COMMUNICATION IS A TWO-WAY STREET: INVESTIGATING COMMUNICATION FROM COUNSELORS TO LOW-RISK INDIVIDUALS ON THE CONDITIONAL RISK OF HIV

by

KATRINA M. ELLIS

B.S., Michigan Technological University, 2008

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Gary L. Brase

Abstract

In 2006, the Center for Disease Control and Prevention recommended the revision of state HIV testing laws. With these recommendations, more low-risk individuals are tested regardless of their risk group. However, there is a greater chance of a false positive test result for low-risk individuals than for high-risk individuals. Additionally, previous research found that doctors and HIV counselors in Germany did not accurately communicate the relationship between risk factors and false positive tests (Gigerenzer, Hoffrage, & Ebert, 1998). This study aimed to (1) compare the findings of the 1998 German sample to HIV hotline counselors in the United States in 2011; and (2) to investigate the ability of students to calculate the conditional probability of HIV for a low-risk individual after receiving a positive test, based on idealized transcripts of conversations with HIV hotline counselors. The first study found that HIV hotline counselors use both verbal expressions of risk and percentages to communicate HIV testing statistics. Additionally, 2011 American counselors were more aware of the chance of false positives and false negatives than compared to the 1998 German sample. However, no 2011 American counselors were able to provide an accurate positive predictive value for a low-risk woman. The second study found low performance among students in the calculation of the positive predictive value. Performance was facilitated by a natural frequency format for high numerate individuals. There were different patterns of results for the General Numeracy Scale and the Subjective Numeracy Scale. This would suggest that these two scales might be measuring different constructs. These findings are consistent with the two theories supporting the Frequency Effect, namely the Frequentist

Hypothesis and the Nested Sets Hypothesis. Additionally, this research suggests computation of the conditional risk of HIV is facilitated by a natural frequency format. Teaching techniques have been developed and demonstrate long lasting improvement in health related computations. If a few hours of training is all that it takes to communicate these life and death statistics in a manner that is consistent with reasoning, health practitioners and students should be required to have more education in communicating and computing probabilities.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge the following people for their help in the researching and writing of this thesis: Dr. Gary Brase for his guidance and perspective; Dr. Lester Loschky and Dr. Richard Harris for their suggestions and additions; David Kilby for transcribing the interviews with the HIV counselors; Jaden Lantz and Marina Vilardo for coding the transcripts; and William Hill, Jericho Hockett, Margo Woller-Carter, April Park, Heather Hilgenkamp and Lora Peadan for their support and proofreading skills.

Dedication

I would like to dedicate this work to my family. They have encouraged me from the beginning and I would not be the person I am without their support. I love you Mom, Dad, Nico and Jake.

"In the fields of observation, chance favors only the prepared mind."

- Louis Pasteur

Chapter 1 - Introduction

According to a Congressional Research Service report, $21.2 billion dollars of governmental funds were allocated to human immunodeficiency virus (HIV) and acquired immune deficiency syndrome (AIDS) research, prevention, and treatment in 2007 (Johnson, 2008). Additionally, the Centers for Disease Control and Prevention (CDC) received $879 million in 2007 (CDC, 2007) to support education programs, which inform the public about HIV prevention, transmission, treatment and testing. Among the information provided to the general public about HIV/AIDS, the CDC reports statistical information and recommendations for counseling. Considering the amount of funds reserved for HIV/AIDS research and education, counselors should not only be knowledgeable in counseling for individuals with HIV/AIDS, but also able to communicate information about HIV testing statistics differently to high risk and low risk populations.

Time pressures on physicians, counselors, nurses, and other health professionals, however, limit the amount of information that can be transferred from doctor to patients. Effective and efficient ways to communicate testing statistics are necessary, but it is questionable whether funds are being used to educate health professionals in communicating uncertain risks. Previous research has demonstrated that doctors and counselors in Germany not only had a hard time with numerical information, but they became confused with repeated questions about HIV testing statistics (Gigerenzer et al., 1998). If doctors were more knowledgeable in testing statistics, they would understand the relationship between risk factors and the amount of false positives; specifically, as the risk for HIV/AIDS decreases, the proportion of false positive test results increases, up to 50%—this is a measures of the conditional risk of HIV (for the calculation of this statistics, see Appendix A).

Also in 2006, the CDC recommended that more individuals be routinely tested for HIV (CDC, 2007; Neff & Goldschmidt, 2011). The CDC suggested that all individuals within a population where the prevalence of HIV is .1% be tested for HIV. Additionally, the CDC suggested that health practitioners adopt an opt-out procedure. This out-out procedure gives individuals the ability to decline an HIV test, but patients are no longer required to sign a specific consent form. The consent to medical care is sufficient for consent to HIV testing. Not all states are required to adopt the suggestions by the CDC, but since the 2006 recommendations, only 5 states require a separate informed consent for HIV testing (Neff & Goldschmidt, 2011).

With more testing of individuals, there is an increase in the amount of low-risk individuals who will get tested for HIV. With more low-risk individuals being tested, it is important that they understand the statistics pertaining to their risk group. All individuals testing for HIV should understand that there is a relationship between individual risk and the possibility of a false positive test. Specifically, low-risk individuals are more likely to receive a false positive test than high-risk individuals. There are very serious consequences associated with receiving a false positive test. Case studies of individuals who received a false positive HIV test result suggest a similar pattern of outcomes in the reception of a positive test and the reception of a proceeding negative test result (Bhattacharyaa, Bartonb, & Catalanc, 2008). Initially when individuals receive a positive HIV test result, they make drastic changes to their lifestyle. They may spend all of their money, change jobs, divorce their spouse or lose contact with friends and family because their life expectancy is reduced. Additionally, upon discovering that their test result was a false positive they initially feel elated and relieved, but later experience anger and guilt.

Under the new guidelines for HIV screening, there are more low-risk individuals who will receive a false positive test result. At the same time, however, most brochures available in the waiting rooms of doctors' offices suggest that false positive results rarely occur (Krantzler, 2007; Proudfoot, 2008; Schettler, 2005). The conditional risk of HIV, or positive predictive value (PPV), is the likelihood that an individual has HIV after receiving a positive test result (for a glossary of terms, see Appendix B). To calculate the PPV it is necessary to know particular statistics about the test in relation to an individual's risk group and use Bayesian reasoning in some form. Bayesian reasoning is used to calculate the conditional probabilities of hypotheses based on updating evidence—the probability that a person is HIV positive, after receiving a positive test result. Research has shown that the ability of patients to understand risks is influenced by the format of the numerical information (Hoffrage, Gigerenzer, Krauss, & Martignon, 2002; Ghosh, Crawford, Pruthi, Williams, Neal, Sandhu et al., 2008; Miron-Shatz, Honch, Graef, Sagi, 2009; Schapira, Nattinger, & McHorney, 2001; Sirota & Juanchich, 2011; Yamagishi, 2003). Specifically, a naturally sampled frequency format of numerical information evokes more correct responses on Bayesian reasoning tasks than a probability format (Brase, 2008; Chapman & Liu, 2009; Gigerenzer & Hoffrage, 1995; Gigerenzer, 1996; Gigerenzer & Edwards, 2003; Hoffrage & Gigerenzer, 1998). Research has also found a facilitative effect of frequencies in a comparison of risk task

(Miron-Shatz et al., 2009). Comparisons of risk require fewer calculations than Bayesian reasoning; however, this type of problem is also influenced by information format.

Has the finding that natural frequencies facilitate health computations infiltrated the language used by counselors to communicate HIV testing statistics? The first aim of the present research was to assess the current state of communication between HIV hotline counselors and low-risk individuals. Specifically, the aim was to investigate the ability of HIV hotline counselors to provide accurate HIV testing statistics, the format that HIV hotline counselors use to communicate the testing statistics, and to compare these findings with the 1998 German sample. The second aim of the present research was to investigate the ability of individuals to calculate the posterior probability in a Bayesian reasoning task. Specifically, the aim was to assess students' ability to calculate the PPV and the influence of format, numeracy, interpretation, confidence and clarity on performance of a Bayesian reasoning task with ecologically valid risk.

HIV Statistics

It is often difficult for individuals to think of numbers in a probabilistic way. According to a qualitative study on formats of risk information, individuals would rather think in terms of definite answers rather than receiving risk information with the uncertainty inherent in probabilities (Schapira et al., 2001). Additionally, when calculations are necessary, individuals would rather receive statistical information in a numerical format than a verbal expression of risk (Vahabi, 2010). In calculating health risks, Bayesian reasoning can be used in any instance for which it is necessary to evaluate a hypothesis (diagnosis) given particular evidence (symptoms/test results). In the HIV example, there are only two possible hypotheses: an individual has HIV or an individual does not have HIV. The likelihood of HIV given a positive test—the posterior probability—can be calculated using Bayesian reasoning. The posterior probability is the possibility of the hypothesis (i.e., an individual has HIV) given the data (i.e., a positive test). For the present study, the posterior probability is the positive predictive value (PPV) or the probability of having HIV given the risk factors and rate of incidence. The prior probability is the base rate of individuals with the disease in the population (i.e., prevalence). To get from the prior probability to the posterior probability we need to use the evidence we have on the nature of HIV tests, namely the test-dependent statistics of sensitivity and specificity—the ability of the test to accurately identify individuals with and without HIV respectively.

There are four possible outcomes of HIV tests: true positives, true negatives, false positives and false negatives. What may be more important are the two latter outcomes; the two types of errors that can occur (see Table 1-1). A false negative occurs when an individual receives a negative test result when they do have the disease; whereas a false positive occurs when an individual receives a positive test result when they do not have the disease. These errors can be attributed to the varying criteria on which the tests are assessed in addition to possible human errors. Additionally, it should be noted that the sensitivity and specificity statistics are dependent on the type of test, whereas prevalence and the positive predictive value are dependent on the risk factors of the individual being tested. The test used most frequently to detect HIV is the Enzyme-linked immunosorbent assay (ELISA). This test detects the antibodies in response to HIV virus. The statistics of window period, the prevalence of HIV within a population, the sensitivity of ELISA, and the specificity of the ELISA can be used to answer questions of personal risk including the accuracy of the tests, the risk of the disease, and the PPV.

Window Period

The window period for a test is the time between the possible exposure and the test. The ELISA is an HIV test that detects antibodies responding to the HIV virus. It takes time for these antibodies to respond, which is the reason the window period is necessary. The window period for HIV tests is 90 days. This means that an individual needs to wait at least 90 days since the possible exposure, in order for the HIV test to be able to detect the antibodies responding to the HIV virus.

Prevalence

The prevalence of a disease is the commonality of the disease among a group of individuals; which could be the individuals in a country (e.g., United States), individuals in a state (e.g., Arkansas), or individuals in groups determined by risk factors (e.g. sexual behaviors, sexual orientation, and IV drug use). Without knowing additional evidence (such as the error rates for false positives and false negatives) about a risk group, the likelihood of HIV is equal to the prevalence of the disease within a population. The most recent statistics on the prevalence of HIV come from 2006. The prevalence of HIV in the United States in 2006 was estimated to be over 1,100,000; of those individuals, 23% were female (CDC, 2006). The prevalence varies depending on risk factors. The population for the present study was women with low risk and the prevalence for this risk group is about 1%. This means that out of every 10,000 women, 100 of those women had HIV in 2006. The prevalence of a disease is important to individuals when the statistic reported is specific to the individual’s demographics and risk group.

Sensitivity

Sensitivity refers to the ability of the test to identify individuals with the disease. The equation for sensitivity is

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{1}$$

The sensitivity varies depending on the test; the most widely used HIV test is ELISA. The most current estimate of sensitivity for the ELISA is 99% (George, 1998). This means that, if 100 out 10,000 women have HIV, out of the 100 people who do have HIV and are tested, 99 of them will receive positive tests—one person will receive a false negative test. Note that the explanation of the sensitivity statistic is in reference to 100 people being tested, all of whom are known *to have* HIV.

Specificity

Specificity is the ability of the test to identify those without the disease. The equation for specificity is

$$\text{specificity} = \frac{TN}{TN + FP} \tag{2}$$

The specificity, like the sensitivity, varies depending on the test. The most current estimate of specificity for the ELISA is 99.14% (George, 1998). This means that, if 9,900 out of 10,000 people do not have HIV, of the 9,900 people who do not have HIV and are tested, 9,811 of them will receive negative tests— 89 people will receive false positive tests. Note that the explanation of the specificity statistic is in reference to 9,900 people, all of whom are known *to not have* HIV.

Positive Predictive Value (PPV)

The statistic most relevant to low-risk individuals is the positive predictive value (PPV) of HIV tests. Unfortunately, this statistic is rarely reported in frequencies (Gigerenzer et al., 1998)—if it is reported at all. Additionally, the PPV is often confused with sensitivity. Sensitivity is the ability of a test to accurately identify positive individuals, whereas the positive predictive value is the probability of being infected having received a positive test result. The PPV demonstrates the relationship between risk factors and the false positive rate in its calculation.

To calculate the PPV, it is necessary to use Bayesian reasoning; however, the actual calculation may be different depending on the information given and the way individuals interpret the information. The PPV can be calculated using the prevalence, sensitivity and specificity statistics in Bayes' Theorem (Hunt, 2007):

$$\text{PPV} = \frac{(sensitivity)(prevalence)}{(sensitivity)(prevalence) + (1 - specificity)(1 - prevalence)} \quad (3)$$

Alternatively, if the frequency of true positives and false positives are known for a population, the formula to calculate the PPV can be simplified:

$$\text{PPV} = \frac{TP}{TP + FP} \quad (4)$$

Both calculations suggest that as prevalence decreases, the likelihood that a test result is a false positive increases. The obvious advantage of *Equation 4* is that it requires fewer operations than *Equation 3.*

Empirical Foundation for the Current Research

The present study was a partial replication and extension of a study conducted in Germany by Gigerenzer and colleagues (1998). The researchers documented visits to 20

AIDS counseling centers, each in a different city, to receive pre-HIV test counseling and to receive an HIV test. A researcher portrayed a low-risk individual getting an HIV test. The researcher asked questions about sensitivity, specificity, prevalence, positive predictive value, and the window period regarding the HIV test. The responses of the 20 counselors were compared for similarities.

The counselors’ answers to the questions on sensitivity and window period were simple and accurate explanations of the concepts. However, 13 of 19 counselors suggested that false positives do not occur, and five of 19 counselors suggested that false negatives do not occur. Out of 20 counselors, only one counselor communicated that for low-risk patients, the false positive rate was higher than for high-risk patients. The most uncertainty arose from the question of prevalence. Sixteen counselors suggested that the prevalence cannot be determined due to unreported cases. Four of the counselors stated that prevalence statistics are not useful for the individual case. (On the contrary, prevalence is necessary in determining the positive predictive value for an individual.)

In addition, this study suggested that counselors are using percentages more than natural frequencies to communicate statistical information about the HIV tests. Additionally, the majority of counselors assumed that false positives never occur or are eliminated through repeated testing, and confused sensitivity and the positive predictive value. Gigerenzer and colleagues (1998) suggested that the counselors’ mistakes might have been avoided if the counselors would have presented information in terms of natural frequencies rather than percentages.

