Numeracy
Advancing Education in Quantitative Literacy

Volume 8 | Issue 2 Article 7

2015

Numeracy Infusion Course for Higher Education (NICHE), 2: Development of Students' Bayesian Reasoning Skill

Frank Wang
LaGuardia Community College, CUNY, fwang@lagcc.cuny.edu

Recommended Citation

Wang, Frank. "Numeracy Infusion Course for Higher Education (NICHE), 2: Development of Students' Bayesian Reasoning Skill." *Numeracy* 8, Iss. 2 (2015): Article 7. DOI: http://dx.doi.org/10.5038/1936-4660.8.2.7

Numeracy Infusion Course for Higher Education (NICHE), 2: Development of Students' Bayesian Reasoning Skill

Abstract

Gerd Gigerenzer's technique of frequency representations for solving the medical diagnosis problem, mammography problem, and other Bayesian reasoning problems is summarized in this paper. Such a method has been introduced to community college students in an elementary statistics course. With repeated practice, many community college students can acquire the skill and avoid reported judgment errors that are commonly committed by medical professionals. However, weaknesses in basic skills such as percentage calculations prevent some students from obtaining the correct probability.

Keywords

cognitive illusion, Bayesian reasoning, medical diagnosis problem, mammography problem

Cover Page Footnote

Frank Wang is Professor of Mathematics at LaGuardia Community College of the City University of New York (CUNY). His research interests include general relativity and dynamical systems.

Introduction

This paper is the second of two papers in this issue on the Numeracy Infusion Course for Higher Education (NICHE) project[1] to promote quantitative reasoning (QR) across disciplines at the City University of New York (CUNY). NICHE is a predominantly online course to train faculty to (1) articulate QR learning goals; (2) create a QR lesson to help students achieve these goals; and (3) develop an instrument to assess student learning. Following enrollment in NICHE, every instructor will teach a QR-infused course to implement the instructional materials that they developed during the faculty training program. In the first paper (Wang and Wilder 2015), we outline the content of a NICHE unit on the phenomena of human intuition leading to faulty judgment, and how to apply cognitive science to QR instruction. This paper describes an effort to use NICHE material to teach the medical diagnosis problem, and more broadly Bayesian reasoning (to be explained shortly), in an elementary statistics course at a community college. Assessment results of student learning from 2013 to 2014 will be reported. Psychological research on Bayesian reasoning, particularly the method of solving the medical diagnosis and mammography problems using the natural frequency representation, provides the backdrop for the student learning activities described in the paper.

Interpreting Medical Test Results: Bayesian Reasoning

The medical diagnosis problem was first published in the *New England Journal of Medicine* (Casscells et al. 1978).

> *If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?*

This is a standard Bayesian reasoning problem, in which the probability of a cause (e.g., disease) has to be inferred from an observed effect (e.g., a positive result). This problem could be approached formally using Bayes' rule, which is a simple mathematical formula shown below. However, Casscells et al. found that only 11 of the 60 subjects (consisting of Harvard Medical School faculty, staff and students) gave the correct probability, 2%. Twenty-seven respondents vastly overestimated the probability by answering 95%. Many other studies show that

[1] NSF TUES 1121844.

clinicians miscalculated the probability in similar problems (Eddy 1982; Hoffrage and Gigerenzer 1998; Gigerenzer 2002).

The medical diagnosis problem, and more broadly Bayesian inference problems, have become the subject of an intense debate among psychologists (Gigerenzer and Hoffrage 1995; Kahneman and Tversky 1996; Gigerenzer 1996). To understand the debate, one needs to know the different interpretations of probability. Although probability as a mathematical axiomatic theory is well established, its interpretation is still an unsettled issue (McGrayne 2011). The "frequentists" interpret probability as a measure of how frequently the event will occur when the experiment is continually repeated. On the other hand, the "Bayesians" regard probability as a subjective measure of belief (hence the name "subjective probability"). The Bayesians allow the assignment of probabilities to a single event (e.g., the outcome of a single toss of a coin, or the probability of having breast cancer after a positive mammography), which is considered meaningless for the frequentists. The paper by Kahneman and Tversky (1974) and follow-up studies suggest that probability theory is systematically counter-intuitive. However, some psychologists argued that if the same problem is expressed in terms of frequencies rather than probabilities, people are more accurate at estimating the probability (Gigerenzer and Hoffrage 1995; Cosmides and Tooby 1996). The disagreement among statisticians and philosophers about the interpretation of probability is in part responsible for the disagreement among psychologists (Kahneman and Tversky 1996; Gigerenzer 1996). Despite the lack of consensus on the cause of biases in assessing subjective probability, the method of obtaining probability from frequency representation has been reported to be successful in training professionals in solving medical and legal problems (Hoffrage and Gigerenzer 1998; Sedlmeier and Gigerenzer 2001; Gigerenzer 2002).

