International Journal of Internet Science
2012, 7 (1), 37–54 ISSN 1662-5544

International Journal of Internet Science

IJIS.NET

Improving Internal Consistency in Conditional Probability Estimation With an Intelligent Tutoring System and Web-Based Tutorials

Christopher R. Wolfe[1], Christopher R. Fisher[1], Valerie F. Reyna[2], Xiangen Hu[3]

[1]*Miami University, USA,* [2]*Cornell University, USA,* [3]*The University of Memphis, USA*

Abstract: Three Web-based laboratory experiments explored the efficacy of three different Web-based tutorials designed to improve performance on Bayesian conditional probability estimation problems. In each experiment, participants estimated the probability of two events, and two conditional probabilities P(A|B) and P(B|A). Problems reflected five distinct relationships between two sets: identical sets, mutually exclusive sets, subsets, overlapping sets, and independent sets. Performance was measured against two benchmarks: internal inconsistency, a type of fallacy, and semantic coherence, a constellation of estimates of P(A), P(B), P(A|B), and P(B|A) that are consistent with the relationship among sets presented in the problem statement. As predicted by fuzzy-trace theory, in all three experiments, problems depicting identical sets, mutually exclusive sets, and independent sets yielded superior performance with respect to inconsistency and semantic coherence than problems depicting subsets and overlapping sets. In Experiment 1, a Web-based tutorial teaching the logic of the 2 × 2 table reduced internal inconsistency for overlapping sets problems. In Experiment 2, a Web-based tutorial using Euler diagrams was effective in reducing inconsistency and increasing semantic coherence for overlapping sets and subsets problems. Experiment 3 employed *AutoTutor Lite*, the first Web-based Intelligent Tutoring System with two-way interactions with people in natural language (English). AutoTutor Lite is cross-platform enabled with talking animated agents that converse with learners using Latent Semantic Analysis to "understand" natural language. AutoTutor Lite elicits verbal responses from the learner through a textbox and encourages them to further elaborate their understanding. AutoTutor Lite tutorial significantly reduced internal inconsistency on overlapping sets and subsets problems.

Keywords: AutoTutor Lite, semantic coherence, Web-based tutoring, conditional probabilities, intelligent tutoring system

Introduction

Technological advances, and theoretical progress in cognitive science, have led to a point where Artificial Intelligence is beginning to play an important role in Web-based "cognitive technologies" (Wolfe, 2006). We conducted three randomized controlled experiments on the efficacy of three Web-based tutors designed to increase the semantic and logical coherence (Wolfe & Reyna, 2010a, 2010b; Wolfe & Fisher, 2010) of conditional probability estimates. Fuzzy-trace theory (FTT; Reyna & Brainerd, 1995) provides the theoretical underpinning of these interventions. The first intervention was a static Web-based tutorial, the second a more

Address correspondence to Christopher R. Wolfe, Department of Psychology, Miami University, Oxford, OH 45056, USA, Phone: (+1) 513 529-5670, WolfeCR@muohio.edu
Updated second version, available since February 9, 2013. Previous version available from http://www.ijis.net/ijis7_1/ijis7_1_wolfe_et_al_v1.pdf

interactive Web-based tutorial, and the third an early version of AutoTutor Lite, a Web-based Intelligent Tutoring System with a talking pedagogical agent that interacts with learners in natural language (Graesser & McNamara, 2010).

Conditional probability estimation is at the core of Bayesian thinking which is becoming the subject of increased interest (e.g., Holyoak & Cheng, 2011). Conditional probability judgment is important because it helps us deal with uncertainty by updating our beliefs as new information is acquired, and allows us to engage in hypothetical thinking about potential future events. More generally, it helps us weigh potential risks and opportunities under varying circumstances. Given the importance of Bayesian thinking and conditional inference in domains such as medical decision making (Reyna, 2008; Reyna, Lloyd, & Whalen, 2001), the goal of this research and development effort was to decrease fallacious reasoning and increase semantic coherence in people's conditional probability judgments.

Semantic coherence in conditional probability estimation

There are two general methods to assess the quality of probability judgments: correspondence and coherence (Hammond, 2000). Correspondence refers to the empirical accuracy of probability judgments (Reyna & Adam, 2003). Coherence refers to the mathematical internal consistency of probability judgments. Unlike research on joint probability judgment, which uses conjunction and disjunction fallacies as coherence benchmarks, there are not corresponding fallacies in conditional probability judgments based on a pair of estimates. Thus, in the case of the famous Linda problem (Tversky & Kahneman, 1983) the joint probability Linda is a feminist bank teller should not exceed the probability that she is a bank teller. However, if the problem is reframed as one of conditional probabilities, as long as the probability that Linda is a feminist is greater than 0 and less than 1, then the probability that she is a feminist given or assuming that she is a bank teller could be anything from 0 to 1.

Although no pair of two estimates can be considered fallacious in and of itself (as is the case with joint probabilities) when people are asked to provide all four estimates of P(A), P(B), P(A|B) and P(B|A), the entire set of estimates can be evaluated for internal consistency relative to Bayes' theorem (see Appendix A). Moreover, a constellation of estimates – P(A), P(B), P(A|B) and P(B|A) – can be consistent with Bayes' theorem and semantically deficient at the same time.