Bayesian Reasoning and Formats of Risk

Since the 1998 German study, a growing body of literature suggests individuals are more accurate in their calculation of Bayesian reasoning tasks when information is presented in terms of naturally sampled frequencies compared to probabilities (Brase, 2008; Chapman & Liu, 2009; Gigerenzer & Edwards, 2003). "In a naturally sampled population, natural frequencies are obtained by counting individuals according to their features (e.g. disease versus no disease, positive test result versus negative test result)" (p. 345, Hoffrage et al., 2002). Think of a doctor giving HIV test results. The doctor may administer over 100 HIV tests in a year. Some patients will have HIV and some patients will not have HIV. Some patients will have a positive test and some will have a negative test. The occurrence of these possibilities is observed over time by natural sampling. The counts of these occurrences by natural sampling are naturally sampled frequencies.

There are multiple explanations for why naturally sampled frequencies facilitate the computation of Bayesian reasoning problems. One explanation for this frequency effect is that naturally sampled frequencies are consistent with the mental framework that we have evolutionarily developed to reason with probabilistic information—the Frequentist Hypothesis (Brase, 2008; Gigerenzer & Hoffrage, 1995). Another explanation as to why naturally sampled frequencies increase performance in Bayesian reasoning problems is that the natural frequency format makes clear the subset relationships between probabilities—the nested sets view (Girotto & Gonzalez, 2001). Although the reason for this effect of format is unclear, it is clear that patients would benefit from natural frequency presentations of risk information.

Figures and Tables

Table 1-1 Contingency Table Displaying All Outcomes for HIV Testing

	HIV +	HIV -
Test +	True positive	False Positive
Test -	False Negative	True negative

Chapter 2 - Experiment 1

The present study assessed the current state of communication of HIV risk from HIV hotline counselors to low-risk individuals. Specifically, this study aimed to answer the following questions:

Q1: What statistical information can HIV hotline counselors provide to patients?

Q2: Do hotline counselors communicate HIV testing statistics accurately?

Q3: What format do HIV hotline counselors use to communicate statistical information in (probabilities, frequencies, or verbal expressions of numerical probabilities)?

It is important to have knowledgeable health counselors who understand the statistical information they are communicating to others. A particular difficulty appears to be communication of an accurate positive predictive value for low-risk individuals. Hopefully, in the more than 13 years since the Gigerenzer et al. (1998) study in Germany, counselors have become aware of the differences in the PPV between high-risk and low-risk individuals.

Hypotheses

The aim of the study was to investigate the similarities and differences in communication of HIV statistics from counselors to low-risk patients between the 1998 German sample and an American sample of HIV hotline counselors in 2011. If counselors have incorporated previous research in the language they use to communicate to low-risk individuals the first hypothesis suggests that 2011 American counselors will use the naturally sampled frequency format to communicate HIV testing statistics, will understand that false negatives and false positives do occur, and will not confuse the sensitivity and specificity. However, it is possible that counselors are not aware of the previous research on the frequency effect and the relationship between risk factors and the positive predictive value. If previous research has not been incorporated into the language of HIV hotline counselors then the alternative hypothesis suggests that 2011 American counselors will still use percentages to communicate HIV testing statistics, they will still believe that false negatives and false positives do not occur, and they will still confuse the sensitivity and positive predictive value statistics.

Method

Participants

Participants of this study were HIV counselors from across the United States. According to the U. S. Department of Health and Human Services website, 49 states have local HIV hotlines as well as the District of Columbia, Puerto Rico, and the Virgin Islands. For the purpose of the study, it was necessary to record the interviews with the HIV counselors. State laws had to be taken into consideration when recording the telephone conversations. In some states, it is required that both parties consent to the recording of telephone conversations. Therefore, to avoid breaking the law, hotlines from these states were not interviewed. These exclusions were based on a concern that if the counselors were asked for their consent, it may have biased their answers to the interview questions. The Kansas State University Institutional Review Board consented to the waiver of the informed consent and debriefing statements for this research because the interview questions should be no different from questions counselors would be asked in their everyday work experience. During the interviews, there was a technical difficulty with the recording equipment; six of interviews had to be discarded because the counselors' responses were inaudible. These interviews were not redone because the same counselor might have answered. This would be a problem because it could have caused suspicion being asked the same questions. Additionally, no personal information about the HIV hotline counselors was collected so it would have not been known if the same counselor was the first available counselor from that state's hotline. The participants of this study were the remaining 28 state hotlines and 1 national hotline.

Materials

One researcher (KE) was the sole interviewer for this study. She presented herself as being a member of the low-risk group for HIV—that is, an individual in a monogamous, heterosexual relationship with no prior history of intravenous drug use. Interviews were conducted on a landline office phone. A telephone handset mini recorder controller was used in tandem with Audacity (a free sound recording program) to record the conversations to electronic files.

KE used the same script of questions on HIV testing for all counselors (see Appendix C). Counselors were asked, in lay language, questions about the specificity of the test, sensitivity of the test, the window period, the PPV, and the prevalence of HIV for the researcher's particular risk group. These questions were adapted from Gigerenzer et al. (1998).

Procedures

Before conducting the interviews analyzed for the current study, practice calls were made to the states that were ineligible for data collection (the states that required both parties to consent to the recording of phone conversations). The purpose of the practice calls was to identify any problems in data collection. After completing 10 practice calls to 10 different state hotlines, the researcher began interviewing the HIV hotline counselors for the current research.

The researcher called and interviewed the first available counselor for each HIV hotline. Once on the phone, KE explained to the counselors that she had been researching information on the Internet about HIV testing statistics, and that she had gotten confused about the testing statistics pertaining to low-risk individuals. Additionally, the interviewer told the counselors that she was planning on getting pregnant and it was necessary for her to get an HIV test. KE asked the questions in the same order for all counselors. In the event of a non-numerical answer, the counselors were not pressed for a more exact answer due to the finding that counselors become irritable after repeated questioning (Gigerenzer et al., 1998). After asking the questions, KE thanked the counselors for their time and ended the interview. All interviews were anonymously recorded such that no personal information about the counselors was gathered. After all the interviews were complete, the recordings of the interviews were transcribed for further analysis (for full transcriptions from the present study, see https://sites.google.com/site/hivtranscripts).

Coding

Two independent coders, blind to the hypotheses of the study (JL and MV), coded the transcripts. Both coders were trained in numerical formats. The initial interrater reliability was $\alpha = .77$ and all discrepancies in coding were resolved by mutual discussion. For each question, the transcripts were coded for the following variables: accuracy, format, knowledge of response, and assessment of risk.

Accuracy of answer for the window period question was determined in two ways; first, counselors' responses were considered accurate if their answer was exactly 90 days, and second, counselors' responses were coded as accurate if 90 days was included in the range of time that they provided. If their answer was not 90 days or they did not give a range that included 90 days, their answer was coded as incorrect. Answers to the sensitivity question were coded as accurate if the counselors responded with either 99% sensitive or 1% false negatives. Answers to the specificity question were coded as accurate if the counselors responded with either 99.14% specific or <1% false positives.

Both coders were trained on numerical formats including: percentages, absolute frequencies, relative frequencies, fractions, and ratios. Counselors' responses were coded for the format of numerical information: frequencies, percentages (and other normalized numbers), non-numerical (or verbal expressions of risk), and other (including ratios and fractions). Non-numerical responses of the sensitivity and specificity questions were coded for being either *definite* or *possible*. Definite responses indicated that the counselor denied the possibility of false positives or false negatives; whereas possible responses from counselors indicated that false positives and negatives might occur. Finer distinctions between the non-numerical responses were not made because previous research suggests that there are differences in the rankings of verbal expressions of probabilistic information and that these rankings vary between contexts (Gonzalez-Vallejo, Erev, & Wallsten, 1994). The majority of these non-numerical responses included phrases such as "very likely" or "not at all likely".

Knowledge of response was determined by whether counselors used additional resources to answer the questions. If a counselor stated that they needed to reference additional information, their response was coded such that they did not use their own knowledge to answer that question. Finally, assessment of risk was determined by whether counselors' asked any questions pertaining to risk (e.g. previous sexual history, present sexual history, IV drug use, etc.).

Results

Descriptive Statistics

Assessing Risk

Nine of 29 counselors (31%) asked questions pertaining to risk in order to assess the interviewer's individual risk. The remaining 20 counselors (69%) relied on the interviewer's assessment of risk (as a woman of low risk).

Window Period

All 29 counselors answered the window period question. Fourteen of 29 counselors (48%) accurately reported the window period as being 90 days or the equivalent (for accuracy by question, see Table 2-1). Some counselors reported the window period as a range of time rather than a cut-off of time; when these counselors' responses were included, 23 of 29 counselors (79%) suggested an accurate window period for HIV tests within their reported range. One of 29 counselors (3%) referenced additional sources in responding to the window period question (see Figure 2-1).

Sensitivity

All 29 counselors answered the sensitivity question. No counselors reported the sensitivity of HIV tests as exactly 99% or a 1% false negative rate. When allowed a ± 5% interval of error on the sensitivity of HIV tests, one of 29 counselors (3%) reported the sensitivity of HIV tests as being between 98.5% and 99.5%. Numerical and non-numerical answers were further coded as definite or possible responses. When looking at the proportion of counselors who suggested that false negatives could definitely not occur, four of 29 counselors (14%) suggested that false negatives do not occur.

As for the format of the sensitivity responses, 26 of 29 counselors (90%) used a non-numerical format to convey the sensitivity information. Of the three counselors that did use a numerical format, two counselors used percentages to respond and one counselor used a relative frequency format (to compare percentages and frequencies for the sensitivity and specificity questions, see Table 2-2). One of 29 counselors (3%) used additional sources in responding to the sensitivity question.

Specificity

Twenty-eight of 29 counselors (97%) answered the specificity question. No counselors reported the specificity of HIV tests as 99.14%. When allowed a ± 5% interval of error, one of 28 counselors (3%) reported that the specificity was 98.9%. Numerical and non-numerical answers were further coded as definite or possible responses. When looking at the proportion of counselors who suggested that false positives could definitely not occur, seven of 28 counselors (25%) incorrectly suggested that false positives do not occur.

As for the format of the specificity responses, 20 of 28 counselors (71%) used a non-numerical format to convey the specificity information. Of those counselors that did use a numerical format, seven of 28 counselors (25%) used percentages to respond and one of 28 counselors (3%) used a relative frequency format. Three of 28 counselors (11%) used additional sources to respond to the specificity question.

Positive Predictive Value (PPV)

Twenty-seven of 29 counselors (93%) answered the PPV question. Ten of 27 counselors (37%) misunderstood the PPV question for a question about false positive tests. The remaining 17 of 27 counselors (63%) suggested that the PPV could not be determined. As for the format of the PPV responses, 19 of 27 counselors (70%) used a non-numerical format to convey the PPV information. The remaining eight of 27 counselors (30%) used probabilities to covey the PPV information. One of 27 counselors (4%) used additional sources to respond to the PPV question.

Prevalence

Twenty-six of 29 counselors (90%) answered the prevalence question. Additionally, only half of the counselors who answered the prevalence question gave a numerical response. The reference class used for the prevalence statistic varied between counselors and none of the counselors reported a prevalence specific to the interviewer's stated risk group (i.e., the prevalence for white women with low risk). Of those that gave numerical answers, nine of 26 counselors (35%) used absolute frequencies to answer the prevalence question, three of 26 counselors (12%) used percentages, and two of 26 counselors (8%) used relative frequencies. One counselor (4%) responded to the prevalence question using both absolute frequencies and percentages. Five of 26 counselors (19%) used additional sources to respond to the prevalence question.

Inferential Statistics

The data was further explored using a series of one-tailed difference of proportions tests with the alpha level set at .05. Difference of proportion tests are useful in finding effects with nominal variables and are more powerful than the alternative chi-square test (Cohen, 1988). Counselors used verbal expressions to describe risks significantly more than numerical expressions of risk for the sensitivity (z = 6.04, p < .000, η = 1.83), specificity (z = 3.21, p = .001, η = 0.89), and PPV (z = 2.99, p = .002, η = 0.84) questions (see Figure 2-2). There was no difference between the proportion of counselors that used frequencies and the proportion of counselors that used percentages to describe the sensitivity statistic (z = 0.59, p = .278, η = 0.16). However, a significantly greater proportion of counselors used percentages than frequencies to describe the specificity statistic (z = 2.29, p = .013, η = 0.67; see Figure 2-3).

For the specificity question, seven of 28 counselors (25%) erroneously suggested that false positives do not occur—this proportion is significantly smaller (z = 2.95, p = .002, η = 0.9) than the 13 of 19 German counselors (65%) who suggested that false positives do not occur (see Figure 2-4). This outcome suggests that American hotline counselors are more knowledgeable about the risks of false positives than the German counselors were over a decade ago. However, American counselors had similar errors in reasoning to the German counselors, in that the American counselors suggested false positives rarely occur and that false positives can be eliminated through repeated testing.

For the sensitivity question, four of 29 counselors (14%) erroneously suggested that false negatives do not occur—this proportion is not significantly different (z = 1.09, p = .141, η = 0.32) from the 5 of 19 German counselors (26%) who suggested that false negatives do not occur (see Figure 2-5). Although the current research did not find significance, this outcome suggests that 2011 American hotline counselors are more aware than the 1998 German sample that false negatives occur. Additionally, American counselors had similar errors in reasoning to the German counselors such that the American counselors suggested that false negatives only occur in the window period—

before the antibodies to the HIV virus have had time to increase to be detected by the ELISA.

In addition, 10 of 27 American counselors (37%) confused the sensitivity with the PPV—this proportion is not significantly different ($z = 0.50$, $p = .310$, $\eta = 0.15$) from the 8 out of 18 German counselors (44%) who confused the sensitivity with the PPV (see Figure 2-6). This suggests that 2011 American HIV hotline counselors are less likely to confuse the sensitivity and PPV but the difference was not significant. Additionally, no American HIV hotline counselors were able to provide an accurate PPV.

Together, these results suggest that 2011 American HIV hotline counselors are more likely to communicate that false positives and false negatives could occur, but still confuse the sensitivity and the positive predictive value. Additionally, the 2011 American HIV hotline counselors have similar errors in reasoning as the 1998 German sample in that they still believe that false positives can be eliminated through repeated testing and that false negatives only occur during the window period. In sum, some findings from research have been incorporated into the information that counselors communicate to low-risk individuals; however, HIV hotline counselors need to make clear the relationship between risk factors and the risk of a false positive test result.

Discussion

The purpose of the first study was to assess the information communicated from HIV hotline counselors to patients about HIV testing and statistics of risk. The first hypothesis—to demonstrate a difference in knowledge between American and German counselors—was partially supported. American counselors used percentages more than frequencies to describe risk information, used verbal expressions of risk (e.g., very little chance, it happens sometimes, likely), erroneously suggested that false positives and false negatives never occur, and confused the sensitivity and PPV.

Even with the numerous studies that suggest presenting risk in terms of naturally sampled frequencies facilitates Bayesian reasoning (Brase, 2008; Chapman & Liu, 2009; Gigerenzer & Hoffrage, 1995; Gigerenzer, 1996; Gigerenzer & Edwards, 2003; Hoffrage & Gigerenzer, 1998) and comparisons of risk (Miron-Shatz et al., 2009); American counselors are still using verbal expressions of risk to convey HIV testing statistics. Previous research suggests that individuals prefer to receive risk information in numerical formats when making computations (Vahabi, 2010). In a pilot study, a student sample responded they would be more likely to consult their general practitioner (M = 5.02, SD = 1.12) than a hotline counselor (M = 1.71, SD = 1.09); t (47) = 10.85, p < .000. The knowledge of general practitioners is beyond the scope of the present study, however, the numerical formats used by general practitioners in their explanations of risk warrants further investigation.