The probability of having a disease given a positive medical test result is a conditional probability, which can be formally approached using Bayes' rule,

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

where the symbols H and D are for the hypothesis (e.g., disease) and data obtained (e.g., positive result), respectively. We refrain from further elaborating this abstract formula, as a mathematical approach can be found in any probability textbook (e.g., Ross 2014). Some researchers contend that Bayesian reasoning does not necessarily mean inserting probabilities into the formula; they suggest that Bayesian reasoning can occur naturally. To appreciate such a point of view, consider this example (Gill et al. 2005). Patient 1 is an obese 72 year old man with poorly controlled hypertension, and Patient 2 is a 28 year old, 44 kg, non-

smoking, vegan woman who competes regularly in triathlons. Patient 1 was rushed to the emergency after crushing chest pressure, and Patient 2 came in because she felt dizzy after running 20 km in hot weather. Both have an abnormal electrocardiogram. Logically, clinicians would suspect a heart attack for Patient 1, and consider such a diagnosis very unlikely for Patient 2. Bayesian reasoning takes the prior probability (patients' background) into account, and the posterior probability of heart attack diagnosis (after electrocardiogram) can be obtained from the formula, whereby "the pre-test odds of a hypothesis being true multiplied by the weight of new evidence (likelihood ratio) generates post-test odds of the hypothesis being true" (Gill et al. 2005).

The failure of Bayesian reasoning reported in the psychological literature is often attributed to research subjects' "base rate neglect," the condition that the prior probability is ignored or significantly underweighted (Bar-Hillel 1980). However, Cosmides and Tooby (1996) presented evidence that people are "good" at Bayesian reasoning when they are given frequencies as input and asked for frequencies as output. They argue that instead of saying 8%, it is easier for the human mind to consider "8 out of 100 people" or "80 out of 1000 people," based on their research on evolutionary theory. They rewrote the medical diagnosis problem by Casscells et al. (1978) as the following.

> *1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive (i.e., the "true positive" rate is 100%). But sometimes the test also comes out positive when it is given to person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease (i.e., the "false positive" rate is 5%).*
>
> *Imagine that we have assembled a random sample of 1000 Americans. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people.*
>
> *Given the information above:*
>
> *On average,*
>
> *How many of these 1000 people will have the disease?*
>
> *How many of the 1000 people will have the disease AND test positive for it?*
>
> *How many of the 1000 people will be healthy AND test positive for the disease?*
>
> *How many of the 1000 people will test positive for the disease, whether they have the disease or not?*
>
> *How many people who test positive for the disease will actually have the disease? ___ out of ___*

Cosmides and Tooby show that this frequency representation of the medical diagnosis problem elicited correct Bayesian reasoning from 92% of subjects and helped to eliminate base rate neglect.[2]

Similarly, Gerd Gigerenzer and Ulrich Hoffrage (1995) performed an experiment on the "mammography problem." The standard probability format reads:

> *The probability of breast cancer is 1% for a woman at age forty who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?*

Eddy (1982) reported that 95 out of 100 physicians estimated the probability to be between 70% and 80%, much greater than the correct answer, 7.8%. But Gigerenzer and Hoffrage discovered that by using the frequencies, more students and physicians could obtain the correct probability for the mammography problem in accordance with Bayes' rule (Gigerenzer and Hoffrage 1995). They represented the problem as the following.

> *10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ___ out of ___*

In his book *Calculated Risk*, Gigerenzer (2002) summarizes and clarifies his findings and presents a systematic method to guide clear thinking of subjective probability. (The method can be applied to many other situations. For example, Gigerenzer used it to analyze the relationship between wife battering and spousal murder in the United States to refute Alan Dershowitz's claim in the O. J. Simpson case.) The key is to think in terms of natural frequencies, simple counts of events, rather than in more abstract notations of percentages, odds, or probabilities. (Incidentally, Steven Strogatz (2012) reported that a few of his Cornell students discovered this trick independently and would avoid using Bayes'

[2] A *Numeracy* reviewer pointed out that if subjects are given a detailed method for solving a problem, one might expect a better success rate. I thank the reviewer for raising some issues about the Bayesian/frequentists debate among psychologists. To avoid misunderstanding, Cosmides and Tooby (1996) used the term bayesian reasoning (with a small "b") to refer to any cognitive procedure that causes subjects to reliably produce answers that satisfy Bayes' rule, whether that procedure operates on representations of frequencies or single-event probabilities. In this way, they can, without contradiction, ask the question, "Do frequentist representations elicit bayesian reasoning?"

