QUESTION FORMAT AND REPRESENTATIONS: DO HEURISTICS AND BIASES APPLY TO STATISTICS STUDENTS?

JENNIFER J. KAPLAN
Michigan State University
kaplan@stt.msu.edu

JUAN DU
Kansas State University
dujuan@ksu.edu

ABSTRACT

Researchers in the field of psychology studying subjects' reasoning abilities and decision-making processes have identified certain common errors that are made, particularly on probability questions standard in introductory statistics courses. In addition, they have identified modifications to problems and training that promote normative reasoning in laboratory subjects. This study attempts to replicate, in the context of a statistics classroom, the results of one particular type of probability question, a two-stage conditional probability problem. The psychology literature suggests two possible implications for teaching probability. Although no effect for format modification was found, the representations training effects were replicated. The implications of these results for teaching and directions for future research are discussed.

Keywords: *Statistics education research; Probability; Representations; Question format*

1. INTRODUCTION

Researchers in the field of psychology have been conducting studies designed to test the decision-making skills of humans since Peter Wason's work in the early 1960s (Evans & Newstead, 1995). In the early 1970s Tversky and Kahneman began to publish their now famous "heuristics and biases" literature. Although researchers who have followed in the wake of Wason and Kahneman and Tversky have disputed the original findings (Hertwig & Gigerenzer, 1999), the fact remains that subjects are prone to making errors in judgment on the designed tasks (Stanovich, 1999).

The psychology results on human reasoning, studied through the lens of statistics education, will inform statistics teaching because the irrationalities discovered by psychologists may represent misconceptions held by statistics students. The research presented in this paper is centered on two-stage conditional probability problems that are found in standard introductory statistics textbooks. This type of problem was chosen because it has a rich history of having been studied by psychologists and is easily embedded into an introductory statistics course because of the prevalence of the topic across textbooks designed for those courses (see, for example, De Veaux, Velleman, &

Statistics Education Research Journal, 8(2), 56-73, http://www.stat.auckland.ac.nz/serj

Bock, 2006, and Moore, 2003). This paper explores the research on the effects of changing the format of such items and the effects of isomorphic questions with different presentation formats on student performance. Furthermore, the findings of studies in which subjects are trained to use representations to solve probability and statistics problems are used to suggest a blueprint for teaching similar concepts to students. The purpose of this research, therefore, is two-fold. First, it introduces a body of literature into statistics education research. Second, it describes a classroom experiment based in the body of research and compares the results from the classroom to the results from the psychology laboratories.

2. BACKGROUND

2.1. THE ORIGINAL PROBLEM

Casscells, Schoenberger, and Grayboys (as cited in Cosmides & Tooby, 1996) created the following problem in 1978 for a study at the Harvard Medical School:

> *Medical Diagnosis Problem (Original form)*
> 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 a positive result actually has the disease, assuming you know nothing about a person's symptoms or signs? ___%

Of the sixty medical students and faculty used as subjects in the original study, only 18% gave the correct answer of 2% (Cosmides & Tooby, 1996). Forty-five percent of their subjects responded that 95% of those who test positive actually have the disease, appearing to have used the complement of the false positive rate as the true positive rate.

This question is a brief version of a standard two-stage conditional probability problem of the type found in many textbooks used in introductory statistics classes. The correct solution can be worked out using the tree diagram in Figure 1.

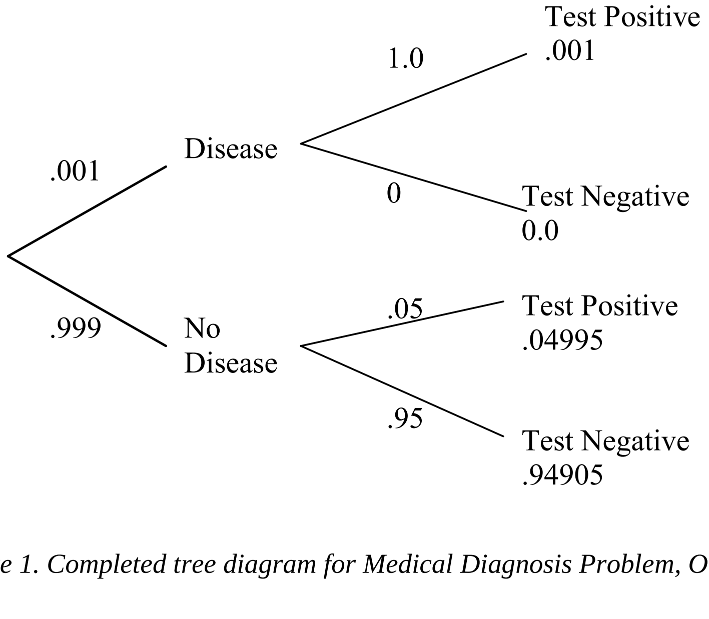

Figure 1. Completed tree diagram for Medical Diagnosis Problem, Original form

The question asks for the probability that someone who tests positive for the disease actually has the disease, P(Disease | Test Positive). P(Disease | Test Positive) = P(Disease AND Test Positive)/ P(Test Positive) = .001/(.001+.04995) ≈ .0196 or approximately 2%.