American counselors still erroneously believe that false positives and false negatives do not occur. However, a significantly smaller proportion of American counselors suggested that false positives never occur than the proportion of German counselors who suggested that false positives do not occur. These counselors believed that false positives either did not occur or are eliminated through repeated testing. The majority of American counselors understood the risk of a false positive test but were not able to explain that the risk of false positives increases as the amount of risk for an individual decreases. Additionally, a smaller (but not significant) proportion of American counselors suggested that false negatives do not occur than the proportion of German counselors who suggested that false negatives never occur. Similiar to the German sample, the American counselors believed that false negatives occur only during the window period for HIV tests. This data indicates that 2011 American counselors' knowledge about HIV testing is somewhat more accurate than the 1998 German sample's knowledge.

Additionally, a smaller proportion of 2011 American counselors confused the sensitivity and PPV than the German sample. This difference was not significant, but the effect size was large which indicates that with a larger sample significant effects might have been found. These findings would indicate that American counselors are more aware of the PPV than the 1998 German sample, however no American hotline counselors were able to report an accurate PPV for an individual at low-risk for HIV. Again, the PPV can be computed and comprehended using Bayesian reasoning. However, the ability to solve Bayesian reasoning problems and reasoning with probabilistic information is only one component of health numeracy, a subset of health literacy (Baker, Williams, Parker, Gazmarian, & Nurss, 1999). Health numeracy has been described as "the degree to which individuals have the capacity to access, process, interpret, communicate and act on numerical, quantitative, graphical, biostatistical and probabilistic health information needed to make effective health decisions" (p. 375, Golbeck, Ahlers-Schmidt, Paschal, & Dismuke, 2005). Future research should focus on the differences in health outcomes based on decisions from low numerate and high numerate patients. Additionally, future research should investigate the individual differences contributing to differences in health numeracy.

There were not enough accurate responses from counselors to investigate further the effect of format on the accuracy of counselors' responses. No counselors answered the PPV question correctly or used natural frequencies in describing the PPV. One counselor used relative frequencies to describe the specificity of HIV tests and accurately reported the FP rate. One other counselor used relative frequencies to describe the sensitivity of the HIV tests but did not accurately report the FN rate. One explanation for the low accuracy rate is that counselors used verbal expressions of probabilities to express risk instead of accurate numerical expressions of probabilities. Again, research has shown that patients prefer numerical expressions of risk especially when required to make computations based upon the data (Gonzalez-Vallejo, Erev, & Wallsten, 1994). Counselors were not pressed for numerical answers because Gigerenzer and colleagues (1998) found that counselors would get agitated with repeated questioning and the current research did not want to arouse suspicion and possibly biases in responding. Future research should investigate the ability of general practitioners, nurses, counselors and other health professionals to provide accurate statistics on the sensitivity, specificity, prevalence and positive predictive value for individual risk groups.

Figures and Tables

Table 2-1 Proportion of counselors responding and proportion of accurate responses by question

			Correct Responses			
	n		Exact		Range +/- 5%	
Question	%	Freq.	%	Freq.	%	Freq.
Window Period	100%	29/29	48%	14/29	79%	23/29
Sensitivity	100%	29/29	0%	0/29	3%	1/29
Specificity	97%	28/29	0%	0/29	25%	7/28
PPV	93%	27/29	0%	0/29	0%	0/29
Prevalence	90%	26/29	0%	0/26	0%	0/26

Figure 2-1 Raw counts of counselors who referenced additional sources during the interview by question

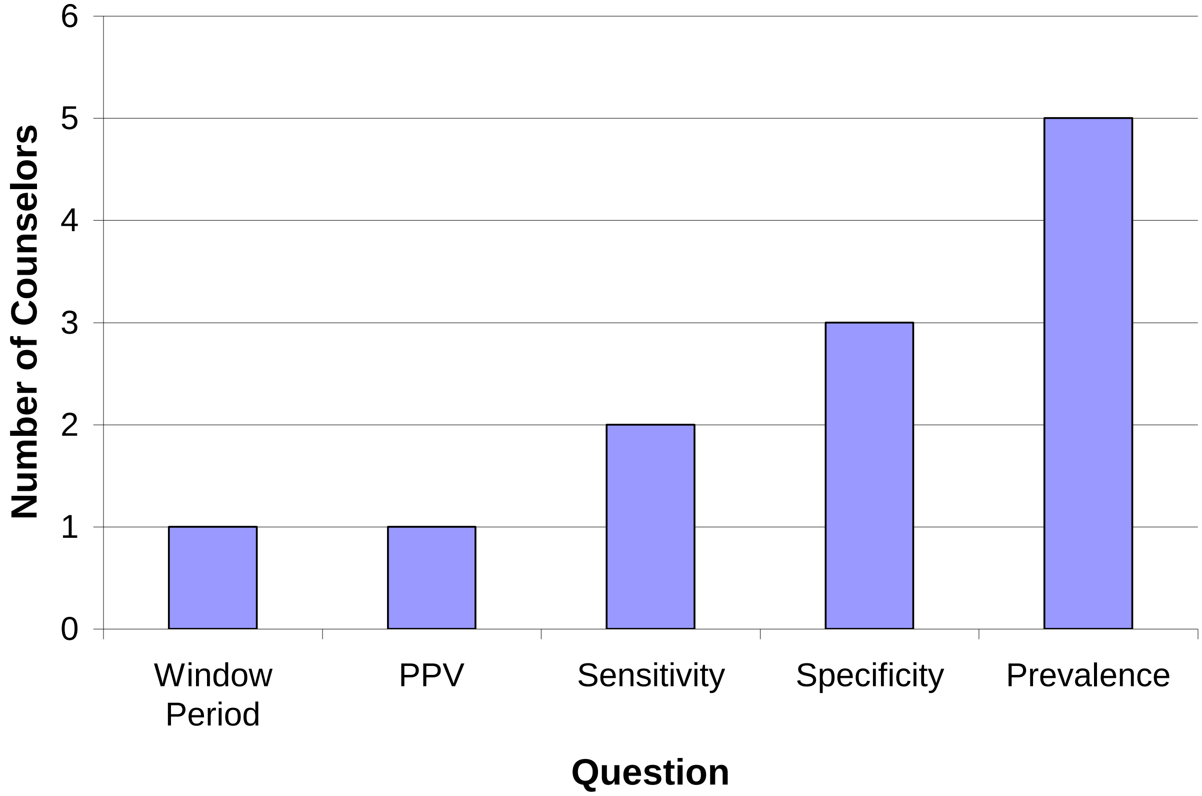
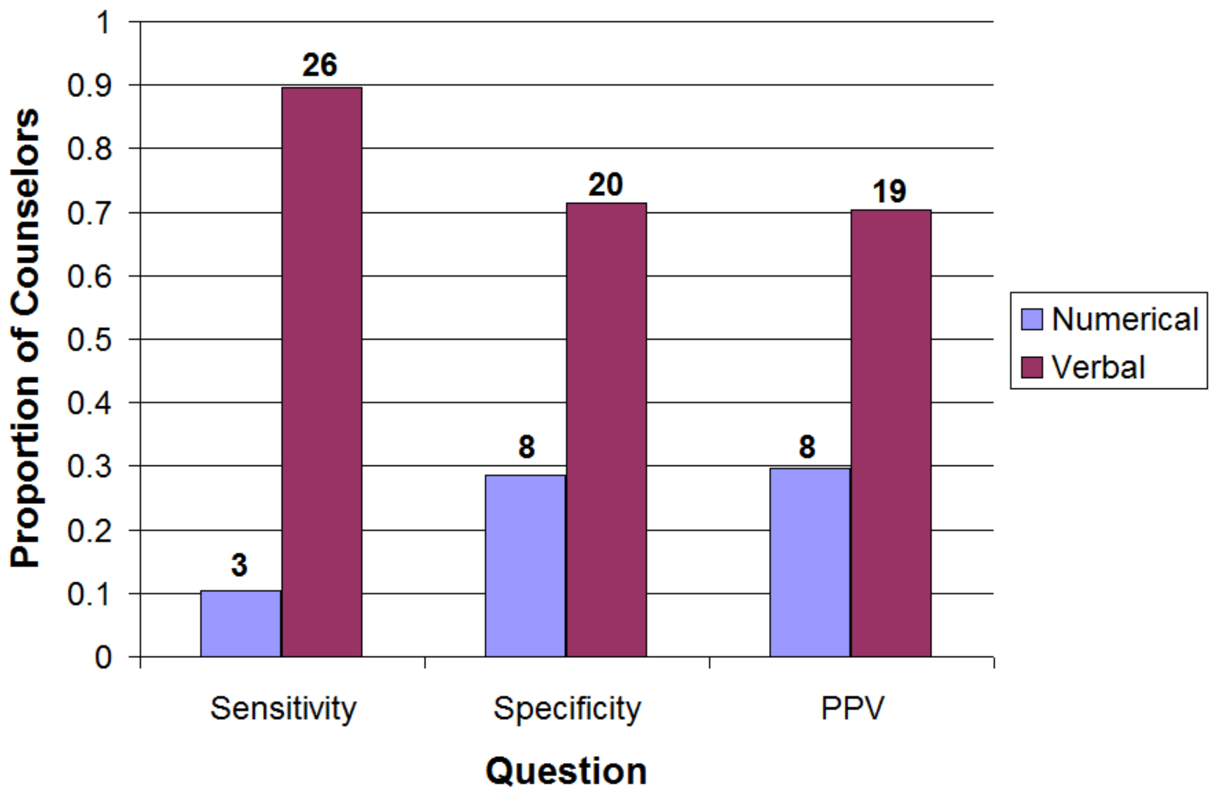


Table 2-2 Proportion of counselors using each format by question

			Format of Responses					
	n		Frequency		Percentages		Non-numerical	
Question	%	Freq.	%	Freq.	%	Freq.	%	Freq.
Window Period	100%	29/29	-	-	-	-	-	-
Sensitivity	100%	29/29	3%	1/29	7%	2/29	90%	26/29
Specificity	97%	28/29	4%	1/28	7%	2/28	93%	26/28
PPV	93%	27/29	0%	0/29	28%	8/29	66%	19/29
Prevalence	90%	26/29	42%	11/26	12%	3/26	50%	13/26

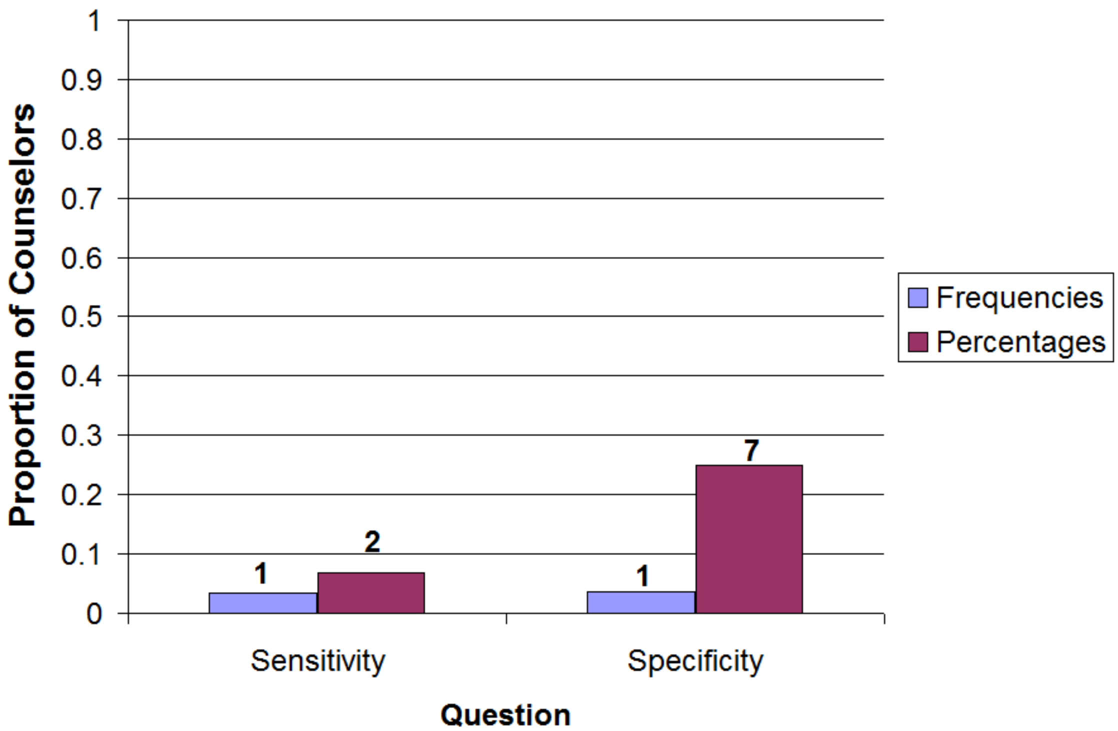
Note: This table does not distinguish between absolute and relative frequencies used by HIV hotline counselors.

Figure 2-2 Proportion of counselors using numerical and verbal expressions of risk for the sensitivity, specificity, and PPV questions



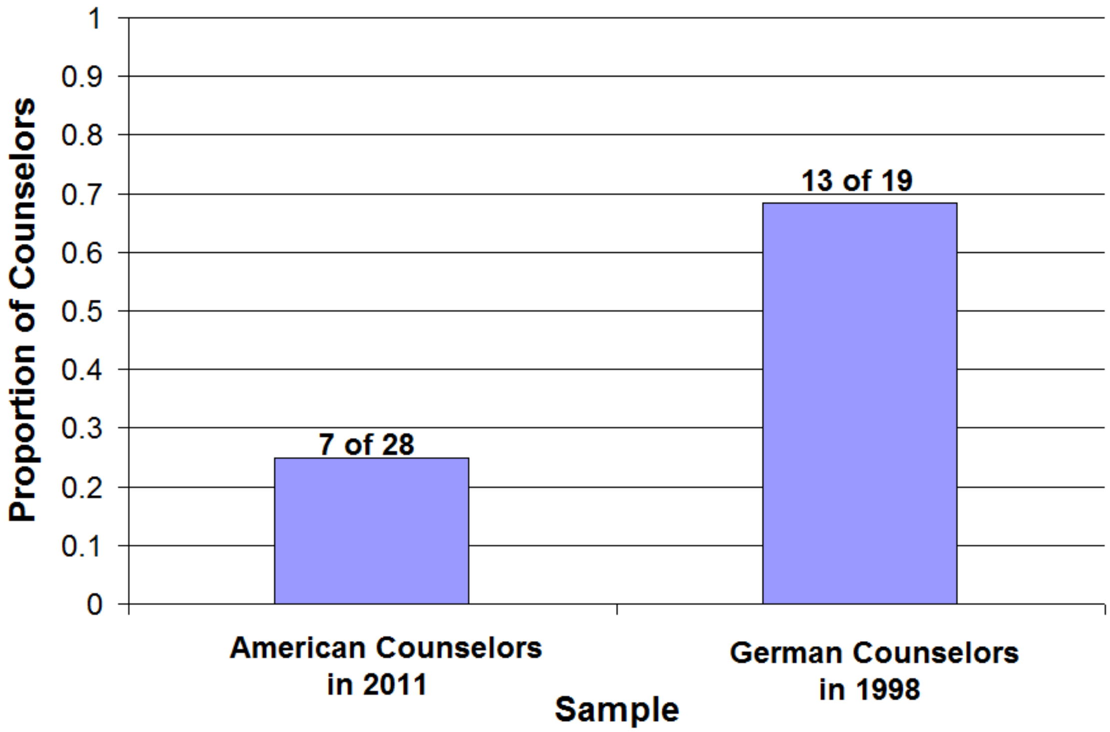
Note: The numerical values above the bars represent the counts of counselors using each format by question.