DOI: http://dx.doi.org/10.5038/1936-4660.8.2.7

rule and solve the problems by an equivalent method.) Before further discussing the method, we need to define pertinent basic medical terms.

No test is 100 percent accurate, and there will inevitably be "false positives" and "false negatives." "False positive rate" is the proportion of positive tests among people without the disease.[3] "Sensitivity" or true positive rate is the proportion of individuals with a disease who test positive in a test.[4] The "base rate" of an attribute in a population is the proportion of individuals manifesting that attribute. (A synonym for base rate is "prevalence.") For the above mammography problem, the base rate for a woman at age forty is 1%, the sensitivity of mammography is 80%, and the false positive rate is 9.6%. Boersma and Willard's (2008) *Numeracy* paper on false positives has a comprehensive list of terminology. It also provides guidance on how to organize a test's four possible outcomes in a two-way table.

With Gigerenzer's method of a frequency representation, the mammography problem is translated to:

> *Ten out of every 1000 women have breast cancer. Of these 10 women with breast cancer, 8 will have a positive mammogram. Of the remaining 990 women who don't have breast cancer, some 95 will still have a positive mammogram. Imagine a sample of women who have positive mammograms in screening. How many of these women actually have breast cancer?*

Numerical information in this statement can be presented in a tree of natural frequencies (Fig. 1). The information is the same as before (with rounding), but it is much easier to see what the answer is: only 8 of the 103 women who test positive (8 + 95) actually have breast cancer, which is 7.8% when expressed in probabilities.

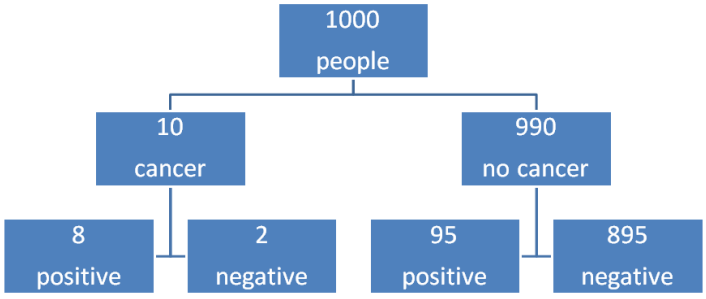

Figure 1. A natural frequency tree for the mammography problem adapted from Eddy (1982).

[3] The term "specificity" used in medical literature is complementary to the false positive rate (i.e., the specificity and the false positive rate add up to 100 percent). It is also called "true negative rate."

[4] The sensitivity and the false negative rate add up to 100 percent.

Below is the method of natural frequencies summarized in four steps (Hoffrage et al. 2000):

1. Select a population and use the base rate to determine how many people in the population have the disease.
2. Take the result and use the test's sensitivity to determine how many people have the disease and a positive test.
3. Take the remaining number of healthy people and use the test's false-positive rate to determine how many people do not have the disease but still test positive.
4. Compare the number obtained in step 2 and the sum of those obtained in steps 2 and 3 to determine how many people with a positive test actually have the disease.

Assessing Students' Learning of Bayesian Reasoning

Before reporting on the NICHE classroom activities, it is relevant to review the rationale for emphasizing the medical diagnosis problem in a QR-infused course. In the first NICHE paper, Wang and Wilder (2015) outlined research results on cognitive illusions showing that a person's beliefs and behaviors do not necessarily correspond to scientific and statistical evidence. There are numerical examples in real life reflecting such a conflict. For example, in 2009, the United States Preventive Services Task Force advised most women in their forties not to have annual mammograms. The response was immediate, and it elicited a public outcry (McGrayne 2011). John Allen Paulos wrote in the *New York Times Magazine* that "both the panel's concerns and the public's reaction to its recommendations may be better understood by delving into the murky area between mathematics and psychology" (Paulos 2009). Many people intuitively believe that early detection through medical tests is invariably beneficial for them, but studies suggested that false positives in mammography often led to unnecessary overtreatment (for an account for the general audience, see Orenstein 2013). Many studies revealed the limits of screening, and the danger of overdiagnosis and overtreatment; these studies have been publicized in the media, (e.g., Parker-Pope 2011 and 2012; Paulos 2012; Kolata 2014).