Given the fact that medical students and faculty fare so poorly with a question in a domain with which doctors should be familiar, is it reasonable to expect undergraduate students to answer this type of question in a beginning statistics course?

In the three decades since Casscells et al. published their original study, much work has been done using modifications of their original problem and other similar two-stage conditional probability problems. Nearly every subsequent administration of such a problem included more explanation about the false positive rate and specified the false negative rate as well. Some of the presentation factors that have been studied using these problems are (a) presentation format - whether the initial information is given as a percentage or as counts; (b) salience of the random sample - whether the question asks about a person, a person selected at random, or the number of expected outcomes in a large group; and (c) the addition of a sub-question in which the subject explicitly calculates the denominator of the conditional probability. A review of these studies is given in the next section and a summary of the findings appears in Table 1.

In addition to the study of factors associated with the problem presentation and their effect on the ability of subjects to give the correct response, several training studies have been associated with this type of problem. In these studies, subjects are trained to use representations to solve similar problems. The results of the presentation and training studies are discussed in the next sections. The presentation formats that are considered in this paper are percentages, partitive counts and non-partitive counts. The presentations of the questions that include sub-questions will be called "two-step" questions. This should not be confused with the general two-stage conditional probability question.

2.2. FACTORS ASSOCIATED WITH THE PROBLEM PRESENTATION

Cosmides and Tooby (1996) published a major work in this branch of the literature using variations of the Medical Test Problem stated above. They first replicated the findings of Casscells et al. using Stanford undergraduates as subjects. Subsequently, they found that they could significantly raise the rate at which respondents gave the correct answer by changing the presentation format of the problem to one that gave information and requested a response in frequencies rather than via a percentage:

> *Medical Diagnosis Problem (Frequency form with explanation)*
> One 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. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease. 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 those people.
>
> Given the information above:
>
> On average, how many people who test positive for the disease will *actually* have the disease? ___ out of ____?

This version of the problem differs from the original in more than use of a frequency presentation and a request for the answer in frequency format. This version also explains the concept of a false positive and specifies the true positive rate. Furthermore, it highlights the salience of the random sample and makes the sample more concrete by enumerating its size (Cosmides & Tooby, 1996). Fiedler (1988) demonstrated that the enumeration of cases, by itself, would improve the correct response rate remarkably. In order to unpack the possible reasons for the rise in correct response rate, Cosmides and Tooby followed up with a study that used a similar text to the frequency form with explanation except the data were presented in percentage format:

> *Medical Diagnosis Problem (Percentage format with explanation)*
> The prevalence of disease X is 1/1000. 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. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, 5% of all people who are perfectly healthy test positive for the disease.
>
> What is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the symptoms or signs?_____%

Subjects did significantly better on this version of the question than those who responded to the original form, with 36% giving a correct response. This indicates that the explanation of false positive did aid performance. The percentage of correct responses to this version, however, was significantly lower than on the frequency format with explanation, demonstrating that the frequency format encourages normative performance as well.

One might note that it is possible that Cosmides and Tooby (1996) have merely replicated Fiedler's (1988) findings that a subject is more likely to give the correct response when the existence of a sample is salient. In the percentage format version, the question asks about the probability that one person who tests positive actually has the disease, whereas in the frequency format version the question posed asks about the number of people in a large group who one would expect to have the disease. In order to test the effect of the specificity of the sample, the following version, in which a probabilistic sample was specified, was also tested:

> *Medical Diagnosis Problem (Percentage format with explanation and random sampling assumption)*
> The prevalence of disease X among Americans is 1/1000. 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. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, 5% of all people who are perfectly healthy test positive for the disease.
>
> Imagine that we have given this test to a random sample of Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.
>
> What is the chance that a person found to have a positive result actually has the disease? _____%

Twenty-eight percent of the subjects gave the correct response. Sixteen percent gave the base rate response of 1/1000. These results are not significantly different from the results found using the "percentage format with explanation" version of the problem. The difference between the two versions is the salience of random sampling. Because there were no significant differences in performance on the two items, it appears that the salience of random sampling was not a factor in subjects' ability to respond to this type of item.

Girotto and Gonzalez (2001) realized that in each of the preceding examples, the frequency format requests a two-step response, providing a space for both the numerator and the denominator of the ratio, whereas the percentage format requests a one-step response whether or not the sample is made salient. They created 8 items, used phrasing similar to the Cosmides and Tooby's (1996) in order to investigate differences due to the salience of the two parts of the ratio. To create the eight problems, they used two presentation formats, frequency of people and frequency of chance:

> *Frequency of people version:*
> 4 out of 100 people tested were infected.
> 3 out of 4 infected people have a positive reaction to the test.
> 12 of the 96 uninfected people also had a positive reaction to the test.
>
> *Frequency of chance version:*
> A person who was tested has 4 chances out of 100 of having the infection.
> 3 out of 4 chances of having the infection were associated with a positive reaction to the test.
> 12 of the remaining 96 chances of not having the infection were also associated with a positive reaction to the test.