To illustrate the relationship between internal inconsistency and semantic coherence, consider the following probability judgments regarding the probability that a student named Anna is taking various courses: P(Calculus) = .30, P(Mathematics) = .40, P(Calculus|Mathematics) = .33, and P(Mathematics|Calculus) = .25. Although this constellation of four judgments conforms to Bayes' theorem (see Appendix A), it clearly demonstrates a failure to incorporate the semantic relationship between calculus and mathematics – i.e., that calculus is a kind or subset of mathematics. Because calculus is a subset of mathematics, the proper conditional probabilities given P(Calculus) = .30, and P(Mathematics) = .40 should be P(Calculus|Mathematics) = .75 and P(Mathematics|Calculus) = 1.00. Internal consistency is a useful normative criterion for evaluating conditional probability judgments, but we can also investigate cognitive representations underlying conditional probability estimates, and set more stringent benchmarks, by extending the concept of semantic coherence from joint probability judgment, as first described in Wolfe and Reyna (2010a, 2010b), to conditional probability judgments.

There are five qualitatively different relationships between two sets and their respective conditional probabilities: identical sets, mutually exclusive sets, subsets, overlapping sets and independent sets (Wolfe & Fisher, 2010). Any constellation of judgments that does not match one of these patterns is inconsistent with Bayes' theorem. The three studies reported here asked participants to make estimates of P(A), P(B), P(A|B), and P(B|A) in controlled experiments where some participants received Web-based tutorials. Those estimates were subjected to a semantic coherence analysis assessing both internal inconsistency and semantic coherence. Appendix A provides worked examples and the semantic coherence formulae for each of these relationships based on Bayes' Theorem.

Theoretical underpinnings: Fuzzy-trace theory

Fuzzy-trace theory (FTT, Brainerd & Reyna, 1990; Reyna, 2004; Reyna & Brainerd, 1995; Wolfe & Reyna, 2010b) is a dual-processes approach to judgment and decision-making (Barbey & Sloman, 2007; Reyna, 2004; Sloman, 1996) that provides the theoretical underpinning for the Web-based interventions described below. From a FTT perspective, people may be characterized as gist processors. FTT uses the word gist much as it is used in everyday speech to mean the essential bottom-line meaning. When people process information, global gist-like patterns are encoded along with more precise verbatim information to create a multifaceted fuzzy-to-verbatim mental representation. Thus, individuals simultaneously hold multiple representations of events, with gist and verbatim representations being the two poles. There is a kinship between FTT and other dual-process theories, including those that propose System 1 and System 2 thinking (Barbey & Sloman, 2007; Kahneman, 2011), heuristic and analytic processing (Evans, 2006), and heuristic and rule-based reasoning (Ferreira, Garcia-Marques, Sherman, & Sherman, 2006). However, there are also important differences, and FTT makes novel predictions that have received strong empirical confirmation (Reyna & Brainerd, 2008).

A critical component of the theory is that people exhibit a strong preference to reason with the most essential, gist-like representations allowable for a given task. This is referred to as the fuzzy processing preference. Interestingly, rather than being a liability, research guided by FTT finds that gist processing is often the hallmark of mature, superior performance in cognitive development and expert knowledge in domains such as medicine. However, FTT also suggests that people are prone to predictable and systematic errors as a result of our cognitive architecture.

One source of such errors is denominator neglect – behaving as if one is ignoring the marginal totals (denominators) in a 2 × 2 table. As Reyna and Brainerd (1993, p. 28) note, people “often base probability judgments on comparisons between numerators, and, thereby, avoid having to keep track of the relationship between those numerators and the denominators in which they are included”. Denominator neglect can take the form of comparing numerators while ignoring relevant denominators, or attending to convenient but inappropriate denominators. Denominator neglect reduces cognitive load. However, it can lead to systematic errors such as ignoring relevant base rates (Wolfe, 1995), committing logical fallacies of conjunction and disjunction (Wolfe & Reyna 2010b), and confusing the conditional probability $P(\mathrm{A}|\mathrm{B})$ with $P(\mathrm{B}|\mathrm{A})$ (Reyna, 2004).

FTT suggests that people make probability judgments by integrating semantic knowledge with naïve probability theory (Reyna & Adam, 2003; Wolfe & Reyna, 2010b). With respect to conditional probability problems, we expect that performance will be relatively poor due to denominator neglect, the tendency to ignore normatively relevant denominators. Confusion about relations among sets or classes is responsible for denominator neglect (Reyna & Brainerd, 2008). To illustrate, consider the conditional probabilities $P(\mathrm{A}|\mathrm{B})$ and $P(\mathrm{B}|\mathrm{A})$. In estimating each of these, the numerator is the same: $P(\mathrm{B}\ \textsc{and}\ \mathrm{A})$. For $P(\mathrm{A}|\mathrm{B})$ the relevant denominator is $P(\mathrm{A})$ whereas for $P(\mathrm{B}|\mathrm{A})$ the relevant denominator is $P(\mathrm{B})$. Ignoring either marginal total (denominator), $P(\mathrm{A})$ or $P(\mathrm{B})$, in estimating $P(\mathrm{A}|\mathrm{B})$ or $P(\mathrm{B}|\mathrm{A})$ often leads to logical fallacies (see Wolfe & Reyna, 2010b for a discussion of denominator neglect in the context of joint probabilities and Reyna & Brainerd, 2008 for a detailed overview).

Predictions

FTT leads to a number of predictions about semantic coherence and internal inconsistency on conditional probability problems. First, we predict that estimating conditional probabilities will be difficult – indeed, even more difficult than estimating joint probabilities. This is because such estimates require reasoning about hierarchically nested sets and normatively, one must estimate the joint probability of two events as part of the process of estimating their conditional probabilities. Second, people will be sensitive to the semantic content of conditional probability problems with problems depicting subsets and overlapping sets being particularly difficult. Estimating conditional probabilities for overlapping sets and subsets requires division because in each case one class or event contains partial information about the other. The third general prediction is that interventions that draw attention to the appropriate denominators for these hierarchically nested categories will increase semantic coherence and reduce internal inconsistency, including teaching people with Euler diagrams or 2 × 2 tables.