NICHE is predicated on the proposition that media articles are useful in promoting quantitative reasoning.[5] Moreover, students' quantitative analysis of

[5] Using media articles to teach quantitative reasoning has been advocated by other educators. For example, Madison et al. (2012) is a textbook to guide students to examine public media articles to develop quantitative reasoning skills.

medically related issues will have a profound impact on themselves, their family and their community. *Contextualized use of numbers and data in a matter that involves critical thinking skills* is the working definition of quantitative reasoning for NICHE. In short, the medical diagnosis problem encompasses key skills that we want students to develop. Students need to correctly identify and interpret the base rate, sensitivity, and false positive rate stated in the problem, and they need to process these numerical values into the posterior probability. The calculated result can further prompt them to think about the implication of the probability. Assume, for example, that 11% of patients who test positive for cancer actually have it. What is the cost of treating a perfectly healthy person, and what is the cost of withholding treatment? For the above-mentioned reasons, we contend that an in-depth study of the medical diagnosis problem is highly valuable.

For three semesters, we included the medical diagnosis and mammography problems in the syllabus of an honors elementary statistics course, adapting the approach of natural frequency representations developed by Gigerenzer and his colleagues. Although students in the honors section are highly motivated, there is still a considerable variation in their interest and ability. Some engineering majors have completed calculus courses and are comfortable with calculations, yet some students who just exited remedial algebra experience difficulty in conversion among fractions, percents and decimals. Gigerenzer reported that more than 90% of students obtained the correct rate after training students from the University of Chicago and the University of Munich (Sedlmeier and Gigerenzer 2001). It is interesting to see whether such a method is effective for a less-uniform student population. What follows should not be regarded as a rigorous psychological experiment: our sample was nonrandom and our primary purpose was to help students learn. We followed the standard human subjects research protocols approved by CUNY's Institutional Review Board, and most students granted us their consent to allow us to share data collected in this classroom activity: among 56 students over three semesters, only 4 declined.

The gist of natural frequencies is covered in a *Scientific American Mind* article "Knowing Your Chances" by Gigerenzer et al. (2008). When covering the probability sections of the statistics course, students were assigned to read this article and were informed that they would be tested on it. A week later, in an open-book and open-note quiz, students were given the mammography problem by Eddy (1982) (the standard probability format in Gigerenzer and Hoffrage 1995) shown earlier, with a slightly different base rate for breast cancer for women in their forties obtained from the website of the Centers for Disease Control and Prevention.[6]

[6] Breast Cancer Risk by Age: http://www.cdc.gov/cancer/breast/statistics/age.htm (accessed September 15, 2014).

When students were quizzed on the mammography problem for the first time, essentially none of them had any idea about how to solve it. Only one student in the 2013 Fall semester successfully applied Gigerenzer's method to solve this problem on the first attempt. Most of the students simply tried to search for some (irrelevant) formulas in the book, and plugged in the numbers they found in the problem into formulas to produce a nonsensical answer. To ease students' anxiety, they were told that they would be given credit for trying to understand the problem (the first of the four principles in George Pólya's problem solving strategy[7]).

A detailed analysis of the mammography problem immediately followed the quiz. Although it was shocking for many students to learn that medical tests are frequently inaccurate, after an initial disbelief it was not too difficult for them to accept the possibility of two types of positive results, true and false, quantified by sensitivity (true positive rate) and false positive rate, respectively. Students were shown the CDC site to understand base rates with real data. They were then guided to translate the problem into frequency format, and construct a tree similar to Figure 1.

After this lesson, students were given another quiz in the next class, with a different base rate for a different group of women (e.g., women in their fifties instead of forties) using the data from the CDC. For the second quiz, some students could construct the tree of natural frequencies and obtain the correct answer. Many got the general idea but failed to execute the calculations accurately in each step. A common difficulty is a lack of proficiency in percentage calculations. For example, some students cannot translate 1.5% of 1000 people as 15 people. This situation is quite common among community college students who need math remediation before undertaking a college-level course. Again there was a review of Gigerenzer's method and its application to the quiz problem immediately after the quiz. This activity—a low-stakes quiz with a follow-up discussion and analysis—was repeated several times until most students became comfortable with the problem. The medical diagnosis problem is included in one of the three high-stakes examinations and the final examination.

The mammography problem is not the only example of our Bayesian reasoning training. We also asked students to calculate the probability of pregnancy given a positive test result (e.g., Fig. 2), the probability of admission to a prestigious university given a high SAT score, or the probability that a boyfriend is cheating given that a pair of woman's underwear was found in his drawer (which most students found amusing).

.

[7] In his influential book *How to Solve It* (1945), George Pólya suggested the following steps when solving a mathematical problem: (1) understand the problem; (2) devise a plan; (3) carry out the plan; (4) look back.

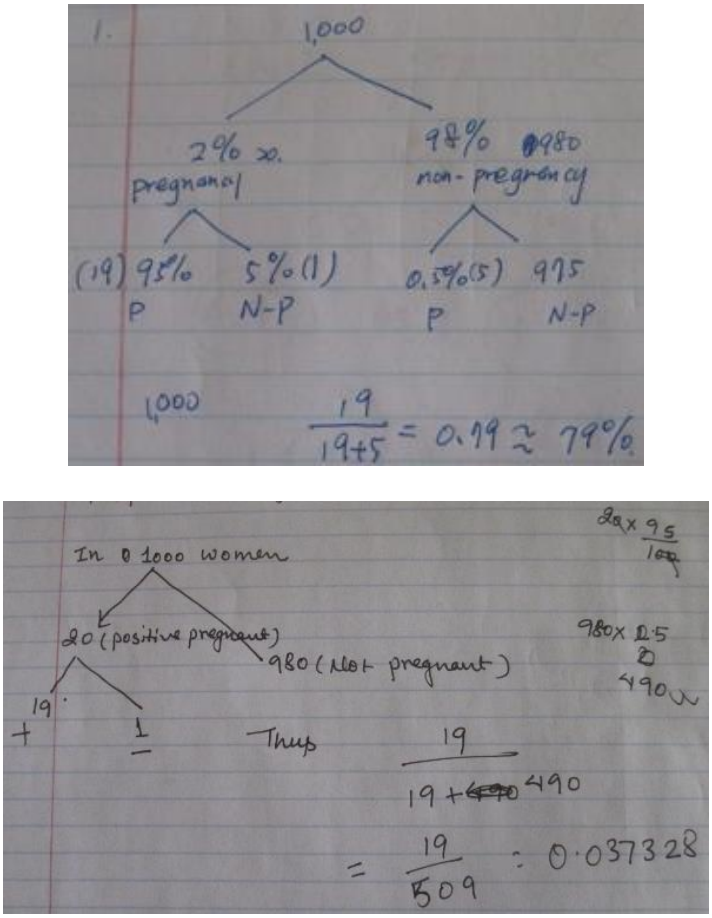

Figure 2. Two students' solution to estimation of the probability of pregnancy given a positive test result with the following information: 2% of women are pregnant; a test has a sensitivity rate of 80% and false positive rate of 0.5%. On the top, a student obtained the correct probability, but on the bottom a students confused 0.5 and 0.5% and obtained the incorrect answer

Figure 3 shows the result of student performance collected in the fall semester of 2013. The area under the blue line represents the number of students who successfully obtained the correct probability. "Partially correct," under the red line, represents students who correctly identified the base rate, sensitivity and false positive rate, but failed to execute the calculations (mostly due to faulty calculations exemplified in Figure 2). "Incorrect" describes students who were unfamiliar with the nature of medical diagnosis, or confused over the three different rates given in the problem. Non-constancy of the number of total students is due to absence. From Figure 3, the gradual improvement in the

performance of students as a group is evident. On the date of the second examination (November 5, 2013), 14 out of 19 students (74%) obtained the correct answer. The correct rate dropped slightly, to 68%, on the date of the final examination (December 10, 2013), perhaps because some students forgot the method or because there were slight variations in the composition of students present at the testing dates.

The charts for the other two semesters show a similar pattern of gradual improvement (Figure 4). The correct rates on the date of the final examination were 71% (14 students in total) and 50% (16 students in total) for 2013 spring and 2014 spring, respectively. Toward the end of the semester, typically only 1 or 2 (out of approximately 20) students remained entirely unfamiliar with the medical diagnosis problem (primarily due to their excessive absences). A striking feature for the three semesters is that deficiency in elementary arithmetic skills continues to prevent some students from obtaining the correct answer, despite their comprehension of the meaning of prevalence, sensitivity and false positive rates. This deficiency likely explains the discrepancy of correct rates between community college students and those from the University of Chicago and University of Munich reported in Sedlmeier and Gigerenzer (2001).

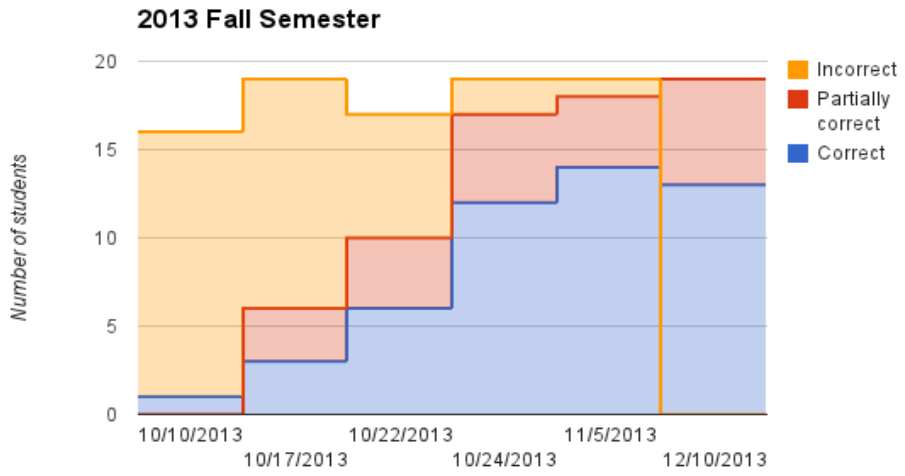

Figure 3: Students' performance on the mammography problem and other Bayesian inference problems in 2013 Fall.

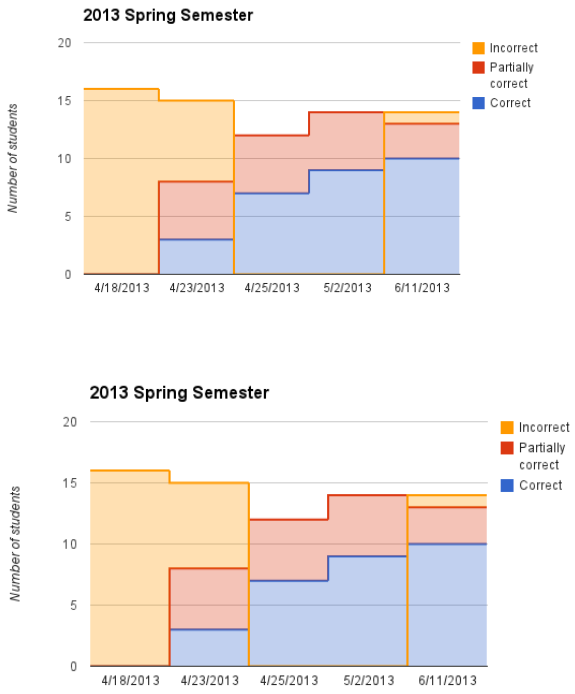

Figure 4: Students' performance on the mammography problem and other Bayesian inference problems in 2013 Spring and 2014 Spring.

In addition to these in-class quizzes and examinations, a problem on the hemoccult test for colorectal cancer (from Hoffrage and Gigerenzer 1998) was given to students as a written assignment. Students were instructed to explain the uncertain nature of a cancer screening program and the base rate, sensitivity and false positive rate in plain language so that their parents and friends could understand. They needed to show their calculations leading to the posterior probability, and most importantly articulate how a personal decision would be made based on the probability. Based on their writing, most students were able to appreciate the need to weigh the costs and benefits associated with a test. Some examples of concluding paragraphs from two students are the following:

> *A little under 5% of people who test positive actually have the cancer, this can lead to expensive treatments when in fact you don't even have the cancer. On the flip side if the cancer is caught at an early stage it is definitely worth it.*
>
> *4.8% is still much higher than the 0.3% base rate but instead of assuming the worst you just have to look at all the numbers and think carefully and thoroughly. It is honestly a useful message for almost all aspects of life.*

Bloom's taxonomy divides educational objectives into three domains: psychomotor, cognitive, and affective (Bloom 1984). The medical diagnosis problem that we used to teach Bayesian reasoning largely addresses the first two

domains.[8] For the third domain, we administered a survey at the end of the semester to gauge students' attitudes toward medical testing after exposure to the mammography and medical diagnosis problems, We selected three questions from a Dartmouth Medical School study "Enthusiasm for Cancer Screening in the United States" (Schwartz et al. 2004). Using Google Drive's survey tool, students anonymously responded to the following three questions.

1. If there was a kind of cancer for which nothing can be done, would you want to be tested to see if you have it?
2. Routine screen means testing healthy persons to find cancer before they have any symptoms. Do you think routine cancer screening tests for healthy persons are almost always a good idea?
3. Would you prefer a total-body CT scan or receiving $1000 in cash?

Students' responses are summarized in Figure 5. "Still want to know" means "Yes" answer to the first question. "Always a good idea" means "Yes" answer to the second question. "Choose CT" means that respondents prefer a total-body CT scan in the third question. Standard errors for Schwartz et al. were provided by Dr. Steven Woloshin, based on their national telephone interview of adults conducted from December 2001 through July 2002 (S. Woloshin, personal communication, November 10, 2013).

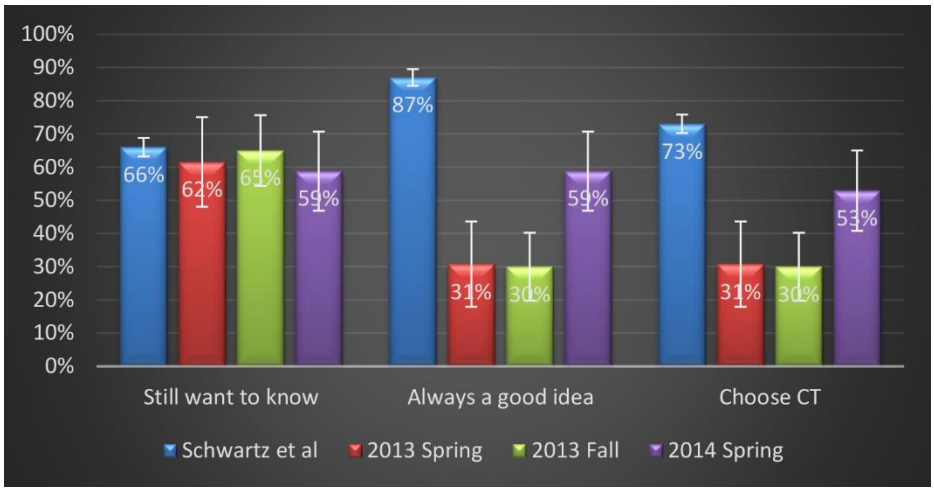

Figure 5: Student responses to 3 survey questions and their comparison with a national survey. The sample size is n=500 for Schwartz et al. (2004); n=13, 20, and 17 for 2013 Spring, 2013 Fall, and 2014 Spring, respectively.

[8] The interpretation of Bloom's taxonomy has been evolving since 1956. Here we adapt the view of Linda Suskie (2009), who categorized learning into three domains: (1) knowledge and conceptual understanding; (2) thinking and other skills; (3) attitudes, values, dispositions, and habits of mind. Chapter 8 of Suskie's book is the reading for NICHE's second unit—QR Learning Goals.

For three semesters, most students still want to know whether they have a cancer even if nothing can be done; their need for certainty is consistent (within the error bar) with the national survey results. However, most students appear to realize that routine cancer screenings for healthy persons are not necessarily always a good idea, and students' responses were significantly different from the national survey results. A total-body CT scan gives a very detailed picture of our body; it has the potential to find many diseases but at the same time it can create many false alarms. Again, students' choices between receiving a full-body CT scan versus cash deviate from the national survey results as the community college students were less likely to opt for the former. Although there are many possible explanations for the differences (e.g., college students tend to be younger; community college students may have a greater need for money; and public attitudes might have changed since 2002), it is plausible that the medical diagnosis problem made students think more carefully and become less enthusiastic about indiscriminate cancer screening.

Conclusions

In an elementary statistics course, the method of estimating conditional probability (specifically solving the mammography and other Bayesian reasoning problems) using natural frequency representations developed by Gerd Gigerenzer and his collaborators was introduced to community college students. For this group of students, deficiencies in elementary arithmetic skills are common. Nevertheless, with repeated practice, many students have mastered the technique and avoided the bias in judgment that is prevalent among medical professionals reported in the literature. This result is very encouraging. Bayesian reasoning is a crucial skill to navigate in the modern world, which is full of information expressed in probabilistic terms, and our case study suggests that it is feasible to teach such a skill that is built on psychological principles to underprivileged students—the group of students for whom NICHE was designed.

Acknowledgment

Support for the NICHE project has been provided by the National Science Foundation's (NSF) Transforming Undergraduate Education in Science, Technology, Engineering and Mathematics (STEM) (TUES) award #1121844. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily represent the views of the National Science Foundation.

References

Bar-Hillel, M. 1980. The base-rate fallacy in probability judgments. *Acta Psychologica* 44: 211–233. http://dx.doi.org/10.1016/0001-6918(80)90046-3

Bloom, B. S. 1984. *Taxonomy of Educational Objectives Book 1: Cognitive Domain*, 2nd edition. New York: Longman.

Boersma, S., and T. Willard. 2008. False positives and referral bias: Content for a quantitative literacy course. *Numeracy* 1(2): Article 5. http://dx.doi.org/10.5038/1936-4660.1.2.5

Casscells, W., A. Schoenberger, and T. B. Graboys. 1978. Interpretation by physicians of clinical laboratory results, *New England Journal of Medicine* 299(18): 999–1001. http://dx.doi.org/10.1056/NEJM197811022991808

Cosmides, L., and J. Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58: 1–73. http://dx.doi.org/10.1016/0010-0277(95)00664-8

Eddy, D. M. 1982. Probabilistic reasoning in clinical medicine: Problems and opportunities. In *Judgment under Uncertainty: Heuristic and Biases*, ed. D. Kahneman, P. Slovic, and A. Tversky, 249–267. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477.019

Gigerenzer, G. 1996. On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review* 103(3): 592–596. http://dx.doi.org/10.1037/0033-295X.103.3.592

———. 2002. *Calculated Risks*. New York: Simon & Schuster.

———, and U. Hoffrage. 1995. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102(4): 684–704. http://dx.doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G., W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. 2009. Knowing your chances. *Scientific American Mind* 20: 44–51. http://dx.doi.org/10.1038/scientificamericanmind0409-44

Gill, C. J., L. Sabin, and C. H. Schmidt. 2005. Why clinicians are natural Bayesians. *British Medical Journal* 330: 1080–1083. http://dx.doi.org/10.1136/bmj.330.7504.1390-d

Hoffrage, U., and G. Gigerenzer. 1998. Using natural frequencies to improve diagnostic inferences. *Academic Medicine* 73: 538–540. http://dx.doi.org/10.1097/00001888-199805000-00024

Hoffrage, U., S. Lindsey, R. Hertwig, and G. Gigerenzer. 2000. Communicating statistical information. *Science* 290: 2261–2262. http://dx.doi.org/10.1126/science.290.5500.2261

Kahneman, D., and A. Tversky. 1996. On the reality of cognitive illusions. *Psychological Review* 103(3): 582–591. http://dx.doi.org/10.1037/0033-295X.103.3.582

Kolata, G. 2014. Study points to overdiagnosis of thyroid cancer. *New York Times* November 6: p. A4.

Madison, B. L., S. Boersma, C. L. Diefenderfer, and S. W. Dingman. 2012. *Case Studies for Quantitative Reasoning: A Casebook of Media Articles*. 3rd edition. Boston, MA: Pearson Learning Solutions.

McGrayne, S. B. 2011. *The Theory That Would Not Die*. New Haven, CT: Yale University Press.

Orenstein, P. 2013. Our feel-good war on breast cancer. *New York Times Magazine* April 28: 36–43, 68–71.

Parker-Pope, T. 2011. Mammogram's role as savior is tested. *New York Times* October 24: p. D1.

———. 2012. Prostate test advice appears unheeded. *New York Times* April 25: A17.

Paulos, J. A. 2009. Mammogram math. *New York Times Magazine* December 13: 19–20.

———. 2012. Weighing the positives: Breaking down the latest mammogram math. *Scientific American* 306(1): 20. See also "Letters to the Editor" in reply to this piece. 2012. *Scientific American* 306(5): 8.

Pólya, G. 1945. *How to Solve It*. Princeton, NJ: Princeton University Press.

Ross, S. 2014. *A First Course in Probability*. 9th edition. Upper Saddle River, NJ: Prentice Hall.

Schwartz, L. M., S. Woloshin, F. J. Fowler, and H. G. Welch. 2004. Enthusiasm for cancer screening in the United States. *Journal of the American Medication Association* 291(1): 71–78. http://dx.doi.org/10.1001/jama.291.1.71

Sedlmeier, P., and G. Gigerenzer. 2001. Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General* 130(3): 380–400. http://dx.doi.org/10.1037/0096-3445.130.3.380

Strogatz, S. 2012. *The Joy of X*. New York: Houghton Mifflin Harcourt.

Suskie, L. 2009. *Assessing Student Learning: A Common Sense Guide*. San Francisco: Jossey-Bass.

Wang, F., and E. I. Wilder. 2015. Numeracy Infusion Course for Higher Education (NICHE), 1: Teaching Faculty How to Improve Students' Quantitative Reasoning Skills through Cognitive Illusions. *Numeracy* 8(2): Article 6. http://dx.doi.org/10.5038/1936-4660.8.2.6