Figure 2-3 Proportion of counselors using frequencies and percentages for the sensitivity and specificity questions.



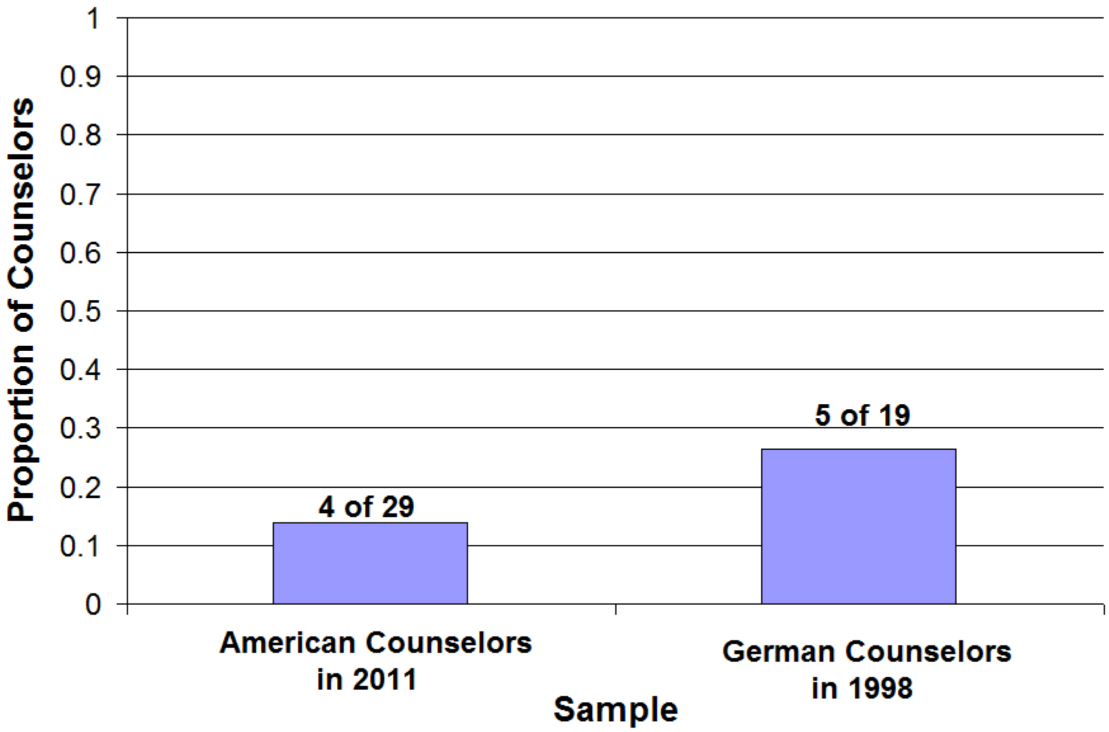
Note: The numerical values above the bars represent the counts of counselors using each format by question.

Figure 2-4 Proportion of counselors suggesting that false positives do not occur by sample



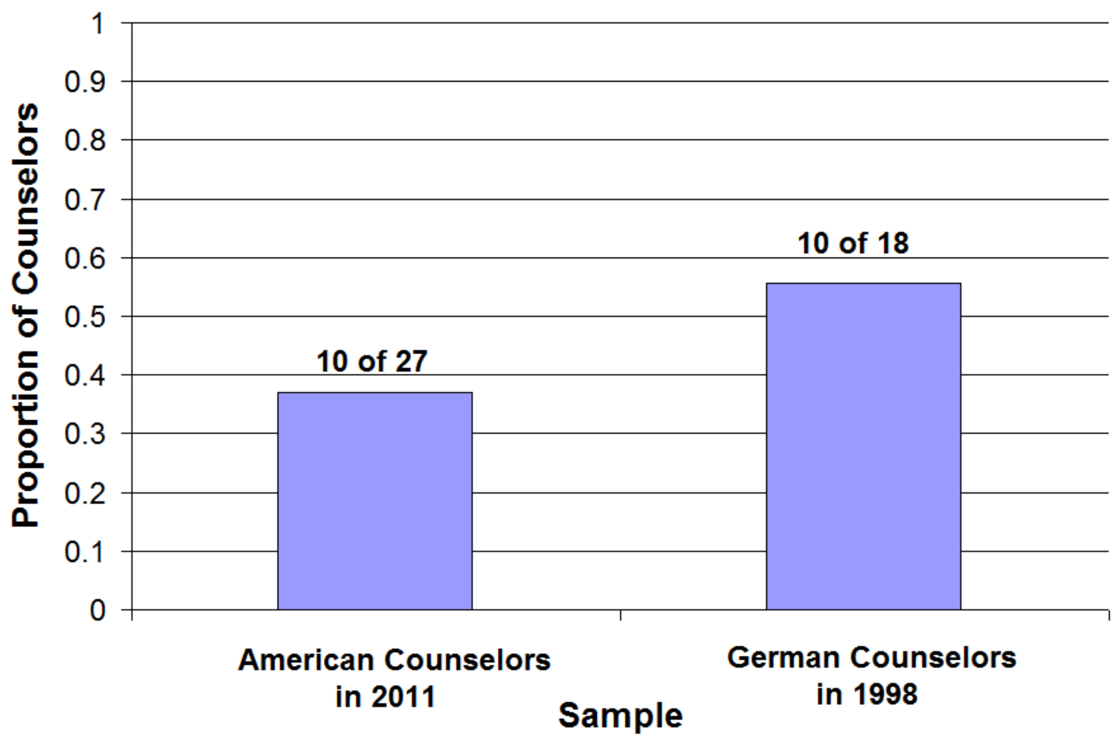
Note: The numerical values above the bars represent the counts and reference class size of counselors for each sample.

Figure 2-5 Proportion of counselors suggesting that false negatives do not occur by sample



Note: The numerical values above the bars represent the counts and reference class size of counselors for each sample.

Figure 2-6 Proportion of counselors confusing the sensitivity of the ELISA and the positive predictive value for low-risk individuals by sample



Note: The numerical values above the bars represent the counts and reference class size of counselors for each sample.

Chapter 3 - Experiment 2

Study 1 found that 2011 American HIV hotline counselors were more aware of the presence of false positive and false negative test results, and were less likely to confuse the sensitivity and PPV than the 1998 German sample; which is indication that research is influencing how counselors communicate to individuals. However, counselors (1) overwhelmingly use verbal expressions of risk to communicate HIV testing statistics; (2) use percentages more often than frequencies to report testing statistics of HIV tests to low risk patients; and (3) could not provide an accurate positive predictive value for low-risk women. Therefore, in order for patients to know the PPV, they would need to calculate the PPV from the other information that the counselors provided. The aim of Experiment 2 was to assess the ability of students to calculate the PPV from transcripts of counselors describing HIV statistics. A secondary aim was to investigate the influence of numeracy, interpretation, clarity of information, and confidence on the facilitating effect of natural frequencies.

Gigerenzer and colleagues (1998) found that German HIV counselors did not provide accurate HIV testing statistics to a low-risk patient. In the discussion of their findings, the researchers suggested ideal answers to the window period, sensitivity, specificity, prevalence, and positive predictive value questions in a naturally sampled frequency format. These suggested responses were used to create idealized transcripts of conversations with HIV counselors for the current study. One version of the transcript used the suggested natural frequency format. The other version of the transcript used the same outline as the natural frequency format, but contained probabilistic information in a probability format using normalized percentages.

Previous research has shown that performance on Bayesian reasoning tasks is related to numeracy such that the effect of the natural frequency format on performance is stronger in high numerate individuals (Chapman & Liu, 2009; Sirota & Juanchich, 2011) when measured using the General Numeracy Scale (GNS; Lipkus, Samsa, & Rimer, 2001). Additionally, Lipkus and colleagues (2001) suggest that the GNS measures the global construct of numeracy because even highly educated samples have difficulty with some of the problems. The GNS includes 11 questions, which involve mathematical operations and converting between numeric formats. This measure of numeracy has been criticized because individuals feel stress and frustration after completing a math test. Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry and Smith (2007) developed an alternative measure of numeracy that does not involve mathematical operations. The subjective numeracy scale (SNS) is based on eight Likert scale ratings on preference for numerical data. The creators of the SNS found that it correlates with the GNS and that participants report lower levels of stress and frustration after completing the SNS compared to the GNS. Additionally, participants took less time to complete the SNS versus the GNS (Fagerlin et al., 2007). This is an advantage of the SNS over the GNS because one purpose of these measures is to assess quickly participants' numeracy level in order to tailor messages of risk to their understanding. However, there is still a debate as to the definition of numeracy.

There are serious consequences associated with not clearly communicating risks to patients, including suicide (due to misinformation about false positive test results), overmedication (due to unclear dosing instructions), and uninformed decision making in trade-off decisions. The implications for communicating risks to low numerate patients are even more serious. Previous research has found that people with low numeracy are less willing to participate in decision making about health (Galesic & Garcia-Retamero, 2011), have less accurate perceptions of risk (Woloshin, Schwartz, Black, & Welch, 1999), and are less likely to recall health information, and understand graphical and textual depictions of risk (Zikmund-Fisher, Smith, Ubel, & Fagerlin, 2007). The current study used the General Numeracy Scale (Lipkus et al., 2001) and the Subjective Numeracy Scale (Fagerlin, et al., 2007) to measure numeracy. Measuring numeracy, both objectively and subjectively, allowed for comparisons of both as predictors of performance on the Bayesian reasoning task with ecologically valid risks.

Another variable that influences accuracy on Bayesian reasoning tasks is the interpretation of the numerical information. Brase (2008) gave participants one of three versions of a Bayesian reasoning task: a naturally sampled frequency format (e.g., 1 out of 10 people), a normalized chances format (e.g., a person with 1 out of 100 chances), and chances with natural sampling (e.g., a group of people with a 10% chance). Brase (2008) found that a frequency interpretation lead to a larger proportion of accurate participants than a single-event probability interpretation. This finding suggests that when information is presented in an ambiguous format the interpretation of frequencies facilitates performance. The current study aimed to extend these findings to the probability format.

Hypotheses

The aim of the second study was to investigate differences in performance on Bayesian reasoning task due to format, numeracy, interpretation, perceived clarity, and rated confidence. Previous research has looked at some of these variables in Bayesian reasoning, but not all of them together and not in a task with ecologically valid risks.

Format

Previous research has demonstrated the frequency effect in Bayesian reasoning tasks. Specifically, a naturally sampled frequency format facilitates Bayesian reasoning (Brase, 2008; Gigerenzer & Hoffrage, 1995; Gigerenzer, 1996; Gigerenzer & Edwards, 2003; Hoffrage & Gigerenzer, 1998). Therefore, the format hypothesis is that a larger proportion of individuals in the natural frequency format condition will correctly answer the PPV question than those in the probability format condition.

However, previous research has also found a floor effect on Bayesian reasoning tasks (Chapman & Liu, 2007). Due to the difficulty of the mathematical calculations using ecologically valid risks, the alternative format hypothesis is that the naturally sampled frequency format will not facilitate performance such that there will be no difference in the proportion correct on the Bayesian reasoning task between the natural frequency format condition and the probability condition.

Numeracy

Previous research has demonstrated that the effect of natural frequencies on performance of a Bayesian reasoning task depends on the level of numeracy (Chapman & Liu, 2007). Specifically, high numerates were found to outperform low numerates on a Bayesian reasoning task when the risks were presented in a naturally sampled frequency format. Additionally, the researchers found that there was no difference between high and low numerates on the probability version of the Bayesian reasoning task. This result suggests an interaction between format and numeracy, specifically that the frequency effect is stronger in high numerate individuals. Therefore, the numeracy hypothesis for the current study is that a larger proportion of high numerate individuals in the natural frequency format condition will correctly answer the PPV question than the proportion of low numerate individuals in the natural frequency format condition.

However, the same research found low performance in all conditions in a Bayesian reasoning task. In the current study, the Bayesian reasoning task may be more difficult due to the decimal answer resulting from calculating ecologically valid risks. Therefore, the alternative numeracy hypothesis for the current study is that performance on the Bayesian reasoning task will not depend on numeracy.

Interpretation

Previous research has demonstrated that when individuals interpret risk information in an ambiguous "chances" format as frequencies they are more accurate than when they interpret them as single-event probabilities (Brase, 2008). Therefore, the interpretation hypothesis for the current study is that a larger proportion of individuals who interpret the numerical information as frequencies will correctly answer the PPV question than the proportion of individuals who interpret the numerical information as single-event probabilities.

However, previous research on the effect of interpretation did not investigate the influence of interpretation on the probability format (Brase, 2008). Additionally, previous research did not find any additional influence of interpretation on the natural frequency format. Therefore, the alternative interpretation hypothesis is that performance on the Bayesian reasoning task will not depend on the interpretation of the numerical information such that there will be no additional influence of a frequency interpretation on performance.

Method

Participants

Participants were 198 college students recruited from an Introductory Psychology course through a Midwestern University's online recruitment system—Sona Systems (Sona Systems: Human Subject Pool Management Software, n.d.). The online recruitment system tracks participation, allows for online data collection, and is accessible for all student enrolled in the general psychology courses on campus. Sixty-two percent of participants were female (38% male) with ages ranging from 16-51 years old ($M = 19.80$, $SD = 3.78$). The majority of participants reported that they were Caucasian (85%) and that their first language was English (94%). All participants were treated in accordance with the "Ethical Principles of Psychologist Code of Conduct" (American Psychological Association, 2002).

Materials and Procedures

Participants completed the study online through the SONA system. First, participants viewed the informed consent information (see Appendix D). The consent form detailed the participant's role in the study and their rights as a participant, including the right to drop out of the experiment at any time without penalty.

Second, participants read a cover story (see Appendix E) that described a scenario of the participant's friend, Sarah. Sarah is obligated to take an HIV test and wants to know beforehand how likely it would be that she has HIV after receiving a positive test—the PPV. Sarah is a low risk individual who has no prior history of intravenous drug use and is in a monogamous relationship. Being a good friend, the participant calls an HIV hotline and gathers the necessary information to calculate the PPV of HIV tests for American women. This cover story was identical to the cover story used in Study 1 for the HIV hotline counselors. In both cover stories, the individual is required to take an HIV test due to pregnancy. Additionally, both cover stories suggest that the information on the Internet is confusing. Since the participants were not provided with the PPV from the transcript, they were required to calculate the PPV based on the information they were given (prevalence, sensitivity, and specificity) through idealized transcripts from counselors.

The transcriptions (see Appendix F) included information about the sensitivity, specificity, and prevalence of HIV for Sarah's risk group. Participants received one of two transcripts and answered Sarah's question on PPV. The difference between the two transcripts was the format of the numerical information. One transcript presented the numerical information in a probability format and the other transcript presented the numerical information in a natural frequency format. Gigerenzer and colleagues (1998) suggested idealized answers to the sensitivity, specificity and prevalence questions and these suggestions were adapted into the transcript that the participants received. After participants answered the PPV question, they were asked three follow-up questions (see Appendix G). They were asked to rate their confidence in their answers (on a 1-6 Likert scale), rate the clarity of the information contained in the transcript (on a 1-6 Likert scale), and indicate how they interpreted the information they were given (e.g., as frequencies or single event probabilities) using the same structure as Brase (2008).