FTT suggests a conundrum for those wishing to develop a Web-based tutorial intervention for every type of conditional probability estimation problem. Although addressing denominator neglect is clearly helpful, performance on some kinds of problems may be addressed by simple gist-based heuristics whereas others require more attention to specific details. Conditional probability problems involving independent sets are simplified by the bottom line understanding that $P(\mathrm{A})$ provides no information about $P(\mathrm{B})$ and thus $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A})$. The appropriate gist of conditional probabilities for identical sets it that $P(\mathrm{A}|\mathrm{B}) = 1.0$ and the gist of mutually exclusive sets is that $P(\mathrm{A}|\mathrm{B}) = 0$. However, to achieve semantic coherence on problems depicting overlapping sets, one must recognize that $P(\mathrm{A})$ and $P(\mathrm{B})$ provide limited but useful information about the conditional probabilities and reason with the specific verbatim details that $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A}\ \textsc{and}\ \mathrm{B}) / P(\mathrm{B})$ and that $P(\mathrm{B}|\mathrm{A}) = P(\mathrm{A}\ \textsc{and}\ \mathrm{B}) / P(\mathrm{A})$. In the case where A is a subset of B, one may have the intuition (gist) that $P(\mathrm{B}|\mathrm{A}) = 1.0$ however, to achieve semantic coherence one most also reason with the verbatim information that $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A}\ \textsc{and}\ \mathrm{B}) / P(\mathrm{B})$. The conundrum is that interventions that improve performance by drawing attention to the four cells of a 2 × 2 table, $P(\mathrm{A}\ \textsc{and}\ \mathrm{B})$, $P(\mathrm{A}\ \textsc{and}\ \neg\ \mathrm{B})$, $P(\neg\mathrm{A}\ \textsc{and}\ \mathrm{B})$, and $P(\neg\mathrm{A}\ \textsc{and}\ \neg\mathrm{B})$ will generally not help gist-based performance on easier problem types. Thus, the fourth general prediction is that interventions that improve performance with subsets and overlapping sets will be ineffective on problems depicting identical sets, mutually exclusive sets, and independent sets.

Research overview

Below we report the results of three experiments about semantic coherence and internal inconsistency in conditional probability estimation. The basic procedure was to present participants with a number of problems depicting identical sets, mutually exclusive sets, subsets, overlapping sets, and identical sets. These problems were all pilot tested to ensure that participants' beliefs about the relationships among sets were consistent with the problem materials, and most have been used in previously-published research (Wolfe & Reyna, 2010b). For each problem, participants were asked to make four estimates, *P*(A), *P*(B), *P*(A|B) and *P*(B|A) with all estimates being collected on a single computer screen. The constellation of these four estimates was scrutinized for semantic coherence using the formulae and procedures described by Fisher and Wolfe (2011). In each of the experiments, participants were randomly assigned to a control group or one or more experimental groups with the experimental groups receiving a Web-based training intervention motivated by FTT. Thus, each experiment includes between-subject comparisons among interventions and controls and within-subject comparisons among types of conditional probability estimation problems. The first experiment builds directly on the interventions used in Wolfe and Reyna's (2010b) research on semantic coherence and fallacies in estimating joint probabilities.

Experiment 1: Web-based tutorial analogy crossed with 2 ✕ 2 table

This first experiment is similar to the Wolfe and Reyna (2010b) Experiment 3, with the materials changed to reflect conditional probabilities rather than joint probabilities. Wolfe and Reyna's (2010b) Experiment 3 examined semantic coherence in joint probability problems using four versions of a Web-based tutorial, formed by crossing the presence or absence of a 2 ✕ 2 table with the presence or absence of pedagogic analogies. They found that the 2 ✕ 2 table intervention significantly reduced fallacies and increased semantic coherence. The pedagogic analogies increased semantic coherence but did not reduce fallacies. Wolfe and Reyna (2010b) also found a main effect for problem type, with participants most coherent on identical set problems and least coherent on problems depicting subsets.

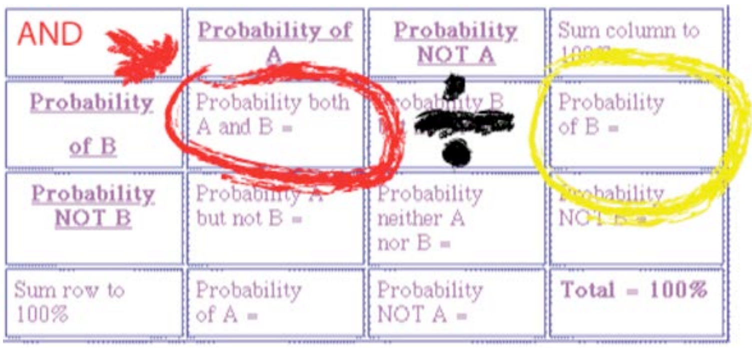
In the present experiment, we predicted a significant main effect for problem type with identical sets, mutually exclusive and independent sets yielding the highest levels of semantic coherence and the lowest levels of inconsistent responding, and subsets and overlapping sets producing the lowest levels of semantic coherence and the highest levels of inconsistent responding. We also predicted that the 2 ✕ 2 table and analogy interventions would increase semantic coherence and reduce inconsistent responding. Finally, we predicted an interaction between the 2 ✕ 2 table intervention and problem type, specifically that the 2 ✕ 2 table would improve performance with subsets and overlapping sets more than on problems depicting identical sets, mutually exclusive sets, and independent sets. We made no such prediction for analogies.