These were then crossed with four question formats, one- and two-step frequency of people and one- and two-step frequency of chance:

> *Two-step Frequency of chance question:*
> Imagine that Pierre is tested now. Out of a total of 100 chances, Pierre has _____ chances of having a positive reaction, _____ of which will be associated with having the infection.
>
> *Two-step Frequency of people question:*
> Imagine that a group of people is tested. In a group of 100 people, one can expect _____ individuals to have a positive reaction, _____ of whom will have the infection.
>
> *One-step Frequency of chance question:*
> If Pierre has a positive reaction, there will be ________ chance(s) out of _______that the infection is associated with his positive reaction.
>
> *One-step Frequency of people question:*
> Among 100 people who have a positive reaction to the test, the proportion that has the infection will be equal to _______ out of _______.

Each subject solved two problems. Both problems had the same question format, but one had frequency of people presentation format and the other had frequency of chance presentation format. Subjects correctly solved two-step versions at a higher rate than the one-step versions, regardless of format of presentation or response. This indicates that drawing attention to the numerator and denominator of the correct response increased the ability of subjects to give that response.

All of Girotto and Gonzalez's (2001) questions are in a format that has come to be called "partitive frequency" format. That is, the number of people in each of the four categories that result at the end of the two-stage problems can be calculated by partitioning the number of people in each category at the end of the first stage of the problem without the need for multiplication. Macchi (1995, 2000) investigated whether the benefit of the frequency format occurred for both partitive and non-partitive formulations. An example of a possible non-partitive frequency format is given below:

Non-Partitive Frequency Format:
2 out of 50 people tested were infected.
3 out of 4 infected people have a positive reaction to the test.
1 of the 8 uninfected people also had a positive reaction to the test.

Note that the two people who are infected cannot be easily partitioned into the two reaction categories, positive and negative. Macchi (1995) found that there was a significant difference in the percentage of subjects who gave the correct answer to her items. Those subjects solving partitive frequency format problems performed much better than those solving non-partitive frequency format problems.

Table 1 provides a summary of the results of the presentation format experiments discussed above. In general, the literature on frequency formats and two-stage conditional probability problems finds the following:

1. Partitive frequency formatted items are easier to solve than either non-partitive frequency format or percentage formatted items.
2. Two-step formatted items are easier to solve than one-step formatted items.
3. There is no difference in difficulty based on the salience of the random sample.

Table 1. Summary of research results in two-stage conditional probability problems

Researchers	Presentation format	Steps	Random sample	% correct	*n*	Type of subjects
Casscells, Schoenberger & Grayboys	Percentages	One	No	18%	60	Physicians, 4th year medical students at Harvard Medical School
Cosmides & Tooby	Percentages	One	No	12%	25	Paid volunteers recruited by advertisement at Stanford University
	Percentages	One	Yes	28%	25	
	Percentages	One	No	36%	25	
	Partitive Counts	Two	Yes	76%	50	
Girotto & Gonzalez	Partitive Counts	One	Yes	18%	40	Undergraduate Psychology students at the University of Provence, France
	Partitive Counts	Two	Yes	58%	40	
Macchi	Non-Partitive Counts	Two	Yes	13%	30	Undergraduate students
	Partitive Counts	Two	Yes	78%	30	

2.3. REPRESENTATIONS TRAINING

Training studies comprise another avenue of research about the two-stage conditional probability questions. These training studies were based on the cognitive perspective claiming that when subjects can create a representation of a situation, they are more likely to compute the correct response to the item. The "mental models" theory of human reasoning specifies that the process of reasoning includes constructing a model (or set of models) based on the premises and using general knowledge to make explicit something only implicit in the premises (Johnson-Laird, 1994). Thus, the ability to create a representation of the problem appears to be a crucial element leading to a correct solution.

In order to create an item that could be represented more easily, Cosmides and Tooby (1996) revised their medical diagnosis item, discussed above, so that

- the sample size was 100 (disease rate 1/100, false positive rate 5/100),
- subjects were given a 10 × 10 grid representing a sample of 100 people,
- subjects were required to circle the boxes representing people who had the disease and to fill in the number of people who would test positive *before* they gave the rate of positive tests that actually indicate the disease.

Under these conditions, 92% of subjects gave the correct response to the item, significantly higher than on any other administration of the task, In fact, the authors also found that many of the subjects who gave the correct response to the percentage version of the problem with explanation of false positive left evidence in their booklet of having enumerated the cases of positive tests. The results from this item indicate that the ability to create representations enhances the ability to provide the correct response to questions of this type.

The notion that training students to create suitable representations in order to complete a standard probability task was also the subject of a research project conducted by Sedlmeier (1999). Sedlmeier provided training for students using either a Venn diagram or a grid approach for probability questions. Another group of subjects was trained in the use of Bayes' formula to solve probability problems. This approach is an example of "rule-based training." In addition to training in the domain of probability, Sedlmeier also trained students to use a flexible urn model to answer questions about sampling distributions. In the flexible urn model, subjects were trained to imagine the sampling step of a hypothesis test process as if they were taking a random sample of balls from an urn. In the case of a test for proportions, the balls were thought to be of two colors. In the case of a test for means, the balls were considered to have values. Sedlmeier found, for all three representations (Venn diagrams, grids, and the flexible urn), that the training was successful on transfer problems and in follow-up tests five weeks after the training. Further, representations training was found to be significantly more successful than rule-based training in helping subjects to obtain the correct response. These results indicate that benefits from representations training may transfer to the classroom.