Next, all participants completed the SNS (Fagerlin et. al, 2007) to assess participants' subjective numeracy. The SNS contains eight items that examine individual comfort and efficacy in using numerical information (see Appendix H). In addition to the SNS, participants completed the GNS (Lipkus et al., 2001) to assess participants' numeracy on an objective level. The GNS contains 11 questions that require individuals to do mathematical computation (see Appendix I). The SNS and the GNS are correlated (Fagerlin et al., 2007) and authors of the GNS suggest that the scale is measuring the global construct of numeracy (Lipkus et al., 2001).

After completing the numeracy scales, participants completed a task that was unrelated to the current study. After completing the online study, participants read a debriefing statement (see Appendix J) which explained the true purpose of the experiment and thanked participants for their participation. Participants were encouraged to contact the CDC for additional information regarding HIV testing.

Results

The researcher analyzed the data using a series of one-tailed difference of proportions tests, one-tailed independent samples t-tests, and a Pearson product-moment correlation with the alpha level set at .05. One-tailed tests were used because previous research has investigated these variables before, but not simultaneously with a Bayesian reasoning tasks with ecologically valid risks. There were 100 participants in the natural frequency condition and 96 participants in the probability condition. Twenty-one of the 196 participants (11%) indicated that they had some training in probabilities.

Format

The accuracy of responses was analyzed using a series of difference of proportion tests. Participants in the natural frequency condition (9/100; 9%) significantly outperformed individuals in the probability condition (0/96; 0%) on the PPV question, $z = 3.01$, $p = .001$, $\eta = .61$ (see Figure 3-1). When allowed a ± 5% error interval around the exact answer, individuals in the natural frequency format condition (12/100; 12%) significantly outperformed individuals in the probability format condition (3/96; 3%) on the PPV question, $z = 2.34$, $p = .010$, $\eta = 0.35$. Since none of the participants in probability format condition answered with the exact response to the PPV question, the remainder of the analyses utilized the performance scores including the ± 5% error interval.

Numeracy

An independent samples t-test was used to compare the levels of objective numeracy (GNS) between the natural frequency and probability conditions. There was no difference in participants objective numeracy total scores between the natural frequency condition ($M = 8.28$; $SD = 2.33$) and the probability condition ($M = 8.56$; $SD = 2.21$), $t(194) = 0.87$, $p = .430$. A median split was performed on the objective numeracy data to create two groups: high numerates ($n = 118$) and low numerates ($n = 78$). A series of difference of proportion tests were used to further probe the data. There was an overall effect of objective numeracy on performance (see Figure 3-2) such that high numerates (14/118; 12%) outperformed low numerates (1/78; 1%) on the PPV question ($z = 2.73$, $p = .003$, $\eta = 0.48$).

The data was divided into four groups for further comparisons: high numerates/natural frequency version ($n = 62$), high numerates/probability version ($n = 56$), low numerates/natural frequency version ($n = 38$), and low numerates/probability version ($n = 40$). High numerates in the natural frequency condition (11/62; 18%) outperformed high numerates in the probability condition (3/56; 5%) on the PPV question ($z = 2.13$, $p = .017$, $\eta = 0.41$). However, there was no difference in performance between low numerates in the natural frequency condition (1/38; 3%) and low numerates in the probability condition (0/40; 0%) on the PPV question ($z = 1.03$, $p = .152$, $\eta = 0.33$). Additionally, high numerates (11/62; 18%) outperformed low numerates (1/38; 3%) in the natural frequency condition ($z = 2.26$, $p = .013$, $\eta = 0.54$); however, there was no difference in performance between high numerates (3/56; 5%) and low numerates (0/40; 0%) in the probability condition ($z = 1.49$, $p = .070$, $\eta = 0.47$).

An independent samples t-test was used to compare the levels of subjective numeracy (SNS) between the natural frequency and probability conditions. There was no difference in participants average subjective numeracy scores between the natural frequency condition ($M = 4.21$; $SD = 0.83$) and the probability condition ($M = 4.15$; $SD = 0.90$), $t\ (194) = 0.45$, $p =.154$. A median split was performed on the subjective numeracy data to create two groups: high numerates ($n = 88$) and low numerates ($n =108$). A series of difference of proportions tests were used to further probe the data. There was no overall effect of subjective numeracy on performance (see Figure 3-3) such that high numerates (8/88; 9%) did not outperform low numerates (7/108; 6%) on the PPV question ($z = 0.68$, $p = .248$, $\eta = 0.10$).

The data was divided into four groups for further comparisons: high numerates/natural frequency version ($n = 47$), high numerates/probability version ($n = 41$), low numerates/natural frequency version ($n = 53$), and low numerates/probability version ($n = 55$). There was no difference in performance between high numerates in the natural frequency condition (6/47; 13%) and high numerates in the probability condition (2/41; 5%) on the PPV question ($z = 1.28$, $p = .101$, $\eta = 0.29$). However, low numerates in the natural frequency condition (6/53; 11%) outperformed low numerates in the probability condition (1/55; 2%) on the PPV question ($z = 2.01$, $p = .024$, $\eta = 0.42$). Additionally, there was no difference in performance between high numerates (6/47; 13%) and low numerates (6/53; 11%) in the natural frequency condition ($z = 0.22$, $p = .412$, $\eta = 0.04$); and, there was no difference in performance between high numerates (2/41; 5%) and low numerates (1/55; 2%) in the probability condition ($z = 0.85$, $p = .198$, $\eta = 0.17$).

The SNS averages and the GNS total scores were correlated using Person product-moment r. The two measures of numeracy were correlated, $r = .36$; which is lower than the correlation that Fagerlin and colleagues (2007) calculated during the creation of the SNS scale ($r = .47$). Additionally, high numerates outperformed low numerates in the natural frequency condition when measured objectively, which is consistent with previous research (Chapman & Liu, 2007). However, there was no difference between high numerates and low numerates on the natural frequency condition when measured subjectively (see Figure 3-4). This correlation, together with the different results from the GNS and SNS on performance, suggests that the GNS and SNS scales are possibly measuring related but significantly different constructs (see Table 3-1).

Interpretation

The data was split into three groups to compare performance across interpretations: frequency interpretation ($n = 112$), single-event probability interpretation ($n = 65$), and other interpretation ($n = 19$). Difference of proportions tests were used to further probe the data. There was an effect of interpretation on performance (see Figure 3-5) such that individuals with an other interpretation (3/19; 16%) outperformed individuals with a single-event probability interpretation (2/65; 3%) on the PPV question ($z = 2.06$, $p = .021$, $\eta = 0.46$). There were no other differences in performance due to interpretation. There was no significant difference in performance between individuals with a frequency interpretation (10/112; 9%) and individuals with a single-event probability interpretation (2/65; 3%) on the PPV question ($z = 1.49$, $p = .069$, $\eta = 0.25$), though there was a trend in the predicted direction. Additionally, there was no difference in performance between individuals with a frequency interpretation (10/112; 9%) and an other interpretation (3/19; 16%) on the PPV question ($z = 0.92$, $p = .178$, $\eta = 0.21$). However, the effect sizes suggest that a significant result may have been found with a larger sample size.

The data was further divided into four groups to compare performance between the single event probability interpretation and the frequency interpretation between the natural frequency and probability conditions (see Table 3-2): frequency interpretation/natural frequency condition (FF; $n = 68$), frequency interpretation/probability condition (FP; $n = 44$); single-event probability interpretation/natural frequency condition (SEF; $n = 21$); and single-event probability/probability condition (SEP; $n = 44$). No differences in performance were found between these groups. There was no difference in performance between the FF group (8/68; 12%) and the FP group (2/44; 5%) on the PPV question ($z = 1.31$, $p = .097$, $\eta = 0.27$); and there was no difference in performance between the SEF group (1/21; 5%) and the SEP group (1/44; 2%) on the PPV question ($z = 0.54$, $p = .294$, $\eta = 0.14$). Additionally, there was no difference in performance between the frequency interpretation (8/68; 12%) and single-event probability interpretation (1/21; 5%) in the natural frequency condition ($z = 0.93$, $p = .177$, $\eta = 0.26$); and, there was no difference between frequency interpretation (2/44; 5%) and single event probability interpretation (1/44; 2%) in the probability condition ($z = 0.59$, $p = .279$, $\eta = 0.13$). Again, the lack of significant differences may be due to the lack of power with such a low accuracy rate on the probability condition.

Confidence and Clarity

An independent samples t-test was used to compare the ratings of confidence between the natural frequency and probability conditions (see Table 3-3). There was no difference in participants' ratings of confidence between the natural frequency condition ($M = 3.44$; $SD = 1.43$) and the probability condition ($M = 3.79$; $SD = 1.54$) regarding the transcript information, $t(194) = 1.66$, $p = .627$. A median split was performed on the confidence data to create two groups of participants: high confidence ($n = 108$) and low confidence ($n = 88$). Difference of proportions tests were used to further probe the data. There was no overall effect of confidence on performance (see Figure 3-6) such that high confidence individuals (8/108; 7%) did not outperform low confidence individuals (7/88; 8%) on the PPV question ($z = 0.14$, $p = .443$, $\eta = 0.02$).

The data was divided into four groups for further comparisons: high confidence/natural frequency version ($n = 50$), high confidence/probability version (n =

58), low confidence/natural frequency version ($n = 50$), and low confidence/probability version ($n = 38$). High confidence individuals in the natural frequency condition (7/50; 14%) outperformed high confidence individuals in the probability condition (1/58; 2%) on the PPV question ($z = 2.43$, $p = .008$, $\eta = 0.50$). However, there was no difference in performance between low confidence individuals in the natural frequency condition (5/50; 10%) and low confidence individuals in the probability condition (2/38; 5%) on the PPV question ($z = 0.81$, $p = .209$, $\eta = 0.18$). Additionally, there was no difference in performance between high confidence individuals (7/50; 14%) and low confidence individuals (5/58; 9%) in the natural frequency condition ($z = 0.89$, $p = .189$, $\eta = 0.17$). Also, there was no difference in performance between high confidence individuals (1/58; 2%) and low confidence individuals (2/38; 5%) in the probability condition ($z = 0.97$, $p = .166$, $\eta = 0.20$).

An independent samples t-test was used to compare the ratings of clarity between the natural frequency and probability conditions. There was no difference in participants' ratings of clarity between the natural frequency condition ($M = 3.82$; $SD = 1.40$) and the probability condition ($M = 4.06$; $SD = 1.40$) regarding the transcript information, $t\ (194) = 1.22$, $p =.896$. A median split was performed on the clarity data to create two groups of participants: high clarity ($n = 116$) and low clarity ($n =80$). Difference of proportions tests were used to further probe the data. There was no overall effect of clarity on performance (see Figure 3-7) such that the high clarity group (8/116; 7%) did not outperform low clarity group (7/88; 8%) on the PPV question ($z = 0.29$, $p = .387$, $\eta = 0.04$).

The data was divided into four groups for further comparisons: high clarity/natural frequency version ($n = 56$), high clarity/probability version ($n = 60$), low clarity/natural frequency version ($n = 44$), and low clarity/probability version ($n = 36$). The high clarity/natural frequency group (7/56; 13%) outperformed high clarity/probability group (1/60; 2%) on the PPV question ($z = 2.30$, $p = .012$, $\eta = 0.46$); however, there was no difference in performance the low clarity/natural frequency group (5/44; 11%) and the low clarity/probability group (2/36; 6%) on the PPV question ($z = 0.91$, $p = .182$, $\eta = 0.21$). Additionally, there was no difference in performance between the high clarity individuals (7/56; 13%) and low clarity individuals (5/44; 11%) in the natural frequency condition ($z = 0.17$, $p = .431$, $\eta = 0.04$). Also, there was no difference between high clarity individuals (1/60; 2%) and low clarity individuals (2/36; 6%) in the probability condition ($z = 1.06$, $p = .146$, $\eta = 0.22$).

Discussion

The purpose of the second study was to assess the ability of students to calculate the PPV from idealized transcripts of conversations with HIV counselors. The format hypothesis was supported such that there was an effect of the natural frequency format on performance of the Bayesian reasoning task; however, in support of the alternative format hypothesis, there was a floor effect in that no participants in the probability condition answered the PPV question correctly as 52.6%. When allowed a ± 5% error interval, three participants in the probability format condition answered the PPV question correctly. Accuracy was low in the natural frequency condition as well, 9 answered with 89 out of 188 and an additional three gave an answer within the ±5% interval. Due to the low accuracy rate, these results are consistent with the observation that performance on Bayesian reasoning tasks is generally low (Chapman & Liu, 2007). Also, these results are consistent with the research that found that a naturally sampled frequency format facilitates performance on Bayesian reasoning tasks (Brase, 2008; Gigerenzer & Hoffrage, 1995; Gigerenzer, 1996; Gigerenzer & Edwards, 2003; Hoffrage and Gigerenzer, 1998). The current research suggests that performance on a Bayesian reasoning task with ecologically valid risk is facilitated by a natural frequency format. Future research should investigate the frequency effect in Bayesian reasoning tasks of ecologically valid risks with a more demographically diverse sample to generalize these findings to the larger population.

The numeracy hypothesis was partially supported such that performance on the Bayesian reasoning task depended on level of objective numeracy, but not subjective numeracy. In the creation of the SNS, the researchers suggested that the SNS and the GNS were correlated and therefore measuring the same construct (Fagerlin et al., 2007). The current study found a correlation between the two measures, but the additional data showed that the GNS predicted performance on a Bayesian reasoning task, whereas the SNS did not. Additionally, further probing of the data showed different patterns of results between the two measures of numeracy. Specifically, high numerates outperformed low numerates in the natural frequency condition using the GNS measure of numeracy, but there was no difference in performance between high numerates and low numerates in the natural frequency condition using the SNS measure of numeracy. This data is consistent with Chapman and Liu's findings that the frequency effect is stronger in high numerate individuals when the objective measure of numeracy is employed for comparisons. However, the current research failed to extend these findings to the standard subjective measure of numeracy. Future research should compare performance on other health numeracy tasks using objective and subjective measures of numeracy.