Materials

The problems used in this experiment were the same as those used in the Wolfe and Reyna's (2010b) Experiment 3, except that, in addition to P(A) and P(B), we asked for the conditional probabilities of P(A|B) and P(B|A) rather than asking for the conjunctive and disjunctive probabilities ("AND" and "OR"). There were two problems for each problem type: identical sets, mutually exclusive sets, subsets, overlapping sets, and independent sets – a special case of overlapping sets where A and B are completely orthogonal such that *P*(A) and *P*(B) provide no information about one another. One independent sets problem was about the relationship between the stock market and a baseball team winning or loosing, and the other was about the relationship between flipping a coin and the probability of rain. With the exception of the new independent sets problems, all of the other problems have been used in previously published research and pilot tested to ensure that participants' perceptions about the relationships among sets (mutually exclusive, overlapping, etc.) match our own. An illustration of our approach to asking for conditional probabilities can be found in Appendix B.

Method

Although the experiment was conducted on the Web, people participated in the laboratory individually or in small groups. Participants were 121 Miami University undergraduates who received partial course credit for their participation. Participants were randomly assigned to one of four groups in a randomized block design. Half of the participants received a Web-based tutorial in the use of a 2 ✕ 2 table, and a picture of a 2 ✕ 2 table with each problem and the other half did not. Crossed with the presence or absence of the 2 ✕ 2 table, half of the participants received an analogy overview and a simple pedagogic analogy with each problem and the other half did not. Figure 1 presents a sample of the 2 ✕ 2 table intervention. It is important to note that the participants did not complete the 2 ✕ 2 tables or have access to scrap paper or calculators. Instead, the 2 ✕ 2 table was provided to aid participants in the underlying logic of its use. Participants in the pedagogic analogies conditions received a brief overview on the use of analogies and then a specific analogy for each problem such as, “Consider that the relationship between cats and mammals is like the relationship between roses and flowers” in case of subsets.



AND	Probability of A	Probability NOT A	Sum column to 100%
Probability of B	Probability both A and B =	Probability B	Probability of B =
Probability NOT B	Probability A but not B =	Probability neither A nor B =	Probability NOT B =
Sum row to 100%	Probability of A =	Probability NOT A =	Total = 100%

Figure 1. 2 × 2 Table used in making conditional probability estimates (from http://think.psy.muohio.edu /Estimation/Both.htm).

Problems were presented in two counterbalanced random orders. For each problem, participants estimated *P*(A), *P*(B), *P*(A|B), and *P*(B|A). For each set of four estimates, we assessed semantic coherence and inconsistent responding using the Fisher and Wolfe (2011) procedures and conducted comprehensive 2 × 2 repeated measures ANOVAs for the table and analogy interventions and their interaction following a similar procedure to that used by Wolfe & Reyna (2010b). In all the analyses reported throughout the paper, responses that were semantically coherent were recorded as one, zero otherwise. The mean semantic coherence was computed per participant per problem type. Thus, values reported in the tables may be interpreted as the proportion of responses that are semantically coherent, and readers can compare results across experiments with different numbers of problems. The same logic was applied to the other analyses including inconsistent sets and conversion errors.

Results

Table 1 presents the mean semantic coherence and mean inconsistent responses by condition and problem type. As predicted, there was a significant main effect for problem type, *F*(3, 351) = 29.92, *p* < .0001. Subset and Overlapping sets problems had very low levels of semantic coherence that were far below those for identical sets, mutually exclusive sets and independent sets, with *p* < .0001 for paired comparisons. Subsets and overlapping sets problems also produced significantly higher levels of inconsistency than the other problem types.

Table 1
Mean Semantic Coherence and Mean Inconsistent Responses per Problem by Problem Type (SD in Parentheses)

	Mean semantic coherence				**Mean inconsistent responses**			
Problem Type	**Analogy & table (*N* =33)**	**Analogy only (*N* = 27)**	**Table only (*N* = 30)**	**None/ control (*N* = 31)**	**Analogy & table (*N* =33)**	**Analogy only (*N* = 27)**	**Table only (*N* = 30)**	**None/ control (*N* = 31)**
Identical sets	.74 (.42)	.78 (.35)	.65 (.42)	.82 (.33)	.12 (.28)	.09 (.24)	.13 (.29)	.13 (.32)
Mutually exclusive sets	.68 (.45)	.54 (.44)	.58 (.44)	.48 (.44)	.21 (.38)	.24 (.32)	.23 (.34)	.35 (.39)
Subsets	.15 (.27)	.17 (.28)	.03 (.13)	.03 (.13)	.54 (.40)	.69 (.34)	.55 (.36)	.60 (.44)
Overlapping sets	.14 (.29)	.06 (.16)	.10 (.20)	.05 (.15)	.39 (.37)	.59 (.39)	.50 (.37)	.52 (.35)
Independent sets	.65 (.39)	.54 (.39)	.55 (.42)	.71 (.40)	.21 (.35)	.35 (.36)	.30 (.34)	.16 (.30)

The main effect for table was not significant, *F*(1, 117) = 2.15, *p* = .15, nor was the main effect for analogy, *F*(1, 117) < 1, or the table by analogy interaction, *F*(1, 117) = 2.23, *p* = .14. However, as predicted the problem type by table interaction was significant, *F*(3, 351) = 3.43, *p* = .017, η^2 = .029. The table intervention led to significantly higher levels of semantic coherence on overlapping sets problems *F*(1, 117) = 5.41. *p* = .022, η^2 = .043 (see Table 1). The mean semantic coherence on overlapping sets problems with the table intervention was 0.12 compared to 0.05 for overlapping sets problems without the table intervention. No other main effects or interactions were significant.