3. THE EXPERIMENT

3.1. RESEARCH QUESTIONS

The research on two-stage conditional probability questions presented above suggests that psychology subjects are more likely to give the correct response to such questions when (a) the questions are presented in partitive frequency format, (b) the questions are given in a two-step format and (c) the subjects are trained to use a representation to aid in the solution of the problem.

The research study presented in this paper is an example of classroom-based research enacted by the instructor-researcher. Because the prior research suggests that students would have more success in solving such problems if they are presented in partitive frequency format, all example problems in class were completed using partitive frequency format. Students were, therefore, taught to convert from percentage format to partitive frequency format. Based on the positive research results about the use of representations, all class examples were completed using tree diagrams. Tree diagrams were chosen because of their prevalence in textbooks. In fact, the two worked examples in the course textbook contained tree diagrams in the solution. In these ways, the psychology research results informed the teaching of probability in the class under research. This fact was taken into consideration in this research.

As will be discussed below, the data for this study were collected through course assessments. In addition, all of the two-stage probability problems that appeared in the course textbook were presented in two-step percentage format. Given that all prior example problems had been presented in two-step format, it seemed unreasonable to assign a random group of students to solve a more difficult one-step problem on a course exam. For practical reasons, therefore, this research does not investigate the difference between one- and two-step presentations of conditional probability problems. Given the issues discussed in this section, the research presented was designed to focus on the following questions:

1. Do differences in format of a probability question produce different outcomes on two-stage conditional probability from students in an introduction to statistics class?
2. Is a tree diagram a useful representation for students in an introduction to statistics course when solving two-stage conditional probability problems?

These questions form the basis of the research experiment.

3.2. RESEARCH SUBJECTS

The research subjects were students at a large public research university in the Midwest. All students were in the same large lecture Introduction to Statistics class taught by the first author. The students in the class were asked to give consent to the author to use information collected via written assessments as research data. Of the 115 students who completed the course, 24 had not consented to be research subjects. When the number of points earned by the students who gave consent was compared to the number of points earned by the students who did not give consent, those not consenting were found to have statistically significantly lower scores (p-value < 0.01). The generalizability of the results of the study may be affected by two factors: (a) all students in the study had the same teaching treatment, and (b) the difference in performance between the consent group and the non-consent group.

3.3. PROCEDURE

The topic of probability was presented in class during the sixth week of the semester. The focus of the instruction was on using representations, rather than formulas to solve probability problems. The following learning goal for probability was presented to the students:

> Students will be able to find the probability of one- and two-stage events including conditional probability through the use of a table, Venn diagram, tree diagram and/or area model and various "rules" for probability.

The word rule is in quotes in the learning goal because although the instruction indicated that the textbook presented certain rules and formulas, such as the "something must happen" rule and the general addition and multiplication rules, the rule that would be stressed in the classroom was "draw a picture, draw a picture, draw a picture."

The textbook contained two worked examples of two-stage probability problems with tree diagrams in percentage format. In addition, two examples were completed during lecture using partitive frequency format and tree diagrams. The textbook had seven exercises about two-stage probability problems. One was collected as a homework problem and the assignment specified that the students include a tree diagram in their solution. This homework problem was collected during the seventh week of the semester, the week following the classroom instruction. It was corrected by the first author and returned at the beginning of the eighth week of the semester.

The second midterm was given at the end of the ninth week of the semester, and the final exam in the week after the 15th week of the semester. The testing dates were roughly six weeks apart. These two examinations each contained one conditional probability problem. Three versions were created for each problem: one in percentage format, one in partitive frequency format, and the third in non-partitive frequency format. All other aspects of the problems, contexts and values for example, were constant across the formats. The assessment items are reprinted in Appendix A. The examination papers containing one of the three versions of the two-stage conditional probability problem were randomly distributed to the students.

3.4. DATA STRUCTURE

This research used a repeated measures design with a dichotomous outcome variable. The structure of the data is summarized in Table 2. The experimentally manipulated factor in the study was the presentation format of the problem on the examinations: percentages, partitive counts, and non-partitive counts. The fact that the students had been instructed in class in the use of tree diagrams and to convert presentation formats to partitive counts might have influenced the effects of this factor. Therefore, information on these factors was also collected and incorporated into the model. Specifically noted were the subjects' use of a tree diagram and the format in which they chose to work (percentages or not percentages). The use of a tree diagram was coded as completed, unfinished, or none. A completed tree diagram was one in which the counts or probability of each of the four branches of the tree had been written by the student. Figure 1 above is an example of a "completed" tree diagram. Figure 2 below is an example of an "unfinished" tree diagram because the probabilities at the end of each branch have not been calculated. Unlike the homework problem, the problems on the midterm and final exam did not specify or suggest the use of a diagram.