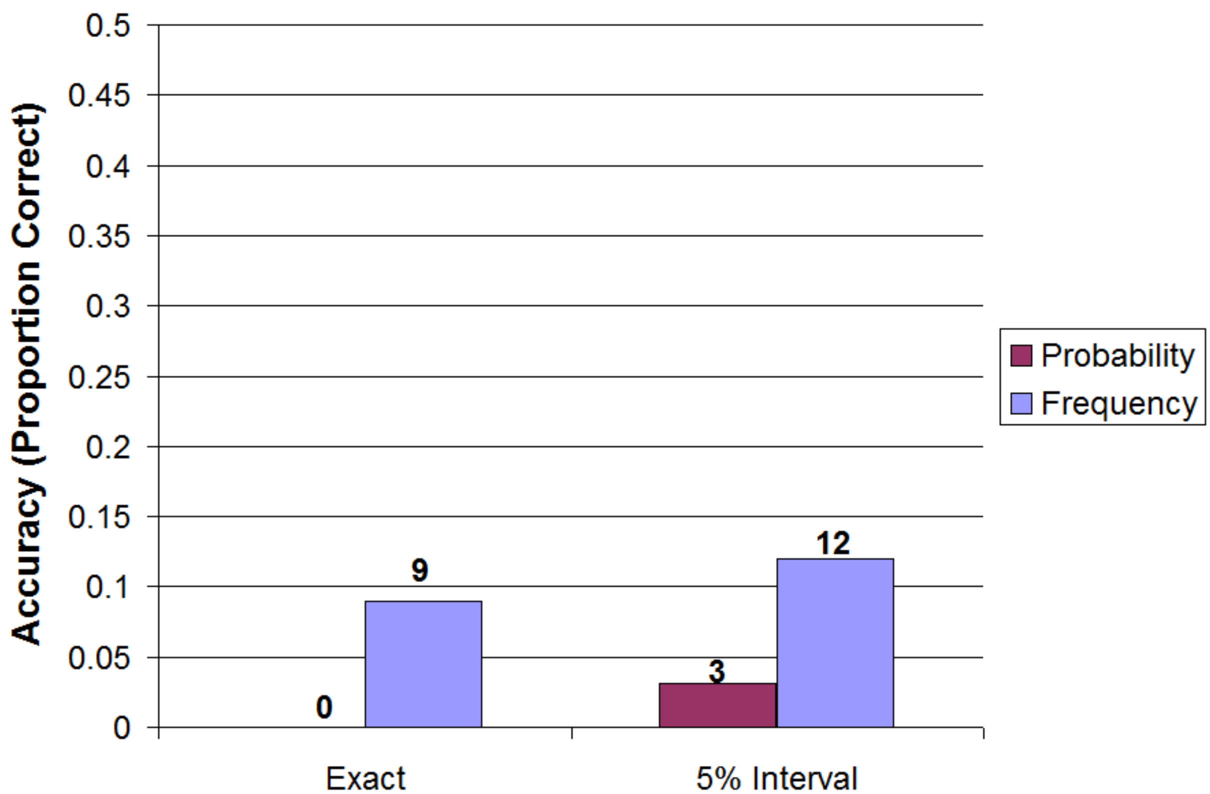
The data failed to support the interpretation hypothesis and supported the alternative interpretation hypothesis such that there was no overall significant effect of interpretation on performance. However, the trend of the data was in the predicted direction such that individuals who interpreted the transcript information as frequencies were more likely to answer the PPV question correctly than individuals who interpreted the information as single-event probabilities. Previous research suggests that when an ambiguous numerical format is used, the interpretation of frequencies facilitates performance (Brase, 2008). In the current study, there was no facilitating effect of interpretation on either the probability format or the natural frequency format. However, there was similar effect sizes between the current study and previous research. This suggests that with enough power and a sufficient sample size that significant effects may have been found in the current data. Alternatively, these findings may suggest that the format is driving the effect of interpretation in Brase's study. Future research should investigate the effect of interpretation using multiple formats in additional Bayesian reasoning tasks and health-related, non-Bayesian reasoning tasks.

Additionally, it was found that there was no additional influence of rated confidence or perceived clarity on performance in a Bayesian reasoning task. Participants who rated their confidence as high did not outperform participants who rated their confidence as low on the Bayesian reasoning task. However, high confidence individuals in the natural frequency condition did outperform high confidence individuals in the probability condition. Although, this effect may be due to format rather than confidence. This suggests that confidence is not a determining factor in Bayesian reasoning tasks. Future research should investigate the role of confidence in performance on other health numeracy tasks. Additionally, there was no overall effect of perceived clarity on performance of the Bayesian reasoning task. Interestingly, there was no difference in ratings of clarity between the natural frequency and probability versions of the Bayesian reasoning task. This suggests that the perception of the difficulty of the problem may not be different between naïve individuals. Future research should look at within-subject effects of format on ratings of clarity for Bayesian reasoning tasks.

Overall, Study 2 found low accuracy on a Bayesian reasoning task with ecologically valid risks, but that a natural frequency formatted version of the task facilitated performance. Additionally, objective numeracy but not subjective numeracy, predicted performance on the Bayesian reasoning task such that high numerates outperformed low numerates. On the other hand, interpretation of numerical format, perceived clarity of the transcripts, and rated confidence in one’s answer did not influence performance on a Bayesian reasoning task with ecologically valid risks. Future research should investigate the influence of interpretation, perceived clarity and rated confidence on Bayesian reasoning tasks with different risks. The Bayesian reasoning task in the current study used ecologically valid risks, which result in a decimal answer. Other Bayesian reasoning tasks have round answers and poor performance in the probability version of the current task may be attributed to the difficulty of the calculation with decimal percentages.

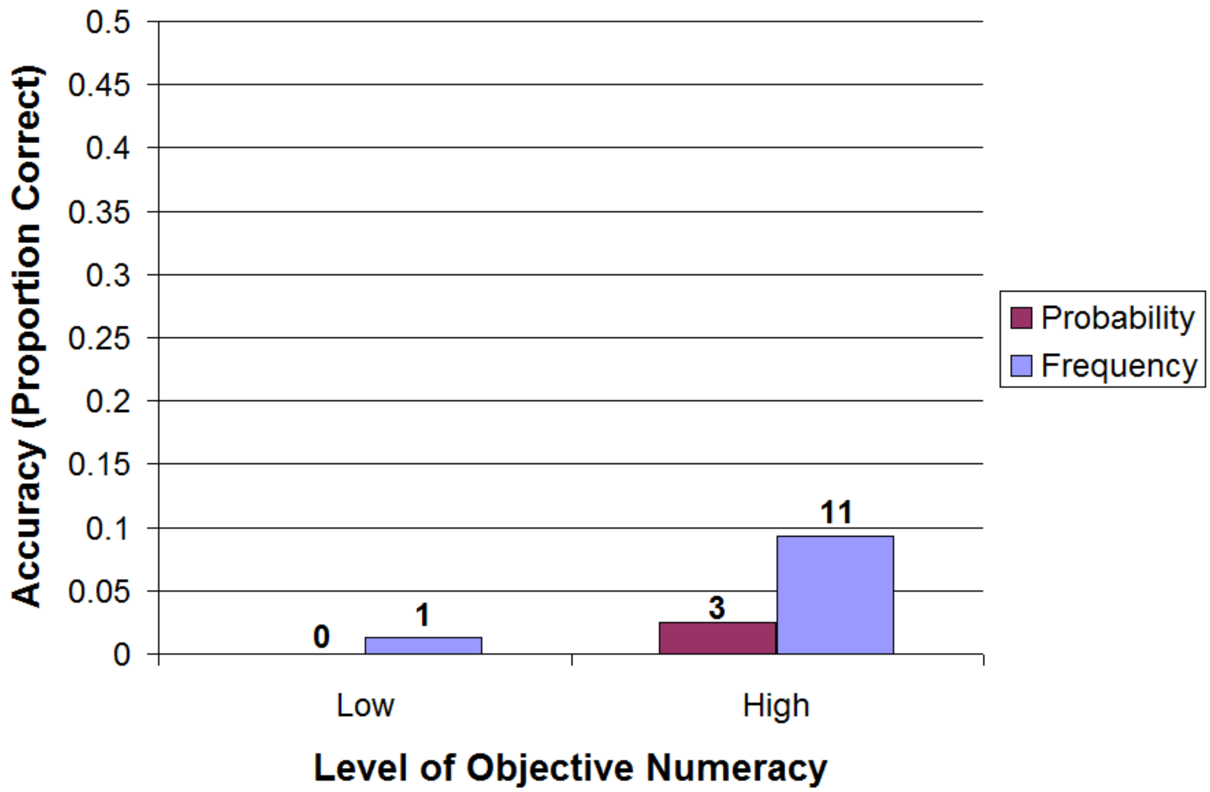
Figures and Tables

Figure 3-1 Proportion of correct answers on PPV question by condition and accuracy level



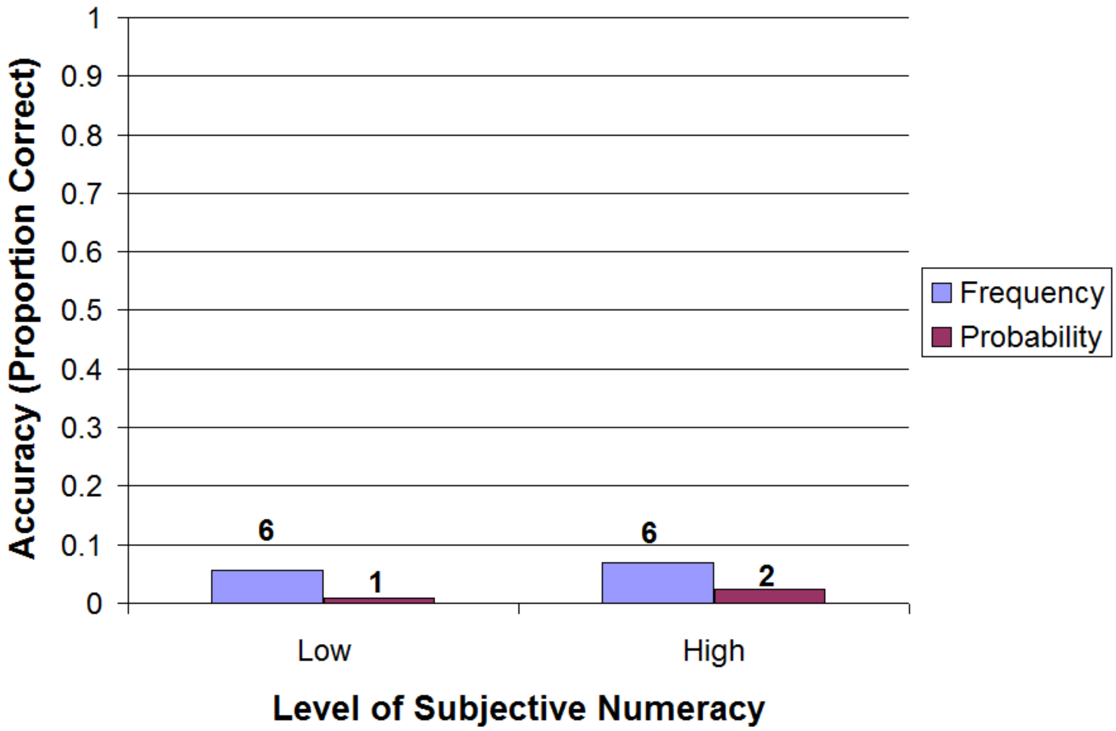
Note: The numerical values above the bars represent the counts of accurate participants by each level of accuracy.

Figure 3-2 Proportion of correct answers on PPV question by condition and level of objective numeracy



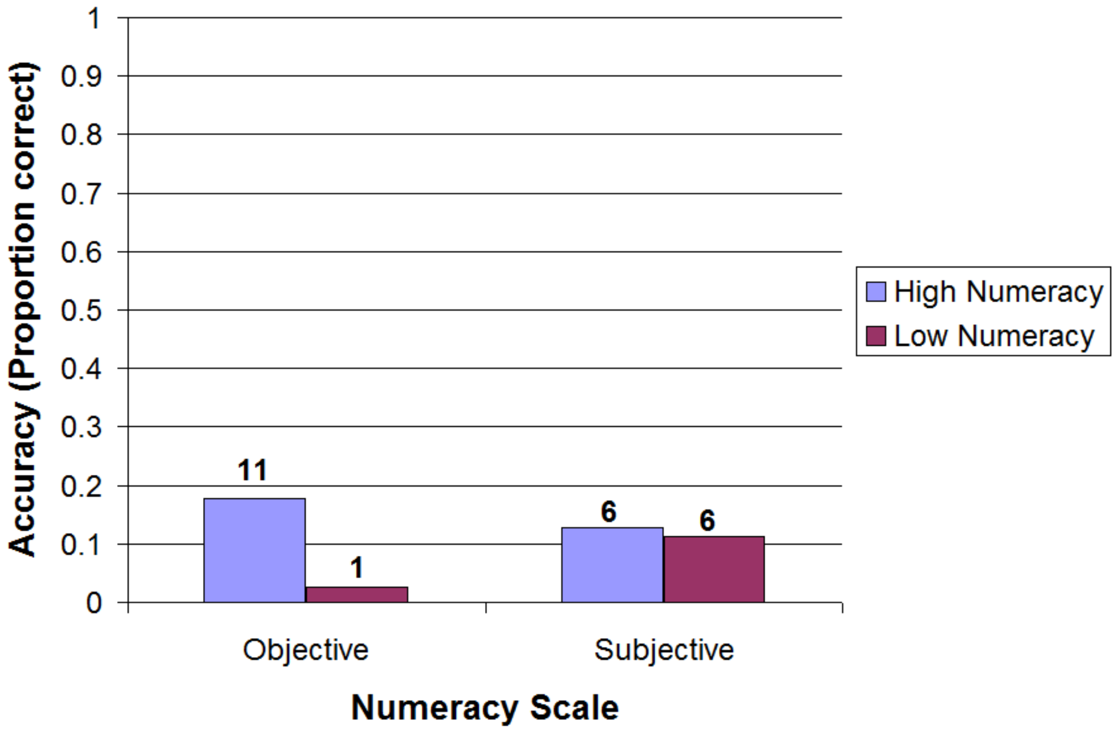
Note: The numerical values above the bars represent the counts of accurate participants by format and level of objective numeracy.

Figure 3-3 Proportion of correct answers on PPV question by condition and level of subjective numeracy



Note: The numerical values above the bars represent the counts of accurate participants by format and level of subjective numeracy.

Figure 3-4 Proportion of correct responses on natural freqeuncy version of PPV question by level of numeracy and numeracy scale

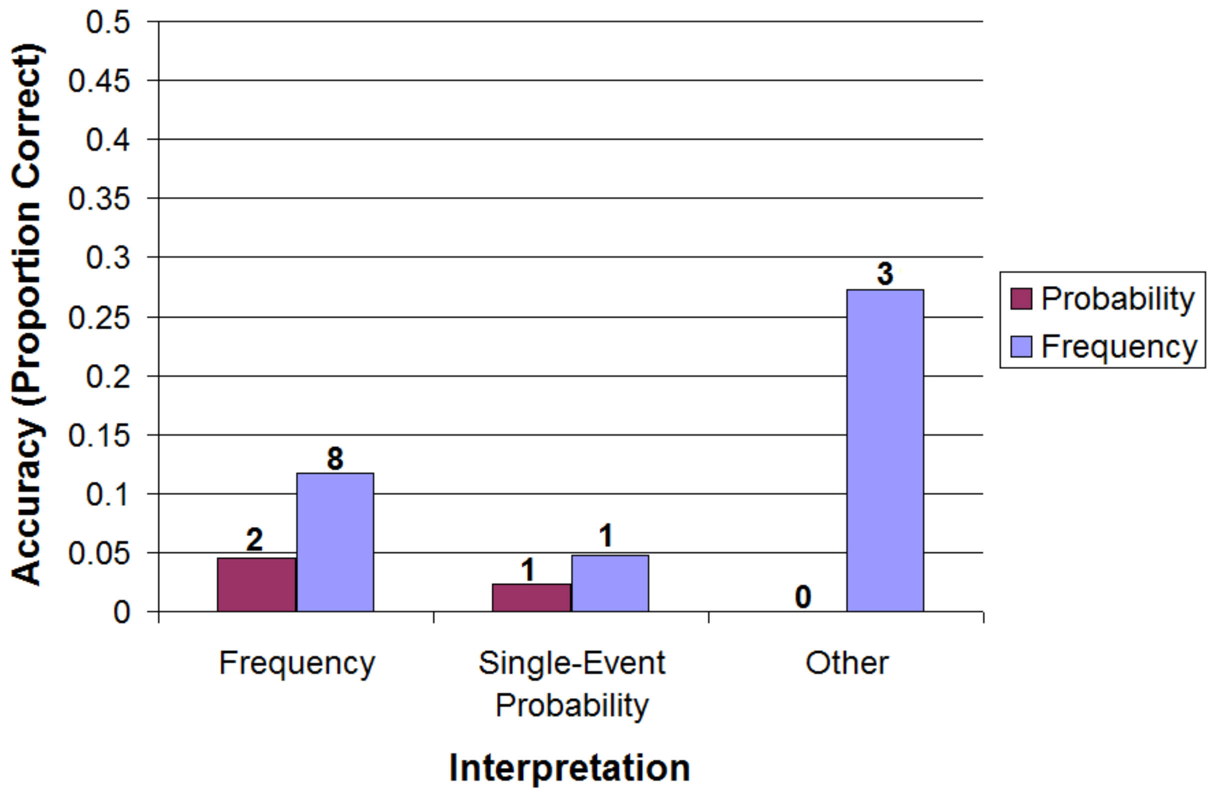


Note: The numerical values above the bars represent the counts of accurate participants by level of numeracy and condition.

Table 3-1 Proportion of accurate responses by level of numeracy and condition

			Numeracy							
	n		GNS High		GNS Low		SNS High		SNS Low	
Presentation Format	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.
Natural Frequency	51%	100/196	18%	11/62	3%	1/38	13%	6/47	11%	6/53
Probability	49%	96/196	5%	3/56	0%	0/40	5%	2/41	2%	1/55

Figure 3-5 Proportion of correct answers on PPV question by condition and interpretation



Note: The numerical values above the bars represent the counts of accurate participants by format and interpretation.

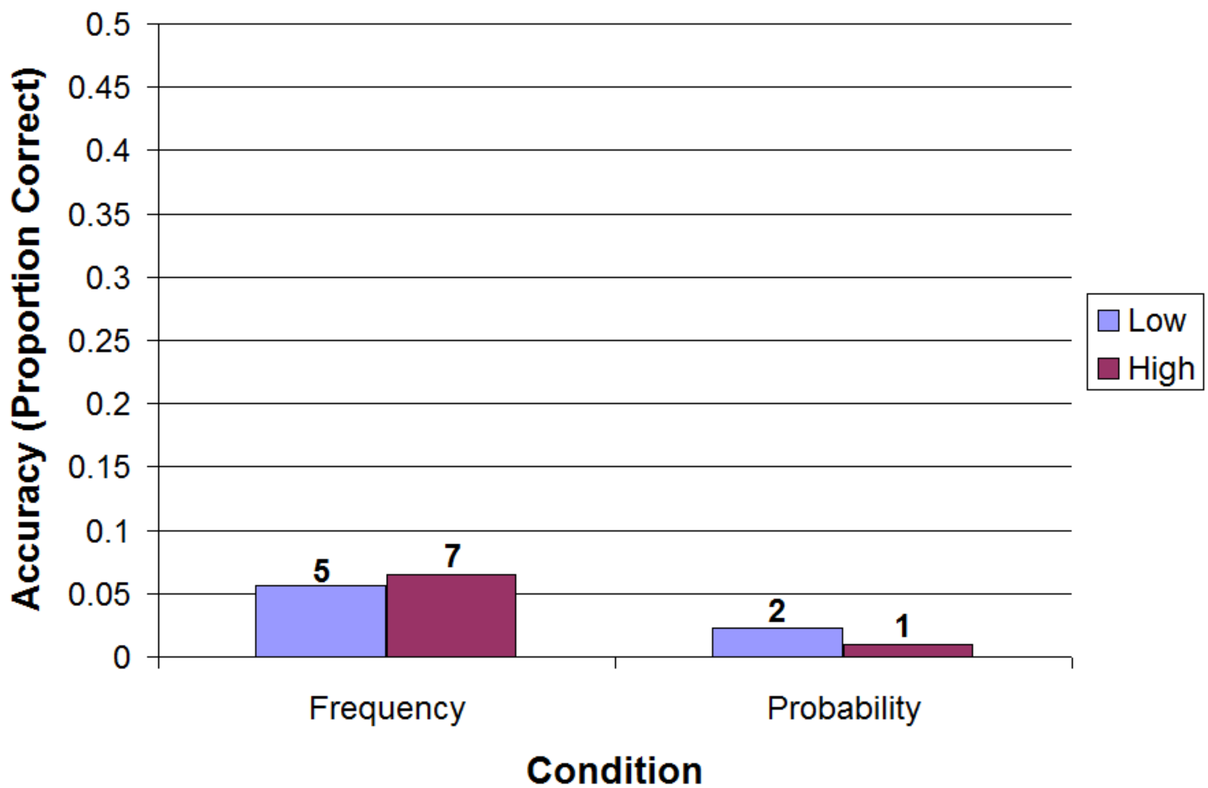
Table 3-2 Proportion of accurate responses by interpretation and condition

			Interpretation					
	n		Frequency		Single Event		Other	
Presentation Format	%	Freq.	%	Freq.	%	Freq.	%	Freq.
Natural Frequency	51%	100/196	68%	68/100	21%	21/100	11%	11/100
Probability	49%	96/196	46%	44/96	46%	44/96	8%	8/96

Table 3-3 Proportion of accurate responses by condition, level of confidence, and perceived clarity

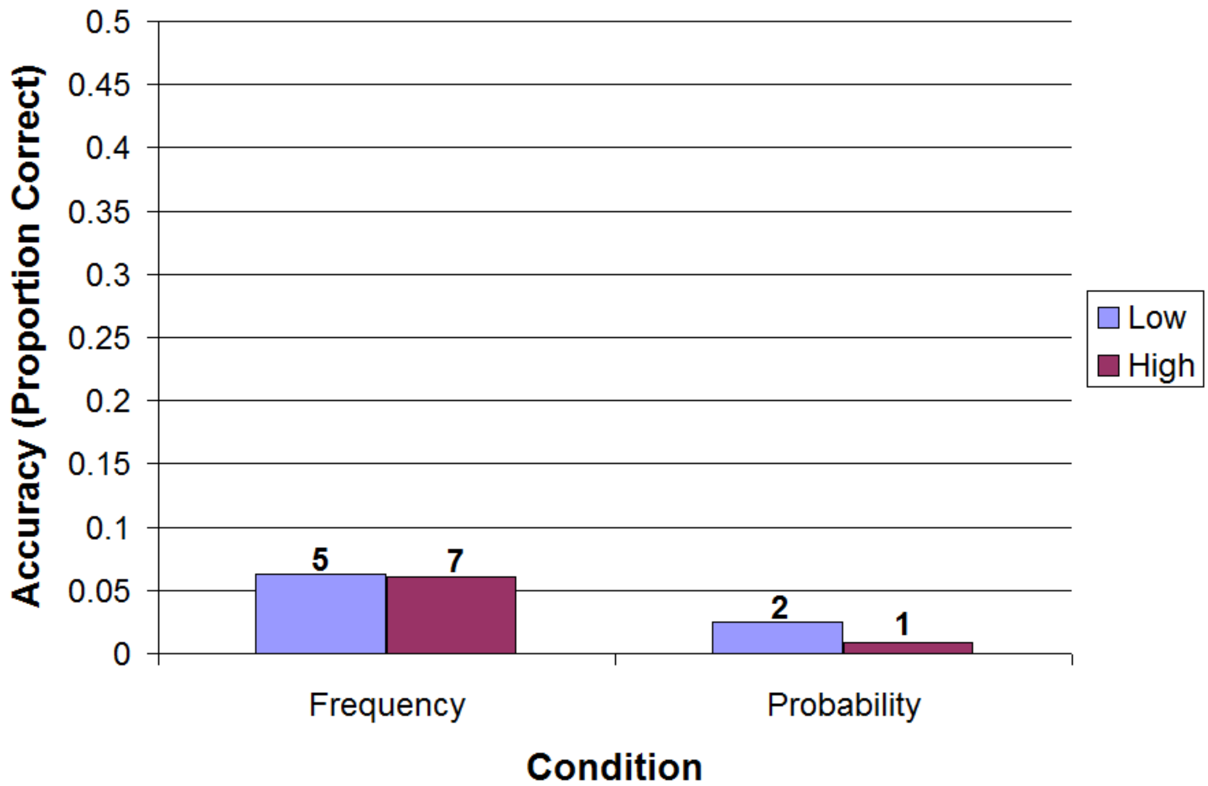
			Confidence				Clarity			
	n		High		Low		High		Low	
Presentation Format	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.
Natural Frequency	51%	100/196	14%	7/50	10%	5/50	13%	7/56	11%	5/44
Probability	49%	96/196	2%	1/58	5%	2/38	2%	1/60	6%	2/36

Figure 3-6 Proportion of correct answers on PPV question by condition and level of confidence



Note: The numerical values above the bars represent the counts of accurate participants by format and level of confidence.

Figure 3-7 Proportion of correct answers on PPV question by condition and level of rated clarity



Note: The numerical values above the bars represent the counts of accurate participants by format and level of rated clarity.

Chapter 4 - General Discussion

The aim of the first study was to investigate the state of communication between HIV hotline counselors and low risk individuals about HIV testing statistics. The results from the first study indicate that counselors use verbal expressions of risk to communicate testing statistics, use probabilities more than frequencies to communicate risk, and erroneously believe that false negatives and false positives do not occur. Nevertheless, these results are actually moving in the right direction relative to the previous study conducted over 13 years ago in Germany. American HIV hotline counselors in 2011 were less likely than the 1998 German sample to suggest that false positives and false negatives do not occur.

One thing that should be noted is that the full time job of these HIV hotline counselors might not be to inform patients of individual risks. HIV hotlines are available to the public to ask all types of questions about HIV testing and the responses of the counselors indicated that they were not accustomed to answering questions about the HIV tests, but were more accustomed to providing resources to individuals with HIV. It is not surprising that counselors had a difficult time providing individualized risks due to the lack of individualized models of risk. HIV is not the only disease that does not have an individualized model of risk—in fact, there are very few models of individualized risk and this information may not be readily available to these counselors. The trend in health care is in the direction of shared/informed decision making between doctors and patients. For patients to be fully informed they should be aware of their individualized risks and should be able to get this information from knowledgeable health professionals. Future research should investigate the different resources detailing individualized risks that are available for doctors, nurses, counselors and other health professionals.

The aim of the second study was to investigate the influence of format, numeracy, interpretation, confidence and rated clarity on performance of an ecologically valid, health-related (HIV risk) Bayesian reasoning task. The first study found that counselors could provide accurate statistics on the sensitivity, specificity, and prevalence; however, no counselors could provide an accurate positive predictive value for low-risk women. Therefore, in order for individuals to know their PPV they need to be able to calculate it from the statistics that the counselors can provide (sensitivity, specificity, and PPV). When students were given idealized transcripts of conversations with HIV hotline counselors, very few were able to compute the PPV from either the natural frequency format or the probability format. No participants were able to compute the exact PPV from the probability version of the transcript. The results suggest that the frequency effect is present in Bayesian reasoning tasks with real world risks; and that high numerates (objectively measured) are more sensitive to the frequency effect in Bayesian reasoning tasks with real world risks than low numerates (objectively measured). Additionally, it was found that subjective numeracy, interpretation, rated confidence, and perceived clarity had no influence on the frequency effect.

Even with a facilitating effect of a naturally sampled frequency format, overall performance on the Bayesian reasoning task was very low. This finding is consistent with previous findings of a floor effect in Bayesian reasoning tasks (Chapman & Liu, 2007). Participants in the second study were young college students—an arguably well-educated sample. For the general population with an average intelligence lower than college students, performance on Bayesian reasoning tasks may be even lower because of difficulty or could be higher due to motivation. Studies on perceptions of risk with participants of a general intelligence level indicated that low numerate individuals tend be less accurate in estimating risk than high numerate individuals (Schapira et al., 2001). Additionally, research has shown that patients preferred multiple graphical depictions of risk (Bayesian trees, bar graphs, and icons) rather than a single depiction of risk (Gottlieb, Weiss, & Chapman, 2007), but these graphical depictions are still rated as confusing (Shapira et al., 2001). This suggests that low numerate individuals may benefit from not only multiple textual formats of risk but also multiple graphical formats of risks as well. Future research should investigate the ability of the general population to perform Bayesian reasoning tasks and the influence of numeracy on performance with graphical and textual representations of risk.

Conclusions

Clear communication of risks between doctors and patients is necessary for shared/informed decision-making. The CDC has recommended that more individuals be tested for HIV. This is increasing the amount of low-risk individuals who are being tested for HIV. Low-risk individuals are more likely to receive a false positive test result than high risk individuals. Previous research has shown that counselors do not communicate the relationship between risk factors and false positive tests to low-risk individuals (Gigerenzer et al., 1998). The current study found that some counselors gave accurate answers to the sensitivity, specificity and prevalence questions; however, no counselors were able to accurately report the PPV of HIV tests for low-risk individuals. Therefore, for individuals to know their conditional risk of HIV they will need to calculate it from the information that the counselors can provide. There are two ways to calculate the PPV—from either natural frequency information or normalized percentages. Previous research has repeatedly demonstrated that presenting information in terms of naturally sampled frequencies facilitates computation of conditional risk (Brase, 2008; Chapman & Liu, 2009; Gigerenzer & Hoffrage, 1995; Gigerenzer, 1996; Gigerenzer & Edwards, 2003; Hoffrage & Gigerenzer, 1998) and comparisons of risk (Miron-Shatz, et al., 2009). Yet, the current research found that counselors use verbal expressions of risk to communicate risk to low risk patients even though research has found that patients prefer numerical expressions versus verbal expressions when computing risks (Vahabi, 2010). It has been shown that Bayesian reasoning is relatively simple to teach and that a little teaching results in long lasting differences in performance on Bayesian reasoning tasks (Sedlmeier & Gigerenzer, 2001). In order for doctor-patient communication to be effective, doctors and other health practitioners should know how to communicate risks to patients of different risk levels and education levels and do so. Additionally, probabilistic reasoning should be taught to children at a young age. This research is an example of how math can be useful in a real world problem. If children can understand Bayesian reasoning when information is presented in naturally sampled frequencies, it indicates that health information should be communicated in a natural frequency format. Additionally, if health professionals have a better understanding of the risks, they will be able to communicate individualized risks to patients effectively.

References

American Psychological Association. (2002). American Psychological Association ethical principles of psychologists and code of conduct. Retrieved September 27, 2011, from http://www.apa.org/ethics/code/index.aspx

Baker, D. W., Williams, M. V., Parker, R. M., Gazmarian, J. A., & Nurss, J. (1999). Development of a brief test to measure functional health literacy. *Patient Education and Counseling, 38,* 33-42.

Bhattacharyaa, R., Bartonb, S., & Catalanc, J. (2008). When good news is bad news: Psychological impact of false positive diagnosis of HIV. *AIDS Care, 20,* 560-564.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review, 15*, 284-289.

Center for Disease Control and Prevention (2006). *New estimates of U.S. HIV prevalence, 2006.* Retrieved October 20, 2010, from http://www.cdc.gov/hiv/topics/surveillance/resources/factsheets/pdf/prevalence.pdf

Center for Disease Control and Prevention (2007). *HIV counseling with rapid tests.* Retrieved August 8, 2010, from http://www.cdc.gov/hiv/topics/testing/resources/factsheets/print/rt_counseling.htm

Chapman, G. B. & Liu, J. (2009). Numeracy, frequency, and bayesian reasoning. *Judgment and Decision Making, 4,* 34-40.

Cohen, J. (1988). Differences between proportions. In, *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*, 672-680.

Galesic, M. & Garcia-Retamero, R. (2011). Do low-numeracy people avoid shared decision making? *Health Psychology, 30,* 336-341.

George, J. R. (1998). Alternative specimen sources: Methods for confirming positives. *Proceedings of Conference on the Laboratory Science of HIV, USA*, 47-60.

Ghosh, K., Crawford, B. J., Pruthi, S., Williams, C. I., Neal, L., Sandhu, N. P., Johnson, R. E., Wahner-Roedler, D., Britain, M. K., Cha, S. S. & Ghosh, A. K. (2008). Frequency format diagram and probability chart for breast cancer risk communication: A prospective, randomized trial. *BMC Women's Health, 8*, n.p.

Gigerenzer, G. & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, *327*, 741-744.

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

Gigerenzer, G. (1996). The Psychology of Good Judgment: Frequency Formats and Simple Algorithms, *Medical Decision Making*, *16*, 273–280.

Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). Aids counseling for low-risk clients. *AIDS Care*, *10*, 197-211.

Girotto, V. & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition, 78,* 247-276.

Golbeck, A. L., Ahlers-Schmidt, C. R., Paschal, A. M., & Dismuke, S. E. (2005). A definition and operational framework for health numeracy. *American Journal of Preventative Medicine, 29,* 375-376.

Gonzalez-Vallejo, C. C., Erev, I., & Wallsten, T. S. (1994). Do decision quality and preference order depend on whether probabilities are verbal or numerical? *The American Journal of Psychology, 107*, 157-172.

Gottlieb, D. A., Weiss, T., & Chapman, G. B. (2007). The format in which uncertainty information is presented affects decision biases. *Psychological Science, 18,* 240-246.

Hoffrage, U. & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73,* 538-540.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84,* 343-352.

Hunt, E. (2007). *The mathematics of behavior.* Cambridge, NY: Cambridge University Press.

Johnson, J. A. (2008). *AIDS funding for federal government programs: FY1981-FY2009*. (Order Code RL30731). Washington, DC: Congressional Research Service.

Krantzler, N. (2007). *HIV facts* [Brochure]. USA: ETR Associates.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37-44.

Miron-Shatz, T., Hanoch, Y., Graef, D., & Sagi, M. (2009). Presentation format affects comprehension and risk assessment: The case of prenatal screening. *Journal of Health Communication, 14,* 439-450.

Neff, S. & Goldschmidt, R. (2011). Centers for disease control and prevention 2006 human immunodeficiency virus testing recommendations and state testing laws. *Journal of the American Medical Association, 305,* 1767-1768.

Proudfoot, L. (2008). *Incredible HIV facts* [Brochure]. USA: ETR Associates.

Schapira, M. M., Nattinger, A. B., & McHorney, C. A. (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Medical Decision Making, 21,* 459-467.

Schettler, J. (2005). *HIV after the test* [Brochure]. USA: ETR Associates.

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

Sirota, M., & Juanchich, M. (2011). Role of numeracy and cognitive reflection in bayesian reasoning with natural frequencies. *Studia Psychologica, 53,* 151-161.

Sona Systems: Human Subject Pool Management Software. (n.d.). Retrieved April 26, 2012 from Sona Systems: http://www.sona-systems.com/

Vahabi, M. (2010). Verbal versus numerical probabilities: Does format presentation of probabilistic information regarding breast cancer screening affect women's comprehension? *Health Education Journal, 69,* 150-163.

Woloshin, S., Schwartz, L. M., Black, W. C., & Welch, H. G. (1999). Women's perceptions of breast cancer risk: How you ask matters. *Medical Decision Making, 19,* 221-229.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Journal of Experimental Psychology*, *50*, 97-106.

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., Fagerlin, A. (2007). Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making, 27,* 663-671.

Appendix A - Two Calculations of the Positive Predictive Value using Probabilities and Frequencies

Information for Probability Equation

Sensitivity = 99% or 0.99

Prevalence = 1% or 0.01

Specificity = 99.1% or 0.991

Calculation of Probability Equation

$$\text{PPV} = \frac{(sensitivity)(prevalence)}{(sensitivity)(prevalence) + (1 - specificity)(1 - prevalence)}$$

$$\text{PPV} = \frac{(0.99)(0.01)}{(0.99)(0.01) + (1 - 0.991)(1 - 0.01)}$$

$$\text{PPV} = \frac{0.0099}{0.0099 + (0.009)(0.99)}$$

$$\text{PPV} = \frac{0.0099}{0.0099 + 0.00891} = \frac{0.0099}{0.01881} = 0.526 \text{ or } 52.6\%$$

Information for Natural Frequency Equation

Sample: 10,000 people

Base rate: 1%

HIV positive: 100 people

HIV negative: 9900 people

True Positive: 99 people

False Positive: 89 people

True Negative: 9811 people

False Negative: 1 person

Calculation of Natural Frequency Equation

PPV = $\frac{TP}{TP + FP}$

PPV = $\frac{99}{99 + 89}$ = $\frac{99}{188}$ = 0.526 or 52.6%

Appendix B - Glossary of Statistical Terms

Base rate – (or prior probability) refers to the probability in a population of a hypothesis that is unconditional of additional evidence

ELISA – Enzyme-linked immunosorbent assay is a blood test for HIV that detects antibodies in response to the HIV virus

False negative (FN) – a test result that indicates that you have a disease or disorder when you do not have the disease or disorder

False positive (FP) – a test result that indicates that you do not have a disease or disorder when you do have the disease or disorder

Frequencies – a numerical format used to convey probabilistic information that focuses on the occurrences of events (e.g. 200 people, 4 out of 5 counselors)

Naturally sampled frequencies – frequencies gathered by natural sampling (e.g., 27 out of 346 tests) "In a naturally sampled population, natural frequencies are obtained by counting individuals according to their features (e.g. disease versus not disease, positive test result versus negative test result)" (p. 345, Hoffrage, Gigerenzer, Krauss, & Martignon, 2002).

Positive predictive value (PPV) – (or posterior probability) the probability that an individual has a disease or disorder given that they have received a positive test result

Posterior Probability – (or PPV) in Bayesian reasoning it refers to the probability of a hypothesis in a population given the evidence

Prevalence – the proportion of individuals with a disease or disorder at a given point in time within a population or risk group

Prior Probability – (or base rate) in Bayesian reasoning refers to the probability of a hypothesis in a population that is unconditional of additional evidence

Probabilities – a numerical format used to convey probabilistic information that divides the number of occurrences by the sample and multiplies that result by 100 (e.g., 99%, 2.5%).

Sensitivity – is a measure of the performance of a test that describes the ability of the test to accurately identify individuals with a certain disease or disorder (1-sensitivity is the FP rate). The sensitivity is equal to TP/(TP+FN).

Specificity – is a measure of the performance of a test that describes the ability of the test to accurately identify individuals without a certain disorder (1- specificity is the FN rate). Specificity is equal to TN/(TN+FP).

Appendix C - Interview Questions for HIV Hotline Counselors

Counselor: HIV Hotline—how can I help you?

Researcher: Hi, I recently took an HIV test and I had some questions. I took the test because I am thinking about becoming pregnant. I was looking online for information about HIV testing and the information was confusing so I wanted to talk to a real person, can you help me?

Counselor: <<response>>

Researcher (Window Period): So, how long do you need to wait between the last possible time you could have gotten HIV and the test before HIV can be detected?

Counselor: <<response>>

Researcher (PPV): How likely is it for women with low risk to have HIV after a positive test?

Counselor: <<response>>

Researcher (Specificity): If I don't have HIV is it possible for me to test positive? How likely would it be for low-risk women?

Counselor: <<response>>

Researcher (Sensitivity): If I do have HIV is it possible for me to test negative? How likely would it be for low-risk women?

Counselor: <<response>>

Researcher (Prevalence): How likely is it for a woman with low risk to have HIV?

Counselor: <<response>>

Researcher: That is all the questions I had for you. I appreciate it. Have a nice day.

Counselor: <<response>>

Appendix D - Informed Consent Form

Participation in this study will involve reading scenarios and answering questions based on the material you have read. Please work hard and answer the questions to the best of your ability. You may use a blank piece of paper to help you with any calculations.

Please do not use a calculator.

There are no anticipated risks associated with this study. The study should take no more than 30 minutes; and in exchange for your participation, you will receive 0.50 research credits. The alternative assignment for Psych 110 students is a critical review and analysis of a psychology journal article.

In addition to experimental credit, you will learn about topics in applied cognitive psychology and judgment and decision making research. The field of psychology may benefit from your participation in this study.

If you are under the age of 18 and wish to participate in this study, you must attain parent/guardian consent before continuing. Please let the researcher know that you are under the age of 18.

Your information will remain anonymous. We are interested in patterns of responses across many people. The responses you provide today will be combined with responses provided by other participants. In no way can your responses be linked back to you. This consent form and the study sign-up sheet will be the only records of your participation in this study.

If you have any questions or concerns about this study, please contact Dr. Gary Brase at (785) 532-0609 or gbrase@ksu.edu. You may also contact Rick Scheidt, Committee on Research Involving Human Subjects, 203 Fairchild Hall, Kansas State University, Manhattan, KS 66506, (785) 532-3224, who can explain the rights of research participants if you have any questions or complaints.

TERMS OF PARTICIPATION: I understand this project is research, and that my participation is voluntary. I also understand that if I decide to participate in this study, I may withdraw my consent at any time, and stop participating at any time without explanation, penalty, or loss of benefits, or academic standing to which I may otherwise be entitled.

Appendix E - Cover Story

Your friend Sarah is pregnant and is required to take an HIV test. She has been reading information on the internet about different HIV tests, but is starting to get really confused about what the testing statistics mean.

She has been reading about the prevalence of HIV among women, and about the sensitivity and specificity of HIV tests. She understands that the <u>prevalence</u> of any disease is the number of cases of that disease within a population. She also understands that the <u>specificity</u> of a test is the ability of a test to correctly identify individuals *without* the disease; whereas the <u>sensitivity</u> of a test is the ability of the test to correctly identify individuals *with* the disease. She also has figured out that tests are usually not perfect, there are usually some people who test positive for a disease even when they do not have it (that is, specificity is not 100%), and there are usually some people who test negative for a disease even when they really do have it (that is, sensitivity is not 100%).

However, Sarah is unsure how to put all this information together for her situation taking an HIV test. *If Sarah receives a positive test result, how likely is it that she does have HIV?*

This seems like a straightforward question. You have some free time on your hands so you decide to help your friend Sarah out. You call up your local HIV hotline to find the answer to her question. Unfortunately, the hotline cannot talk to you about Sarah's exact test results. However, you are able to calculate the likelihood that Sarah will have HIV after a positive test result from the information the counselor gives you. The transcript of the conversation you had with the counselor is on the next page and contains the necessary information to answer Sarah's question:

Appendix F - Transcript Conditions: Questions and Answers

Natural Frequency Version

Question: How many women are living with HIV in the United States?

Frequency Answer: About 100 out of 10,000 American women have HIV.

Question: What is the sensitivity of HIV tests?

Frequency Answer: Of the 100 American women with HIV, 99 will receive a positive test.

Question: What is the specificity of HIV tests?

Frequency Answer: Of the 9,900 American women without HIV, 89 will receive a positive test.

Frequency PPV Question: If Sarah were to receive a positive test, how likely is it that she does indeed have HIV? **______ out of ______**

Probability Version (using Percentages)

Question: How many women are living with HIV in the United States?

Probability Answer: About 1% of American women are infected with HIV.

Question: What is the sensitivity of HIV tests?

Probability Answer: 99% of American women with HIV will receive a positive test.

Question: What is the specificity of HIV tests?

Probability Answer: 99.1% of American women without HIV will receive a negative test.

Probability PPV Question: If Sarah were to receive a positive test, how likely is it that she does indeed have HIV? **___________**

Appendix G - Follow-up Questions on Bayesian Reasoning Task

1. How confident are you in your calculation of the likelihood that Sarah has HIV after she receives a positive test result? (circle one number)

1	2	3	4	5	6
Not at all Confident					Extremely Confident

2. How clear is the information contained in the transcripts from the HIV counselor? (circle one number)

1	2	3	4	5	6
Not at all Clear					Extremely Clear

3. Which of the following best describes how you thought about the information and reached your answer to this problem? (check one)

- ☐ I thought about the information as a large number of people, either with or without HIV, and some getting positive test results and some getting negative test results

- ☐ I thought about the information as a single person, with some possibility of having HIV or not having HIV and some possibility of getting a positive or negative test result.

- ☐ Other: I thought about the information as: ______________________________

Appendix H - Subjective Numeracy Scale

1. How good are you at working with fractions?

1	2	3	4	5	6
Not at all Good					Extremely Good

2. How good are you at working with percentages?

1	2	3	4	5	6
Not at all Good					Extremely Good

3. How good are you at calculating a 15% tip?

1	2	3	4	5	6
Not at all Good					Extremely Good

4. How good are you at figuring out how much a shirt will cost if it is 25% off?

1	2	3	4	5	6
Not at all Good					Extremely Good

5. When reading the newspaper, how helpful do you find tables and graphs that are parts of a story?

1	2	3	4	5	6
Not at all Helpful					Extremely Helpful

6. When people tell you the chance of something happening, do you prefer that they use words ("it rarely happens") or numbers ("there's a 1% chance")?

1	2	3	4	5	6
Always Prefer Words					Always Prefer Numbers

7. When you hear a weather forecast, do you prefer predictions using percentages (e.g., “there will be a 20% chance of rain today”) or predictions using only words (e.g., “there is a small chance of rain today”)?

1	2	3	4	5	6
Always Prefer Percentages					Always Prefer Words

8. How often do you find numerical information to be useful?

1	2	3	4	5	6
Never					Very Often

Appendix I - General (Objective) Numeracy Scale

1. Imagine that we flip a fair coin 1,000 times. What is your best guess about how many times the coin would come up heads in 1,000 flips? ___________

2. In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize is 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket to BIG BUCKS? ___________

3. In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car?

____

4. Which of the following numbers represents the biggest risk of getting a disease? *(please check one)*

___ 1 in 100, ___ 1 in 1000, ___ 1 in 10

5. Which of the following numbers represents the biggest risk of getting a disease? *(please check one)*

___ 1%, ___ 10%, ___ 5%

6. If Person A's risk of getting a disease is 1% in ten years, and person B's risk is double that of A's, what is B's risk? ___________

7. If Person A's chance of getting a disease is 1 in 100 in ten years, and person B's risk is double that of A's, what is B's risk? ___________

8. If the chance of getting a disease is 10%, how many people would be expected to get the disease:

A: Out of 100? ___________

B: Out of 1000? ___________

9. If the chance of getting a disease is 20 out of 100, this would be the same as having a ___________% chance of getting the disease.

10. The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected? ___________

Appendix J - Debriefing Statement

Thank you for participating in this study. The purpose of the first task is to assess the current state of communication from counselors to students about HIV testing information. The transcripts contained in this experiment are based on real conversations with HIV hotline counselors. Counselors used probabilities (98%) more than frequencies (98 out of 100) to communicate risk information. It is predicted that calculation of the positive predictive value (the probability of having a disease after a positive test result) is made easier when information is in terms of frequencies rather than probabilities. Additionally, the positive predictive values indicates the relationship that lower risk individuals. The aim of the second task was to investigate the possible ways that people think about the Monty Hall Dilemma and the Bertrand Box Paradox, which are two classic brainteasers. The researchers appreciate your time and effort in this experiment.

Thank you for your participation. Please let the researcher know if you have any questions.