On problems depicting overlapping sets, about 50% of the response sets were internally inconsistent. About 40% of these (20% of all response sets for overlapping sets) were “conversion errors” where participants inappropriately estimated *P*(A|B) = *P*(B|A) (Wolfe, 1995). The proportion of conversion errors was significantly higher on problems depicting overlapping sets than those depicting mutually exclusive sets, subsets, or identical sets $t(120) = 2.73$, $p = .004$, $\eta^2 = .060$. We also found that about 26% of the response sets were internally inconsistent for problems depicting mutually exclusive sets. Of these, about 39% (10% of all response sets for mutually exclusive sets) resulted from errors of excessive overlap errors where both conditional probabilities were estimated to be 0, but the sum of the two mutually exclusive probability estimates exceeded 1.0.

Discussion

As predicted, overlapping sets and subsets were much more difficult than other problem types. Fuzzy-trace theory suggests that this is because conditional probabilities of these types require division and keeping track of multiple potential denominators (Reyna & Brainerd, 2008). Following Fisher and Wolfe (2011) we were able to rule out simple rounding errors by systematically loosening the criteria. We still observed the same pattern of results even when we set the rounding error to ±.05 (i.e., when the expected estimate of P(B|A) was .75 or 75%, responses ranging from .70 to .80 were considered semantically coherent; see Fisher & Wolfe, 2011 for more details). Instead, many participants provided evidence of conversion errors (Wolfe, 1995) incorrectly estimating P(A|B) = P(B|A) particularly on overlapping sets problems. We also found that a large portion of errors on problems depicting mutually exclusive sets were excessive overlap errors. To illustrate, on a problem where Charles was stung by an unidentified bug, Participant 57 estimated P(bee) = .99, P(wasp) = .10, P(bee|wasp) = 0, P(wasp|bee) = 0. It appears that the participant recognized that bees and wasps are mutually exclusive (i.e., that a single sting can not be caused by both a wasp and a bee) but he or she apparently did not recognize that under those circumstances if the probability that it is a bee is .99 then the probability that it is a wasp can not exceed .01.

The Web-based tutorial was less effective than predicted. Although we found the predicted interaction between problem type and intervention, the 2 × 2 table intervention was effective only for overlapping sets problems, not subsets, and the effects on inconsistency were not significant. The interventions also did not reduce conversion errors. Even in the 2 × 2 table conditions, for subsets and overlapping sets absolute semantic coherence was quite low and inconsistency was fairly high. These findings prompted us to try an intervention involving Euler diagrams in Experiment 2. Euler diagrams are attractive from a FTT perspective because they preserve the gist of the problems while exposing the hierarchical subset relationships.

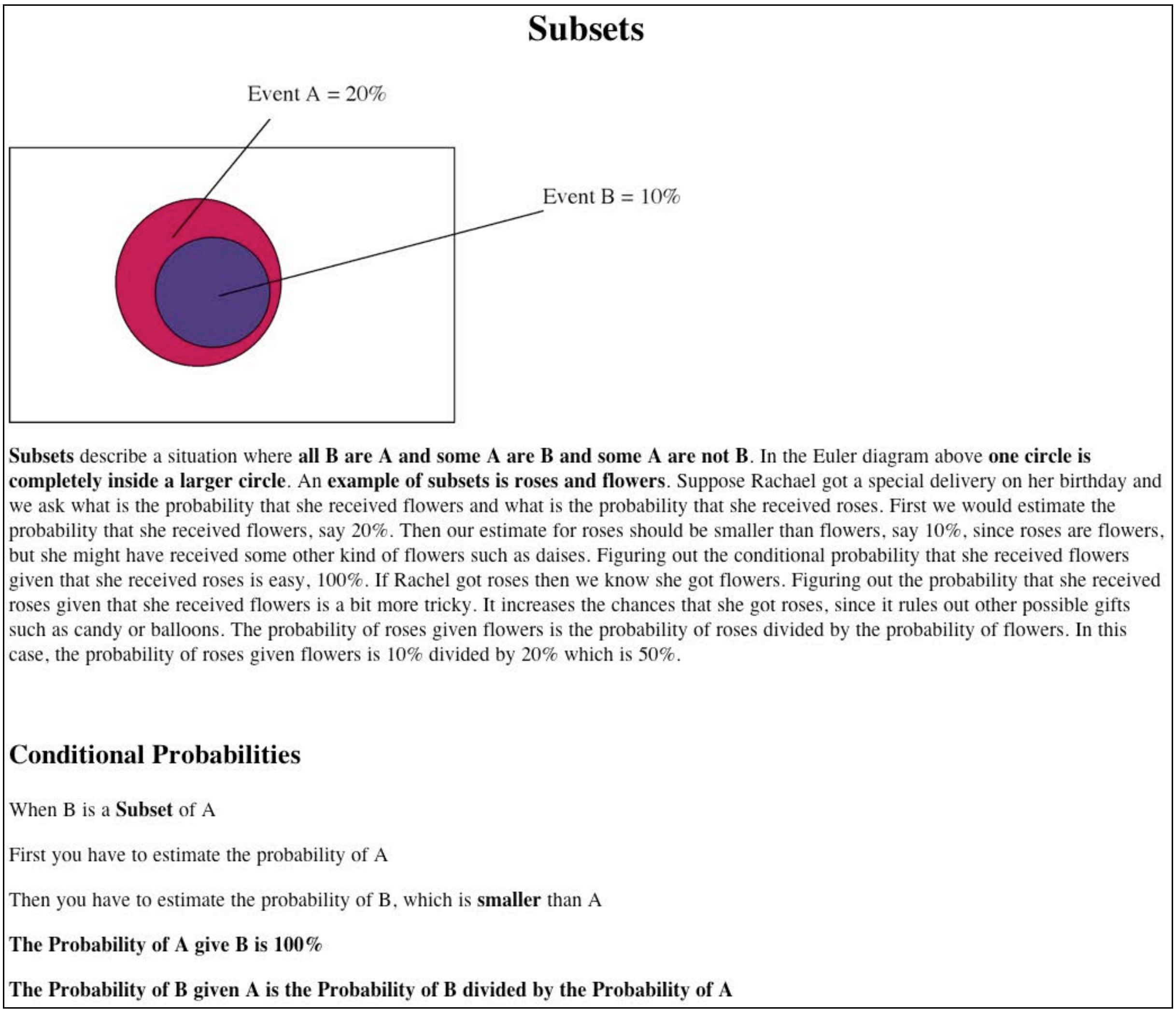
Experiment 2: Web-based tutorial with euler diagrams and practice problems

FTT suggests that semantic coherence is especially difficult for subsets and overlapping sets because the hierarchical nested relationships among classes leads to denominator neglect (Wolfe & Reyna, 2010a, 2010b). Thus, any intervention that draws attention to the correct relationship among classes or sets and reduces denominator neglect should increase semantic coherence. Euler Diagrams were chosen as the basis for this Web-based intervention because they address the gist of conditional probability judgment. Much like a 2 × 2 table, Euler diagrams provide a method for tracking the class inclusion relationships between sets and draw attention to relevant denominators. However, whereas the 2 × 2 table provide for assistance in processing, Euler diagrams preserve the gist of the problem by representing qualitative relationship between sets diagrammatically. Euler Diagrams consist of circles enclosed within a rectangle that represents all possible outcomes. Each circle represents the probability that a particular event will occur. Like Venn diagrams, joint probabilities are represented by the overlapping portions of two circles. Unlike Venn diagrams, the probability of an event is proportionate to the area of the rectangle that each circle occupies.

In this experiment, we predicted that identical sets, mutually exclusive and independent sets would yield the highest levels of semantic coherence and the lowest levels of inconsistent responding, and subsets and overlapping sets producing the lowest levels of semantic coherence and the highest levels of inconsistent responding. We also predicted that a Web-based intervention using Euler diagrams would increase semantic coherence and reduce inconsistent responding relative to a control intervention. Finally, we predicted that the Web-based Euler diagram tutorial would improve performance with subsets and overlapping sets relative to controls more than on problems depicting identical sets, mutually exclusive sets, and independent sets.

Method and materials

The materials used in Experiment 2 were similar to those used in Experiment 1, with the following exceptions. First, to provide further confidence in the generalizability of our findings, additional problems for each problem type were included. In total, there were four problems each for identical sets and independent sets, and five problems each for the more complex problems of interest representing mutually exclusive sets, subsets, and overlapping sets. Second, Experiment 2 included problems featuring a causal relationship between the component events. Many theorists argue that people make essentially rational Bayesian inferences by incorporating causal reasoning (Holyoak & Cheng, 2011). Five problems describing a statistical causal relationship (i.e., neither sufficient nor necessary) between the component events were adopted from Crisp and Fenney (2009). An example of this problem type is the relationship between the taxation and consumption of cigarettes, with the general perception being that an increase in taxes should result in lower cigarette consumption. These problems are mathematically consistent with overlapping sets. Two additional problems we developed described either a sufficient or necessary condition for causation, each of which is mathematically consistent with subsets. As an example of a sufficient condition for causation, consider the relationship between a car starting and an obstructed fuel line. An obstructed fuel line will prevent the car from starting. However, a multitude of other problems could also prevent a car from starting. As an example of a necessary condition for causation, consider the relationship between plant growth and adequate sunlight. A plant will not grow in the absence of sunlight. However, the presence of sunlight does not guarantee the growth of a plant.



Subsets

Subsets describe a situation where **all B are A and some A are B and some A are not B**. In the Euler diagram above **one circle is completely inside a larger circle**. An **example of subsets is roses and flowers**. Suppose Rachael got a special delivery on her birthday and we ask what is the probability that she received flowers and what is the probability that she received roses. First we would estimate the probability that she received flowers, say 20%. Then our estimate for roses should be smaller than flowers, say 10%, since roses are flowers, but she might have received some other kind of flowers such as daises. Figuring out the conditional probability that she received flowers given that she received roses is easy, 100%. If Rachel got roses then we know she got flowers. Figuring out the probability that she received roses given that she received flowers is a bit more tricky. It increases the chances that she got roses, since it rules out other possible gifts such as candy or balloons. The probability of roses given flowers is the probability of roses divided by the probability of flowers. In this case, the probability of roses given flowers is 10% divided by 20% which is 50%.

Conditional Probabilities

When B is a **Subset** of A

First you have to estimate the probability of A

Then you have to estimate the probability of B, which is **smaller** than A

The Probability of A give B is 100%

The Probability of B given A is the Probability of B divided by the Probability of A

Figure 2. Euler diagram used in making conditional probability estimates (from http://think.psy.muohio.edu /Euler/).

Both the control and experimental conditions included tutorials that were designed to be approximately equal in terms of reading time. The Web-based Euler diagram tutorial provided a worked example of each type of relationship with an Euler diagram, didactic text, and a summary of rules for conditional probabilities (see Figure 2 for an example involving subsets). Each of the five possible relationships between sets was described in detail and accompanied by Euler diagrams. To provide participants with practice at categorizing problems, a short multiple-choice test was presented upon the completion of the tutorials. Participants could not proceed until the correct answer was selected, after which point a brief explanation of the correct answer was provided. The multiple-choice test for the Euler tutorial consisted of ten problems, two for each of the five set relationships. Participants identified the set relationship described in each problem by selecting options labeled with text and a corresponding thumbnail picture of an Euler diagram. The control tutorial was adapted from Wolfe, Britt, Petrovic, Albrecht, and Kopp (2009) by slightly modifying the title and introductory paragraph. It pertained only to written argumentation and was approximately equal to the Euler tutorial in terms of complexity.

A total of 144 Miami University undergraduates participated in exchange for partial course credit in introductory psychology. Participants were randomly assigned to the control tutorial or Euler diagram tutorial in a randomized block design. As in Experiment 1, participants completed the Web-based experiment in the laboratory in small groups no larger than six. After the tutorials, participants proceeded to the conditional probability judgment task in which the problem order was randomized for each participant. As with the 2 ✕ 2 table in Experiment 1, participants who received the Euler tutorial applied the logic of Euler Diagrams, but did not draw them.

Results

Table 2 presents the mean semantic coherence and mean inconsistent responses by condition and problem type. As predicted, there were large differences among problem types (see Table 2). For problems depicting identical sets, mutually exclusive sets, and independent sets, over 65% of the responses were semantically coherent, and less than 30% were inconsistent. For problems depicting overlapping sets and subsets – whether or not a causal relationship was depicted in the problem materials – less than 10% of the response sets were semantically coherent and (in the control condition) over 75% of the responses were inconsistent.

Table 2
Mean Semantic Coherence and Mean Inconsistent Responses per Problem by Problem Type (SD in Parentheses)

	Mean semantic coherence		**Mean inconsistent responses**	
Problem type	**Control (*N* = 72)**	**Experimental (*N* = 72)**	**Control (*N* = 72)**	**Experimental (*N* = 72)**
Identical sets	.66 (.31)	.74 (.31)	.19 (.27)	.15 (.23)
Mutually exclusive sets	.73 (.28)	.74 (.31)	.22 (.25)	.16 (.25)
Subsets	.04 (.11)*	.11 (.14)	.83 (.22)*	.73 (24)
Overlapping sets	.07 (.13)*	.12 (.15)	.75 (.27)*	.62 (.29)
Independent sets	.65 (.36)	.59 (.36)	.30 (.33)	.27 (.31)
Causal statistical	.06 (.12)	.08 (.13)	.77 (.24)	.72 (.25)
Causal necessary or sufficient	.02 (.10)	.01 (.08)	.81 (.28)	.81 (.33)

*$p < .05$

As predicted, the Euler diagram intervention significantly improved performance on some types of problems, without having a significant effect on others. On problems depicting overlapping sets, compared to the control group, the Euler diagram increased semantic coherence, $F(1, 142) = 4.59$, $p = .034$, $\eta^2 = .031$ and reduced inconsistency, $F(1, 142) = 8.09$, $p = .005$, $\eta^2 = .054$. Similarly, problems depicting subsets, compared to the control group, the Euler diagram increased semantic coherence, $F(1, 142) = 9.26$, $p = .003$, $\eta^2 = .061$ and reduced inconsistency, $F(1, 142) = 7.04$, $p = .009$, $\eta^2 =.047$. However, the intervention was not significantly effective in improving performance on problems depicting identical sets, mutually exclusive sets, or independent sets. With respect to semantic coherence the Euler diagram intervention did not have a significant effect on problem depicting identical sets $F(1, 142) = 1.20$, $p = 0.141$, mutually exclusive sets $F(1, 142) < 1$, or independent sets $F(1, 142) < 1$. With respect to internal inconsistency the Euler diagram intervention did not have a significant effect on problem depicting identical sets $F(1, 142) = 1.18$, $p = 0.278$, mutually exclusive sets $F(1, 142) = 2.36$, $p = 0.127$, or independent sets $F(1, 142) < 1$. It was also ineffective for problems presenting a causal relationship whether they depicted a sufficient or a necessary relationship, or a causal statistical relationship.

As can be seen in Table 2, in the control group 75% of the responses on problems depicting overlapping sets were inconsistent. Of these, about 28% (21% of all response sets for overlapping sets) were "conversion errors" where participants inappropriately estimated $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{B}|\mathrm{A})$. These were not affected by the experimental intervention, $F(1, 143) = 1.58$ $p = .21$. Conversion errors were significantly more common on problems depicting overlapping sets than for identical sets, mutually exclusive sets, subsets, and independent sets, $t(143) = 3.65$, $p < .001$, $\eta^2 = .089$. Such errors were slightly, but not significantly more common on overlapping sets problems than on problems depicting a statistical causal relationship, $t(143) = 1.75$, $p = .08$. In the control group, 77% of the responses were inconsistent on problems a depicting statistical causal relationship, and of these 19% (21% of all response sets) were conversion errors. Turning to mutually exclusive sets, in the control group we found that about 22% of the response sets were internally inconsistent. Of these, about 46% (10% of all response sets for mutually exclusive sets problems) resulted from excessive overlap errors where both conditional probabilities were estimated to be 0, but the sum of the two mutually exclusive probability estimates exceeded

1.0. The Euler diagram significantly reduced these excessive overlap errors from a mean of .10 (*SD* = .13) to a mean of .05 (*SD* = .11), $F(1, 143) = 5.70$, $p = .018$, $\eta^2 = .039$.

Discussion

As in Experiment 1, we found more errors in problems depicting overlapping sets and subsets than for other problem types. Also as in Experiment 1, we can rule out simple rounding errors as providing the locus of this effect because we obtained comparable results even when setting a very lax criterion of ±.05 for rounding errors (Fisher & Wolfe, 2011). Particularly on problems depicting overlapping sets, a large portion of participants provided evidence of conversion errors (Wolfe, 1995) incorrectly estimating $P(A|B) = P(B|A)$. Once again, on mutually exclusive sets problems we found that a large proportion of responses were excessive overlap errors. However, unlike Experiment 1, in this case the Euler diagram intervention significantly reduced these errors. FTT suggests that conversion errors are the result of denominator neglect (Reyna, 2004; Wolfe, 1995) and thus interventions that help people attend to the relevant denominators should reduce conversion errors and thus reduce inconsistence and increase semantic coherence.

In this experiment we added problems depicting a statistical causal relationship as well as causal necessary and sufficient relationships (Crisp & Fenney, 2009). However, the inclusion of causal relationships had little effect on performance. As can be seen in Table 2, compared to other problems depicting subsets and overlapping sets the degree of inconsistency and semantic coherence is quite comparable for the causal problems. One difference is that the effects of the Euler diagram intervention did not rise to the level of statistical significance for these causal problems. However, given that the direction of the trend is comparable for causal and non-causal problems, it is probably wise not to make too much of this distinction.

The Euler diagram intervention appears to be effective for subset and overlapping sets problems. As in Experiment 1, it was effective in reducing inconsistency for overlapping sets problems. However, in addition, it also increased semantic coherence on overlapping sets problems, and both increased semantic coherence and decreased inconsistency for problems depicting subsets. These results are consistent with FTT. However, FTT also predicts that teaching participants the logic of the 2 × 2 Table (Wolfe & Reyna, 2010b) or relative frequencies should also improve performance providing that the intervention addresses denominator neglect. An alternative explanation is that unlike in Experiment 1, participants in Experiment 2 underwent an answer-until-correct test of their ability to categorize the different problems as part of their tutorial. It is therefore plausible that Experiment 2, participants were better prepared for the final questions for this reason. As previously noted, conditional probabilities are more difficult than joint probabilities because, for subsets and overlapping sets, estimating joint probabilities is part of the process of estimating conditional probabilities.

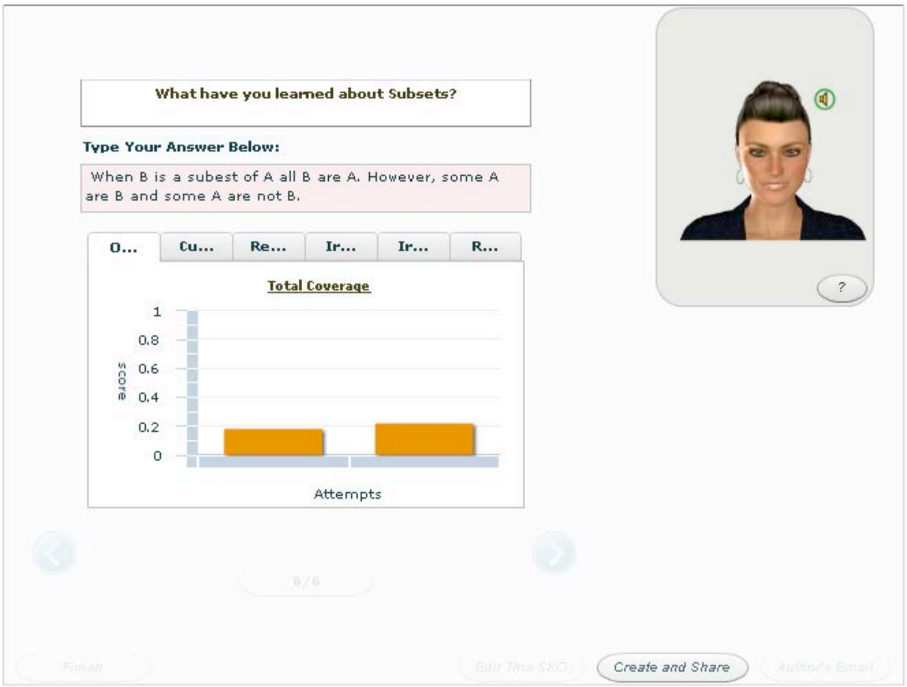
Experiment 3: Web-based intelligent tutoring system

The best way to teach people complex conceptual skills and knowledge is arguable human one-on-one tutoring, with good human tutors often producing gains of two standard deviations over standard classroom practice (Bloom, 1984; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001). Good human tutors engage in active dialogue with the learner encouraging him or her to elaborate their responses. This approach allows the learner to engage in self-explanation and develop rich and subtle cognitive representations. It also permits the tutor to uncover and address misconceptions. AutoTutor Lite is a Web-based intelligent tutoring system that tries to reproduce this success by replicating the kinds of interactions between human tutors and students.

The purpose of Experiment 3 was to test the efficacy of a Web-based Intelligent Tutoring System in reducing internal inconsistency and increasing semantic coherence in conditional probability estimation. AutoTutor Lite (ATL) is an Intelligent Tutoring System that lets people interact with it over the Web using an ordinary browser. ATL is cross-platform enabled and specifically designed to handle large numbers of users across different platforms. ATL uses Artificial Intelligence Markup Language to handle questions, and uses learner's characteristic curves to handle dialog moves (Hu, Han, & Cai, 2008; Hu & Martindale, 2008). ATL has a talking animated agent interface (Graesser & McNemara, 2010). It converses with users based on expectations using hints and elaboration. To the best of our knowledge, ATL is the first Web-based Intelligent Tutoring System that allows learners to interact with it through the use of natural language, in this case English. ATL uses Latent Semantic Analysis to "understand" natural language, present users with images, sounds, text, and video.

ATL elicits verbal responses from the learner and encourages them to further elaborate their understanding. ATL can thus be used to encourage self-explanation (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994). Through a natural language dialogue with the learner, ATL guides the learner toward a set of expectations. With ATL, tutorials are built from units called SKOs (Sharable Knowledge Objects). