about 99% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #2– **Multiple, Probability, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a large city of 5,000,000, 12% of the people in a town who are not the source would nonetheless match the fingerprints found at the scene.

The pathologist also testified that they are right about 70% of the time when indicating someone's fingerprints will match with evidence of this type and they are right about 96% of the time when indicating that someone's fingerprints do not match with evidence of this type.

Case #3– **Multiple, Probability, Large**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a large city of 5,000,000, 26% of the people in a town who are not the source would nonetheless match hair found in the cap.

The pathologist is right about 80% of the time when indicating someone's DNA will match with evidence of this type and right about 97% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #4– **Multiple, Probability, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a large city of 5,000,000, 9% of the people in a town who are not the source would nonetheless match the semen found at the scene.

The pathologist also testified that they are right about 80% of the time when indicating someone's semen will match with evidence of this type and they are right about 95% of the time when indicating that someone's semen does not match with evidence of this type.

Case #1 **– Multiple, Probability, Small**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a small city of 500, 15% of the people in the town who are not the source would nonetheless match the blood found on the shirt.

The pathologist is testified that they are right about 90% of the time when indicating someone's DNA will match with evidence of this type and right about 99% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #2**– Multiple, Probability, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a small city of 500, 12% of the people in the town who are not the source would nonetheless match the fingerprints found at the scene.

The pathologist also testified that they are right about 70% of the time when indicating someone's fingerprints will match with evidence of this type and they are right about 96% of the time when indicating that someone's fingerprints do not match with evidence of this type.

Case #3**– Multiple, Probability, Small**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a small city of 500, 26% of the people in the town who are not the source would nonetheless match hair found in the cap.

The pathologist is right about 80% of the time when indicating someone's DNA will match with evidence of this type and right about 97% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #4**– Multiple, Probability, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a small city of 500, 9% of the people in the town who are not the source would nonetheless match the semen found at the scene.

The pathologist also testified that they are right about 80% of the time when indicating someone's semen will match with evidence of this type and they are right about 95% of the time when indicating that someone's semen does not match with evidence of this type.

Case #1 **– Single, Frequency, Large**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a large city of 5,000,000, the chance the suspect would match the blood found on the shirt if he were not the source is 15 in 100.

The pathologist also testified that they are right about 90 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 99 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #2**– Single, Frequency, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a large city of 5,000,000, the chance the suspect would match the fingerprints found at the scene if he were not the source is 12 in 100.

The pathologist also testified that they are right about 70 out of 100 times when indicating someone's fingerprints will match with evidence of this type and they are right about 96 out of 100 times when indicating that someone's fingerprints do not match with evidence of this type.

Case #3**– Single, Frequency, Large**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a large city of 5,000,000, the chance the suspect would match the hair found in the cap if he were not the source is 26 in 100.

The pathologist is right about 80 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 97 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #4**– Single, Frequency, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a large city of 5,000,000, the chance the suspect would match the semen found at the scene if he were not the source is 9 in 100.

The pathologist also testified that they are right about 80 out of 100 times when indicating someone's semen will match with evidence of this type and they are right about 95 out of 100 times when indicating that someone's semen does not match with evidence of this type.

Case #1 **– Single, Frequency, Small**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a small city of 500, the chance the suspect would match the blood found on the shirt if he were not the source is 15 in 100.

The pathologist also testified that they are right about 90 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 99 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #2**– Single, Frequency, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a small city of 500, the chance the suspect would match the fingerprints found at the scene if he were not the source is 12 in 100.

The pathologist also testified that they are right about 70 out of 100 times when indicating someone's fingerprints will match with evidence of this type and they are right about 96 out of 100 times when indicating that someone's fingerprints do not match with evidence of this type.

Case #3**– Single, Frequency, Small**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a small city of 500, the chance the suspect would match the hair found in the cap if he were not the source is 26 in 100.

The pathologist is right about 80 out of 100 times when indicating someone's DNA will match with evidence of this type and right about 97 out of 100 times when indicating that someone's DNA does not match with evidence of this type.

Case #4**– Single, Frequency, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a small city of 500, the chance the suspect would match the semen found at the scene if he were not the source is 9 in 100.

The pathologist also testified that they are right about 80 out of 100 times when indicating someone's semen will match with evidence of this type and they are right about 95 out of 100 times when indicating that someone's semen does not match with evidence of this type.

Case #1 – **Single, Probability, Large**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a large city of 5,000,000 the chance the suspect would have matched the blood found on the shirt if he were not the source is 15%.

The pathologist is testified that they are right about 90% of the time when indicating someone's DNA will match with evidence of this type and right about 99% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #2– **Single, Probability, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a large city of 5,000,000 the chance the suspect would have matched the fingerprints found at the scene if he were not the source is 12%.

The pathologist also testified that they are right about 70% of the time when indicating someone's fingerprints will match with evidence of this type and they are right about 96% of the time when indicating that someone's fingerprints do not match with evidence of this type.

Case #3– **Single, Probability, Large**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a large city of 5,000,000 the chance the suspect would have matched the hair found in the cap if he were not the source is 26%.

The pathologist is right about 80% of the time when indicating someone's DNA will match with evidence of this type and right about 97% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #4– **Single, Probability, Large**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a large city of 5,000,000 the chance the suspect would have matched the semen found at the scene if he were not the source is 9%.

The pathologist also testified that they are right about 80% of the time when indicating someone's semen will match with evidence of this type and they are right about 95% of the time when indicating that someone's semen does not match with evidence of this type.

Case #1 **– Single, Probability, Small**

Evidence at the Trial:
Police identified three suspects and arrested them. DNA from Suspect A was compared against blood evidence found on the woman's shirt. During a trial, a pathologist testified that the blood found on the shirt matched Suspect A. The pathologist testified that in a small city of 500, the chance the suspect would have matched the blood found on the shirt if he were not the source is 15%.

The pathologist is testified that they are right about 90% of the time when indicating someone's DNA will match with evidence of this type and right about 99% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #2**– Single, Probability, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the fingerprints found on the bottle of lighter fluid matched Suspect A. The pathologist testified that in a small city of 500, the chance the suspect would have matched the fingerprints found at the scene if he were not the source is 12%.

The pathologist also testified that they are right about 70% of the time when indicating someone's fingerprints will match with evidence of this type and they are right about 96% of the time when indicating that someone's fingerprints do not match with evidence of this type.

Case #3**– Single, Probability, Small**

Evidence at the Trial:
Police found the man and tested his hair DNA against hair DNA found in the baseball cap. During the trial, a pathologist testified that the hair found in the cap belonged to the man. The pathologist testified that in a small city of 500, the chance the suspect would have matched the hair found in the cap if he were not the source is 26%.

The pathologist is right about 80% of the time when indicating someone's DNA will match with evidence of this type and right about 97% of the time when indicating that someone's DNA does not match with evidence of this type.

Case #4**– Single, Probability, Small**

Evidence at the Trial:
During a trial, a pathologist testified that the semen found on the woman matched the suspect. The pathologist testified that in a small city of 500, the chance the suspect would have matched the semen found at the scene if he were not the source is 9%.

The pathologist also testified that they are right about 80% of the time when indicating someone's semen will match with evidence of this type and they are right about 95% of the time when indicating that someone's semen does not match with evidence of this type.

Table 1.

Pilot Study Means for Judgments of Verdict Beyond a Reasonable Doubt

	Verdict*	
Crime	Guilty	Not Guilty
Unspecified	90.3	85.4
Murder	93.5	91.2
Sexual Assault	88.0	85.9
Burglary	81.9	77.6
Assault	86.0	81.5
Manslaughter	93.7	90.4
Kidnapping	89.7	87.4
Arson	88.1	80.1

*Subject responded on a scale of 0 (not at all sure) -100 (complete sure).

Table 2.

Judgments of a Probability of a Match Compared to Bayesian Solutions

	Responses	
	Subject Estimates	Bayesian Solutions
Case 1 – Kidnapping	85.0_a	94_b
Case 2 – Arson	76.2_a	76_a
Case 3 – Murder	76.0_a	90_b
Case 4 – Sexual Assault	86.4_a	61_b

Note. Judgements were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at p < .001 in a one-sampled t-test.

Table 3.

Judgments of a Probability of a Match Compared to the Inverse Probability

	Responses	
	Subject Estimates	Inverse Probabilty
Case 1 – Kidnapping	85.0_a	90_a
Case 2 – Arson	76.2_a	70_b
Case 3 – Murder	76.0_a	80_a
Case 4 – Sexual Assault	86.4_a	80_b

Note. Judgements were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at p < .001 in a one-sampled t-test.

Table 4.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 1 (Kidnapping)

Variable	B	S.E.	β	t	p
Step 1					
Strength in Evidence	-.184	.033	-.479	-5.54	.000

Note: $R^2 = .23$.

Table 5.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 2 (Arson)

Variable	B	S.E.	β	t	p
Step 1					
Strength in Evidence	-.174	.031	-.482	-5.56	.000

Note: $R^2 = .23$.

Table 6.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 3 (Murder)

Variable	B	S.E.	β	t	p
Step 1					
Strength in Evidence	-.204	.023	-.654	-8.69	.000

Note: $R^2 = .43$.

Table 7.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 1 (Sexual Assault)

Variable	B	S.E.	β	t	p
Step 1					
Strength in Evidence	-.214	.023	-.683	-9.48	.000

Note: $R^2 = .47$.

Table 8.

Means for Difference Scores for Subjects' Estimates of a Probability of a Match and Bayesian Solutions and Inverse Probabilities

	Difference Scores*	
	Bayesian Solution	Inverse Probability
Case 1-Kidnapping	13.6	11.2
Case 2-Arson	14.1	16.0
Case3-Murder	14.9	12.3
Case 4-Sexual Assault	30.4	14.1

*Difference scores represented here are across all manipulations (i.e., the 8 different possible formulations)

Table 9.

Means for Difference Scores for Subjects' Estimates of a Probability of a Match and Bayesian Solutions

Case Type	Target	Format	Reference	Difference Score
Kidnapping				
	Single	Probability	Small	11.75
	Single	Probability	Large	9.97
	Single	Frequency	Small	18.16
	Single	Frequency	Large	13.92
	Multiple	Probability	Small	13.27
	Multiple	Probability	Large	10.83
	Multiple	Frequency	Small	9.54
	Multiple	Frequency	Large	21.53
Arson				
	Single	Probability	Small	11.04
	Single	Probability	Large	14.82
	Single	Frequency	Small	14.03
	Single	Frequency	Large	15.23
	Multiple	Probability	Small	15.14
	Multiple	Probability	Large	11.66
	Multiple	Frequency	Small	18.65
	Multiple	Frequency	Large	12.85

Table 9 (Cont'd).

Means for Difference Scores for Subjects' Estimates of a Probability of a Match and Bayesian Solutions

Case Type	Target	Format	Reference	Difference Score
Murder				
	Single	Probability	Small	9.86
	Single	Probability	Large	14.33
	Single	Frequency	Small	12.09
	Single	Frequency	Large	14.54
	Multiple	Probability	Small	17.79
	Multiple	Probability	Large	10.69
	Multiple	Frequency	Small	13.81
	Multiple	Frequency	Large	15.73
Sexual Assault				
	Single	Probability	Small	31.46
	Single	Probability	Large	31.06
	Single	Frequency	Small	31.28
	Single	Frequency	Large	29.38
	Multiple	Probability	Small	30.48
	Multiple	Probability	Large	29.69
	Multiple	Frequency	Small	28.29
	Multiple	Frequency	Large	30.47

Table 10.

Judgments of a Probability of a Match Compared to Bayesian Solutions

	Responses	
	Subject Estimates	Bayesian Solutions
Case 1 – Kidnapping	81.7_a	94_b
Case 2 – Arson	78.7_a	76_a
Case 3 – Murder	77.7_a	90_b
Case 4 – Sexual Assault	86.4_a	61_b

Note. Judgments were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at p < .001 in a one-sampled t-test.

Table 11.

Judgments of a Probability of a Match Compared to Bayesian Solutions for Each Case

Case Type	Target	Format	Reference	Subject Estimates	Bayesian Solutions
Kidnapping					
	Single	Probability	Small	84.1_a	94_b
	Single	Probability	Large	84.9_a	94_a
	Single	Frequency	Small	77.5_a	94_b
	Single	Frequency	Large	82.8_a	94_a
	Multiple	Probability	Small	81.5_a	94_b
	Multiple	Probability	Large	85.0_a	94_a
	Multiple	Frequency	Small	85.2_a	94_a
	Multiple	Frequency	Large	73.7_a	94_b
Arson					
	Single	Probability	Small	81.3_a	77_a
	Single	Probability	Large	80.2_a	77_a
	Single	Frequency	Small	81.8_a	77_a
	Single	Frequency	Large	74.9_a	77_a
	Multiple	Probability	Small	79.9_a	77_a
	Multiple	Probability	Large	83.8_a	77_a
	Multiple	Frequency	Small	76.1_a	77_a
	Multiple	Frequency	Large	72.0_a	77_a

Note. Judgments were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at p < .001 in a one-sampled t-test.

Table 11. (Cont'd).

Judgments of a Probability of a Match Compared to Bayesian Solutions for Each Case

Case Type	Target	Format	Reference	Subject Estimates	Bayesian Solutions
Murder					
	Single	Probability	Small	83.5_{a}	90_{a}
	Single	Probability	Large	78.1_{a}	90_{b}
	Single	Frequency	Small	82.8_{a}	90_{a}
	Single	Frequency	Large	77.0_{a}	90_{b}
	Multiple	Probability	Small	74.7_{a}	90_{b}
	Multiple	Probability	Large	81.0_{a}	90_{b}
	Multiple	Frequency	Small	77.5_{a}	90_{b}
	Multiple	Frequency	Large	68.5_{a}	90_{b}
Sexual Assault					
	Single	Probability	Small	90.9_{a}	61_{b}
	Single	Probability	Large	88.9_{a}	61_{b}
	Single	Frequency	Small	88.7_{a}	61_{b}
	Single	Frequency	Large	85.7_{a}	61_{b}
	Multiple	Probability	Small	85.4_{a}	61_{b}
	Multiple	Probability	Large	87.1_{a}	61_{b}
	Multiple	Frequency	Small	84.9_{a}	61_{b}
	Multiple	Frequency	Large	80.2_{a}	61_{b}

Note. Judgments were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at $p < .001$ in a one-sampled t-test.

Table 12.

Means for Difference Scores for Subjects' Estimates of a Probability of a Match and the Inverse Probability

Case Type	Target	Format	Reference	Difference Score
Kidnapping				
	Single	Probability	Small	10.03
	Single	Probability	Large	7.60
	Single	Frequency	Small	15.66
	Single	Frequency	Large	11.77
	Multiple	Probability	Small	9.83
	Multiple	Probability	Large	9.03
	Multiple	Frequency	Small	7.09
	Multiple	Frequency	Large	18.59
Arson				
	Single	Probability	Small	14.17
	Single	Probability	Large	18.18
	Single	Frequency	Small	16.96
	Single	Frequency	Large	15.76
	Multiple	Probability	Small	17.17
	Multiple	Probability	Large	14.52
	Multiple	Frequency	Small	20.55
	Multiple	Frequency	Large	12.38

Table 12 (Cont'd).

Means for Difference Scores for Subjects' Estimates of a Probability of a Match and the Inverse Probability

Case Type	Target	Format	Reference	Difference Score
Murder				
	Single	Probability	Small	9.92
	Single	Probability	Large	10.70
	Single	Frequency	Small	11.40
	Single	Frequency	Large	11.15
	Multiple	Probability	Small	14.69
	Multiple	Probability	Large	7.59
	Multiple	Frequency	Small	11.23
	Multiple	Frequency	Large	19.59
Sexual Assault				
	Single	Probability	Small	13.82
	Single	Probability	Large	13.79
	Single	Frequency	Small	14.97
	Single	Frequency	Large	13.31
	Multiple	Probability	Small	14.10
	Multiple	Probability	Large	12.00
	Multiple	Frequency	Small	13.35
	Multiple	Frequency	Large	15.41

Table 13.

Judgments of a Probability of a Match Compared to the Inverse Probability

	Responses	
	Subject Estimates	Inverse Probabilty
Case 1 – Kidnapping	81.7_a	90_b
Case 2 – Arson	78.7_a	70_b
Case 3 – Murder	77.7_a	80_a
Case 4 – Sexual Assault	86.4_a	80_b

Note. Judgements were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at $p < .001$ in a one-sampled t-test.

Table 14.

Judgments of a Probability of a Match Compared to the Inverse Probability for Each Case

Case Type	Target	Format	Reference	Subject Estimates	Bayesian Solutions
Kidnapping					
	Single	Probability	Small	84.1_a	90_a
	Single	Probability	Large	84.9_a	90_a
	Single	Frequency	Small	77.5_a	90_a
	Single	Frequency	Large	82.8_a	90_a
	Multiple	Probability	Small	81.5_a	90_a
	Multiple	Probability	Large	85.0_a	90_a
	Multiple	Frequency	Small	85.2_a	90_a
	Multiple	Frequency	Large	73.7_a	90_a
Arson					
	Single	Probability	Small	81.3_a	70_b
	Single	Probability	Large	80.2_a	70_a
	Single	Frequency	Small	81.8_a	70_b
	Single	Frequency	Large	74.9_a	70_a
	Multiple	Probability	Small	79.9_a	70_a
	Multiple	Probability	Large	83.8_a	70_b
	Multiple	Frequency	Small	76.1_a	70_a
	Multiple	Frequency	Large	72.0_a	70_a

Note. Judgments were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at p < .001 in a one-sampled t-test.

Table 14. (Cont'd).

Judgments of a Probability of a Match Compared to the Inverse Probability for Each Case

Case Type	Target	Format	Reference	Subject Estimates	Bayesian Solutions
Murder					
	Single	Probability	Small	83.5_{a}	80_{a}
	Single	Probability	Large	78.1_{a}	80_{a}
	Single	Frequency	Small	82.8_{a}	80_{a}
	Single	Frequency	Large	77.0_{a}	80_{a}
	Multiple	Probability	Small	74.7_{a}	80_{a}
	Multiple	Probability	Large	81.0_{a}	80_{a}
	Multiple	Frequency	Small	77.5_{a}	80_{a}
	Multiple	Frequency	Large	68.5_{a}	80_{a}
Sexual Assault					
	Single	Probability	Small	90.9_{a}	80_{b}
	Single	Probability	Large	88.9_{a}	80_{a}
	Single	Frequency	Small	88.7_{a}	80_{a}
	Single	Frequency	Large	85.7_{a}	80_{a}
	Multiple	Probability	Small	85.4_{a}	80_{a}
	Multiple	Probability	Large	87.1_{a}	80_{a}
	Multiple	Frequency	Small	84.9_{a}	80_{a}
	Multiple	Frequency	Large	80.2_{a}	80_{a}

Note. Judgments were made on a scale from 0 (not at all) – 100 (completely). Means in the same row that do not share the same subscript differ at p < .001 in a one-sampled t-test.

Table 15.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 1 (Kidnapping)

Variable	B	S.E.	β	t	p
Step 1					
Strength in Evidence	-.230	.023	-.538	-9.89	.000

Note: R^2 = .29.

Table 16.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 2 (Arson)

Variable	B	S.E.	β	t	p
Step 1					
Strength in Evidence	-.196	.020	-.540	-9.93	.000

Note: $R^2 = .29$.

Table 17.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 3 (Murder)

Variable	B	S.E.	β	t	p
Step 1					
Probability of Murder	-.012	.001	-.550	-10.15	.000
Step 2					
Probability of Murder	-.011	.001	-.513	-9.269	.000
Strength of Evidence	-.016	.006	-.141	-2.55	.011

Note: $R^2 = .32$.

Table 18.

Summary of Stepwise Regression Analysis for Variables Predicting Verdict – Case 1 (Sexual Assault)

Variable	B	S.E.	β	t	p
Step 1					
Probability of Sexual Assault	-.009	.001	-.437	-7.52	.000

Note: $R^2 = .19$.

Figure Caption

Figure 1. Two-way interaction for dependent variable of the difference of estimates of a match and the correct Bayesian solutions for the case of arson.

Figure 2. Three-way interaction for dependent variable of the difference of estimates of a match and the correct Bayesian solutions for the case of murder.

Figure 3. Three-way interaction for dependent variable of the difference of estimates of a match and reliability for the case of murder.

Figure 1. Two-way interaction for dependent variable of the difference of estimates of a match and the correct Bayesian solutions for the case of arson.

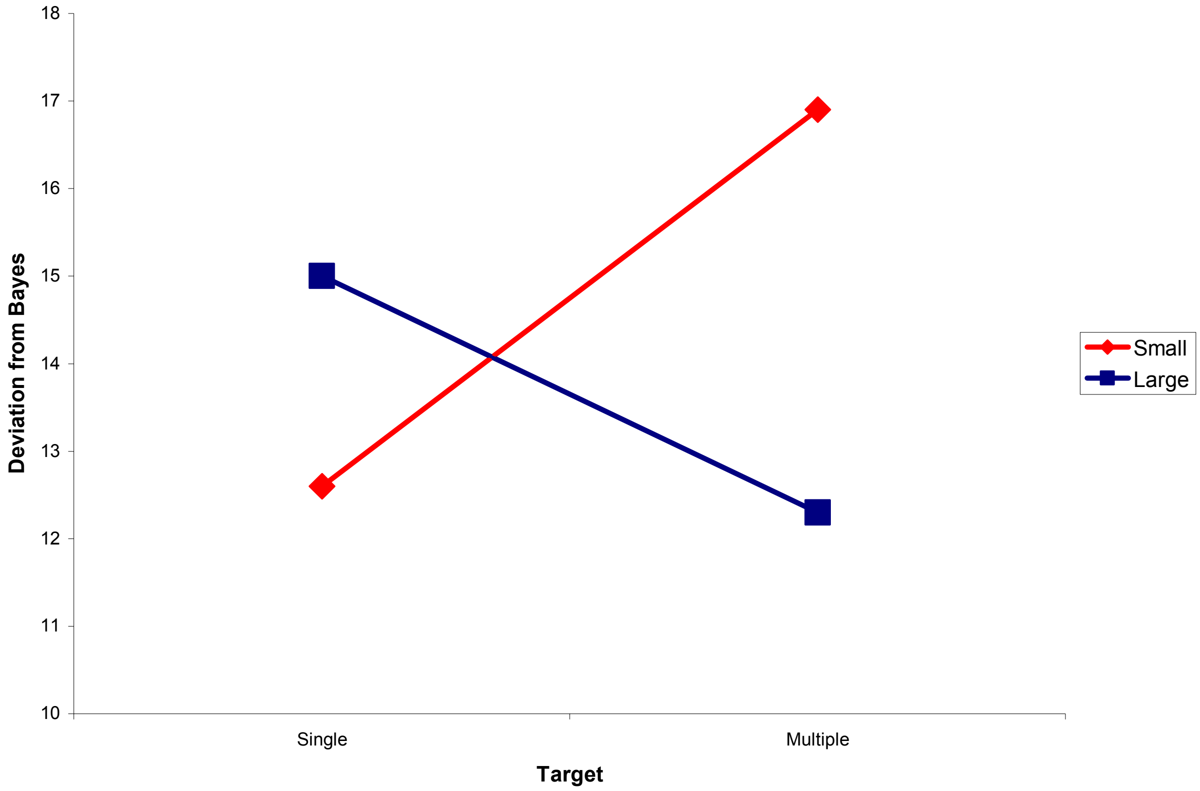


Figure 2. Three-way interaction for dependent variable of the difference of estimates of a match and the correct Bayesian solutions for the case of murder.

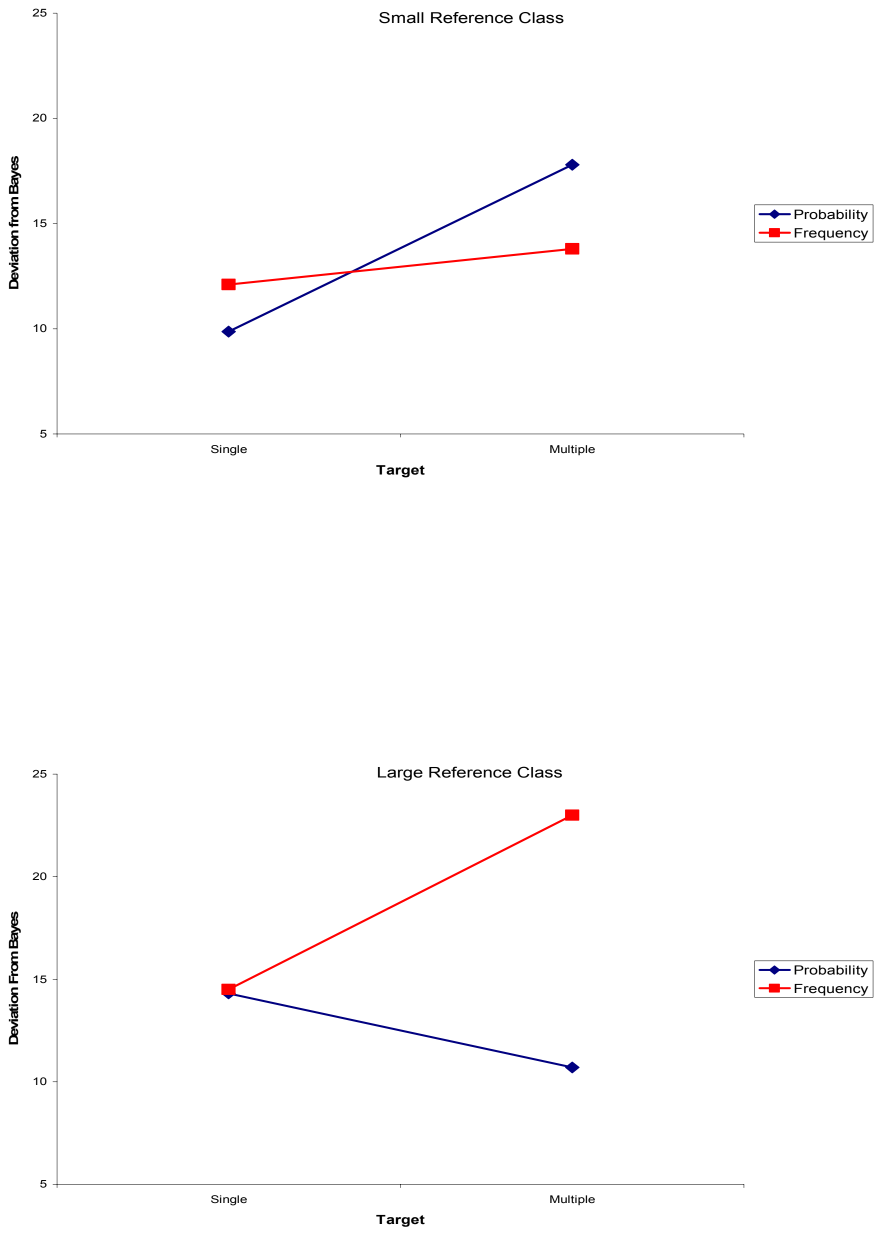
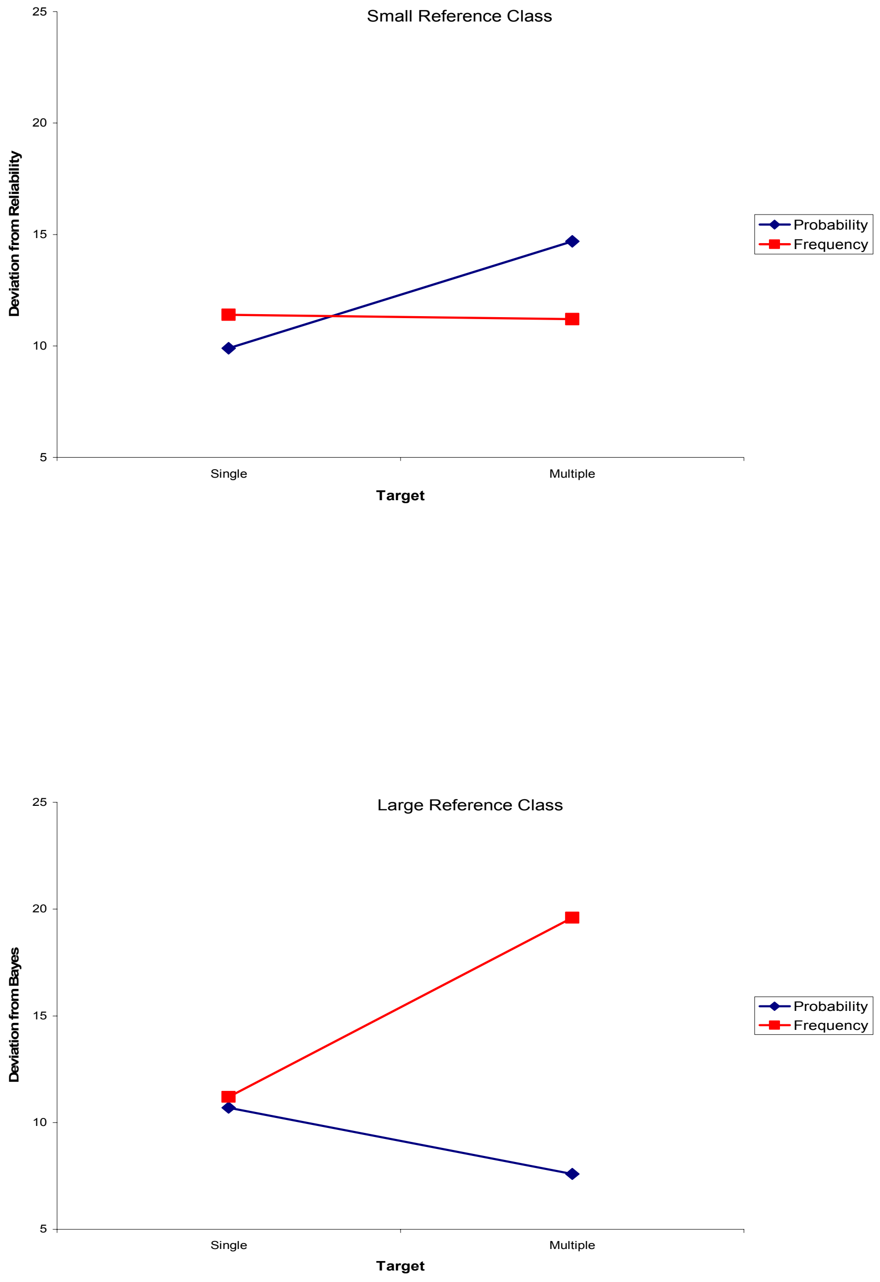


Figure 3. Three-way interaction for dependent variable of the difference of estimates of a match and reliability for the case of murder.


Small Reference Class
Deviation from Reliability
Single
Multiple
Target
Probability
Frequency
Large Reference Class
Deviation from Bayes
Single
Multiple
Target
Probability
Frequency