The final covariate included in the model was the number of points earned by the student over the course of the semester. There were a total of 450 points that students could earn. The distribution of points earned was unimodal with a left skew. Students earned between 183 points and 447 points with a mean of 362 points, and a standard deviation of 52.5 points.

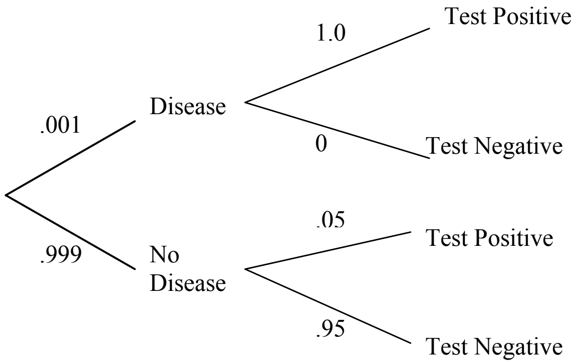

Figure 2. Unfinished tree diagram for Medical Diagnosis Problem, Percentage format

The outcome variable was whether the student solved the problem correctly or incorrectly. Moreover, the outcome variable was repeatedly measured on three occasions: Time 1, the collected homework; Time 2, the second midterm; and Time 3, the final exam.

Table 2. Structure of the data

Variables	Class	Levels	Values
Response	Correct	2	1, 0
Explanatory	Subject	88	1, 2, …, 87, 88
	Time	3	T1_HW, T2_Midterm, T3_Final
	Tree use	3	Completed, unfinished, none
	Work format	2	not percentages, percentages
	Presentation	3	non-partitive counts, partitive counts, percentages
	Total Score	Continuous	Min = 183; Max = 447; Mean = 362; SD = 52.5

3.5. ANALYSIS

Ninety students turned in the homework problem, 92 took the second midterm, and 89 took the final exam. Table 3 gives the percentage correct by presentation format and tree diagram as well as the number of students in each group for each of the three assessments. Note that the percentage correct for two-step non-partitive count format is higher than predicted by the psychology research and the percentage correct for two-step partitive count format is lower than expected.

Recall that this research seeks to ascertain whether the question format affects the percentage of students who answer the probability question correctly. In addition, the research is interested in whether the tree diagram is a useful representation for statistics students when solving this type of problem.

We expect the observations recorded on a student at three time points to be correlated, and we can assume observations between different students to be independent. In view of this, we used logistic regression with correlated outcomes to model the data, the Generalized Estimating Equations (GEEs) approach introduced by Liang and Zeger (1986). Our initial model is a logistic model predicting the log odds of success for the i^{th}

Table 3. Percentage correct and number of students in each Format treatment group

Time	Presentation format	Percent correct	Total number	Tree diagram	Percent correct	*n*
Homework	Percentages	75.6%	90	Completed	92%	50
				Unfinished	65.6%	32
				None	12.5%	8
Midterm	Percentages	42.2%	45	Completed	52.5%	59
	Partitive counts	40.0%	25	Unfinished	31.6%	19
	Non-partitive	36.4%	22	None	0%	14
	Overall	40.2%	92			
Final Exam	Percentages	46.7%	30	Completed	72.4%	58
	Partitive counts	53.6%	28	Unfinished	21.4%	19
	Non-partitive	58.1%	31	None	11.8%	14
	Overall	52.8%	89			

student at time *t* from *treeuse*$_{it}$, *workformat*$_{it}$, *presentation*$_{it}$, *totalscore*$_{it}$, *time*$_{it}$. Following the logistic model expression, each regression coefficient can be interpreted as the increase in the log-odds of correctly solving the problem associated with a one-unit increase in the j^{th} explanatory variable. (And exponentiating that coefficient tells us the multiplicative increase in the odds of correctly solving the problem.)

Further, the within-subject correlation is accounted for by using the working correlation matrix given in Table 4. This is a weighted matrix estimated using an iterative fitting process that is a generalization of the least squares method. Details on the calculation of the correlation matrix are given in Appendix B.

Table 4. Working correlation matrix for GEE analysis

Time	Homework	Midterm	Final
Homework	1	-0.051	-0.0396
Midterm	-0.051	1	0.3114
Final	-0.0396	0.3114	1

The estimates of the correlation parameters suggest that those students who correctly solved the problem on the midterm had a higher probability of having a like outcome on the final exam (correlation is 0.3114). The homework performance, however, has vague negative correlation with the exams; this might be due to the fact that students may use their textbooks, work together or go to a Teaching Assistant for help when completing homework assignments, but do not have such aids on a test.

The algorithm for fitting the specified model using GEEs is given in Appendix B. The analysis of the GEE parameter estimates showed that Tree use, Time, and Total score have a significant relationship with correctly solving the two-stage probability problem, but that Presentation Format and Work Format were not significant. The summary of the Wald statistics is given in Table 5. Although Presentation is not significant, all other factors have significant influence on the response variable.

In order to judge the effect sizes of the significant factors, contrast estimates were calculated using Proc GENMOD in SAS 9.1. See Kutner, Nachtsheim, Neter, and Li (2005) for the more detailed setup of contrast and design matrix. In Table 6, the estimate gives the odds ratio comparing the two situations involved in the contrast and then tests the hypothesis that the two cases will result in the same odds of getting the problem correct. For instance, if we are interested in the difference in the odds between those with a completed tree diagram and those who did not use a tree diagram, we estimate the corresponding coefficient difference (β_{11}- β_{12}). The p-value of 0.0005 shows that the students who completely finished the tree diagram were significantly more likely to get the problem right as opposed to those who didn't use the tree diagram at all.

Table 5. Wald Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
Work format	1	0.14	0.7083
Tree use	2	22.48	< 0.0001
Presentation	2	2.51	0.2855
Time	2	27.18	< 0.0001
Total score	1	12.89	0.0003

Table 6. Contrast estimate results

Label	Estimate	Standard error	Confidence limits		χ^2	Pr > χ^2
Tree_use (completed/none)	3.17	0.9047	1.40	4.95	12.31	0.0005
Exp(Tree_use)	23.90	21.6229	4.06	140.76		
Tree_use (unfinished/none)	1.63	0.8881	-0.11	3.37	3.36	0.0668
Exp(Tree_use)	5.09	4.5242	0.89	29.04		
Tree_use (completed/ not)	1.55	0.388	0.79	2.31	15.87	< 0.0001
Exp(Tree_use)	4.69	1.8205	2.19	10.04		
Time (Final vs. HW)	-1.90	0.4505	-2.79	-1.02	17.83	< 0.0001
Exp(Time)	0.15	0.0672	0.06	0.36		
Time (Midterm vs. HW)	-2.44	0.4679	-3.36	-1.52	27.18	< 0.0001
Exp(Time)	0.09	0.0408	0.03	0.22		
Time (Final vs. Midterm)	0.54	0.2798	-0.01	1.09	3.69	0.0548
Exp(Time)	1.71	0.479	0.99	2.96		
Total score	0.012	0.0045	0.01	0.02	12.89	0.0003
Exp(Total score)	1.0161	0.0045	1.01	1.03		

The estimates in Table 6 show, after adjusting for correlated outcome data using unstructured correlation matrix (Table 4) and controlling for Work format, Presentation Format, Time, and Total score, those who finished a tree diagram were 23.9 times more likely to correctly solve a two-stage probability problem compared to those who didn't use a tree at all (p-value = 0.0005). The odds that students who finished the tree diagram correctly solved the conditional probability problem are more than 4.7 times the odds for those who did not finish the tree diagram (p-value < 0.0001). The odds of solving the conditional probability problem correctly for students who didn't finish the tree diagram are 5 times the odds for those who didn't use a tree at all. This finding, however, was not statistically significant at the $\alpha = 0.05$ level.

For the Time variable, the significant contrasts were between Homework and Midterm, and Homework and Final (p-value < 0.0001 for both cases). The odds of correctly solving the problem on the midterm and final were 9% and 15% of the odds of having solved the homework problem correctly, respectively. Finally, Total score for the semester is significantly correlated with whether the student answered the conditional probability problem correctly. In particular, each one-point increase in the total score increases the odds of correctly solving the conditional probability problem by 1.6%.

Because students did significantly better on the problem when it was done as part of the homework assignment, the GEE analysis was run a second time without the homework time point. In this analysis, Time was no longer a significant factor (p-value = 0.0691). Tree use and Total score were the two significant factors (p = 0.0011 and p = 0.0024, respectively). Presentation format and Working format were still not significant factors (p = 0.2755 and p = 0.8901, respectively). Those who completed a tree diagram were only 15.8 times more likely to correctly solve the two-stage problem compared to those who didn't use a tree at all (p = 0.0018), as compared to 24 times in the above analysis. The odds that a student who completed the tree correctly solved the problem were still more than 4 times the odds for those who did not finish the tree diagram (p = 0.0021).

4. CONCLUSIONS AND FUTURE DIRECTIONS

There are two main findings of this research: (a) neither the presentation format of probability problems nor the format in which they chose to work was associated with students' abilities to provide the correct response on two-stage conditional probability problems, and (b) students who created a completed tree diagram were more likely to give the correct response to a two-stage conditional probability problem than those who either did not attempt or did not complete a tree diagram. It should be stressed that these findings are the result of classroom research enacted by the instructor-researcher. The findings, therefore, may be a result of the particular pedagogy employed by the instructor and may not generalize to other classrooms. This research and its findings should, therefore, be considered as from a pilot study. The significant results, however, indicate that future research in this area might be fruitful. Therefore, directions for future research and the possible implications for teaching suggested by these findings are discussed in the next section.

4.1. FUTURE DIRECTIONS FOR RESEARCH

Although this study did not find significant differences in student performance based on the format in which the data were given, it is possible that the lack of effect is due to the instruction that the students experienced. It would be interesting to try to replicate this finding in other classes, particularly those in which the instructor did not use partitive frequencies when solving class examples. This would aid in answering the question of whether making format issues salient or teaching students to decode problems when given in different formats is a useful teaching technique and would inform statistics instructors about possible avenues of best practices for teaching.

Another research question raised by the findings of this study in the domain of student learning is the usefulness of representations. In this case, students' ability to correctly complete a tree diagram as a representation was associated with the ability to correctly answer the probability question. The first possible follow up study suggested would be to experimentally manipulate the tree diagram factor, either by teaching only some students to use a tree diagram or by requiring only some of the students to include a tree diagram when solving a similar problem. If tree diagrams are indeed shown to have value, this type of research could then be replicated with other representations, such as Venn diagrams and grids. Future research might also consider which representations are useful in statistics learning as well as investigating how representations are helpful to students.

In addition, in the domain of learning inference there is the question of the transferability of Sedlmeier's (1999) flexible urn model. Sedlmeier found that the flexible urn model was productive in helping subjects learn about sampling distributions both of proportions and of means. It would be useful to study whether the flexible urn model Sedlmeier developed for understanding sampling distributions could be used successfully in a large lecture introduction to statistics course. Inference, of which sampling distributions is the basis, is a notoriously difficult subject for students to understand. It would be a benefit for the teaching and learning of statistics if the use of a model such as the flexible urn were shown to be successful in helping students understand the concepts involved in inference.

4.2. IMPLICATIONS FOR TEACHING

The findings about format in this study are counter to those found in psychology studies. This may indicate that instruction can mitigate the format effects found in previous probability studies. It may be that when statistics instructors make format issues salient or teach students to decode problems when given in different formats, as was done for these students, the format effect of the problem does not appear. The results in the psychology literature showing subjects' increased ability to operate when data were given as frequency or counts had implications beyond the topic of probability. For example, in the topic of inference, data for inference about proportions can be represented as counts whereas data for inference about means cannot. The original findings that people reason more normatively when data are given as frequencies, therefore, lead to the hypothesis that students would find it easier to master inference about proportions than they would inference about means. Future research designed to uncover whether attention during instruction to format issues helps students to develop their abilities to master statistics content could, therefore, inform the teaching of statistical inference.

The finding that students' ability to correctly complete a tree diagram is associated with the ability to correctly answer the probability question indicates the possible importance of the inclusion of visual representations in the teaching of probability topics by statistics instructors. Because the use of a tree diagram was not an experimentally manipulated factor, it is not possible to attribute a causal relationship to the tree diagram factor. The large odds ratio associated with the students who completed a tree diagram suggests that there may be value in the tree diagram as a representation and that this is an avenue of research that should be explored further with statistics students. As mentioned previously, Cosmides and Tooby (1996) found a grid diagram to be a significant aid to subjects solving probability questions and Sedlmeier (1999) had success implementing a flexible urn representation when teaching sampling distributions. In particular, the finding of this paper in conjunction with those of Sedlmeier, who found more benefit with representations training than with rule-based training, suggest that instructors and textbooks of introductory statistics should consider providing more focus on diagrams, such as Venn and tree diagrams, and less focus on the rules and formulas for finding probability.

ACKNOWLEDGEMENTS

The authors would like to thank John P. (Jack) Smith III and Michael Steele. Their comments on early versions of the paper helped the authors to present the Bayesian problems in an understandable manner. The authors would also like to thank the two anonymous reviewers and the journal editor for their detailed comments that helped organize the writing of the background literature and ensure that the correct analysis was performed and then fully described in this paper.

REFERENCES

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

De Veaux, R., Velleman, P., & Bock, D. (2006). *Intro Stats* (2nd ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Evans, J. St.B. T., & Newstead, S. (1995). Creating a psychology of reasoning: The contribution of Peter Wason. In S. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason* (pp. 1-16). East Sussex, UK: Lawrence Erlbaum Associates, Ltd.

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

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

Hertwig, R., & Gigerenzer, G. (1999). The ‘conjunction fallacy’ revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12*(4), 275-305.

Johnson-Laird, P. (1994). Mental models and probabilistic thinking. *Cognition, 50*, 189-209.

Johnston, G., & Stokes, M. (1996). *Applications of GEE Methodology Using the SAS System.* Cary, NC: SAS Institute Inc.

Kutner, M. H., Nachtsheim, C.J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill Companies, Inc.

Liang, K.Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes, 82*(2), 217-236.

Macchi, L. (1995). Pragmatic aspects of the base rate fallacy. *The Quarterly Journal of Experimental Psychology, 48A*(1), 188-207.

Moore, D. (2003). *The Basic Practice of Statistics* (3rd ed.). New York: W. H. Freeman and Company.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stanovich, K. (1999). *Who is rational?: Studies of individual differences in reasoning.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Thompson, V. (1996). Reasoning from false premises: The role of soundness in making logical deductions. *Canadian Journal of Experimental Psychology, 50*(3), 315-319.

JENNIFER J. KAPLAN
443 Wells Hall
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824
USA

APPENDIX A

Homework Question (collected Oct 13):

Leah is flying from Boston to Denver with a connection in Chicago. The probability her first flight leaves on time is 0.15. If the flight is on time, the probability that her luggage with make the connecting flight in Chicago is 0.95, but if the flight is delayed, the probability that the luggage will make it is only 0.65.

a) What is the probability that her luggage arrives in Denver with her?
b) Suppose you pick her up at the Denver airport, and her luggage is not there. What is the probability that Leah's first flight is delayed?

Midterm Two Question (administered Oct 27):

Percentage Format:

A large company uses a test to determine whether new employees are likely to continue working at the company for longer than one year. The test predicts that 75% of new employees will stay with the company and 25% will leave. 20% of those predicted to stay actually leave and 80% of those predicted to leave actually leave.

a)Find the probability that an employee will actually leave after one year.
b) If a new employee leaves after the first year, what is the probability that he was predicted to leave?

Partitive Count Format:

A large company uses a test to determine whether new employees are likely to continue working at the company for longer than one year. The test predicts that 75 out every 100 new employees will stay with the company and 25 out of 100 will leave. 15 out of every 75 employees predicted to stay actually leave and 20 out of 25 of those predicted to leave actually leave.

Non-Partitive Count Format:

A large company uses a test to determine whether new employees are likely to continue working at the company for longer than one year. The test predicts that 3 out every 4 new employees will stay with the company and 1 out of 4 will leave. 1 out of every 5 employees predicted to stay actually leave and 4 out of 5 of those predicted to leave actually leave.

Exam Question (administered Dec 13):

Percentage Format:

A manufacturer of laundry detergent has introduced a new product that it claims to be more environmentally sound with a major advertising campaign. In an intensive survey, they find that 40% of people have seen the ad and 60% have not. Of the people who have seen the ad, 30% have bought the product. Of the people who have not seen the ad, 15% have bought the product.

a) Find the probability that a person chosen at random has bought the product.
b) If a person has bought the product, what is the probability that he saw the ad?

Partitive Count Format:

A manufacturer of laundry detergent has introduced a new product that it claims to be more environmentally sound with a major advertising campaign. In an survey of 400 people, they find that 160 people have seen the ad and 240 have not. Of the people who have seen the ad, 48 have bought the product. Of the people who have not seen the ad, 36 have bought the product.

Non-Partitive Count Format:

A manufacturer of laundry detergent has introduced a new product that it claims to be more environmentally sound with a major advertising campaign. In an intensive survey, they find that 2 out of every 5 people have seen the ad and 3 out of 5 have not. Of the people who have seen the ad, 3 out of 10 have bought the product. Of the people who have not seen the ad, 3 out of 20 have bought the product.

APPENDIX B

4.3. CALCULATION OF THE WITHIN SUBJECTS CORRELATION MATRIX

Let the vector of measurements on the i^{th} student be $Y_i = [Y_{i1}, Y_{i2}, Y_{i3}]'$ with corresponding vector of means $\mu_i = [\mu_{i1}, \mu_{i2}, \mu_{i3}]'$ and let V_i be an estimate of the covariance matrix of Y_i. The Generalized Estimating Equation for estimating is an extension of the independence estimating equation to correlated data and is given by

$$\sum_{i=1}^{88} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0$$

In addition V_i is modeled by using working matrix R(**α**) which is fully specified by the vector of parameters **α** in the following way: $V_i = \phi \mathbf{A}_i^{1/2} \mathrm{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, where ϕ is a dispersion parameter, $\mathbf{A}_i$ is an 3 ×3 diagonal matrix with $\mu_{it}(1- \mu_{it})$ as the t^{th} diagonal element, $t = 1, 2, 3$. Here we choose to use the unstructured working correlation matrix, therefore $\alpha_{jk} = \mathrm{Corr}(Y_{ij}, Y_{ik})$. The working correlation matrix is estimated in the iterative fitting process using the current value of the parameter vector, **β**, to compute appropriate functions of the

$$\text{Pearson residual. } r_{it} = \frac{Y_{it} - \mu_{it}}{\sqrt{\mu_{it}(1 - \mu_{it})}}$$

See Liang and Zeger (1986) for the detailed estimation procedure.

4.4. ALGORITHM FOR FITTING THE SPECIFIED MODEL USING GEES

The following is an algorithm for fitting the specified model using GEEs.

1. Compute an initial estimate of β, for example, with an ordinary generalized linear model assuming independence.
2. Compute the working correlations $\mathrm{R}_i(\alpha)$.
3. Compute an estimate of the covariance $V_i = \phi \mathbf{A}_i^{1/2} \mathrm{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$
4. Update β:

$$\beta_{r+1} = \beta_r - [\sum_{i=1}^{88} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta))]^{-1} [\sum_{i=1}^{88} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta))]$$

5. Iterate until convergence.

It is worth mentioning is that the parameter α in the working correlation matrix is regarded as nuisance so that the estimators of the regression coefficients and their standard errors on GEE are consistent and asymptotically normal even with mis-specified covariance structure.

See Johnston and Stokes (1996) for more details.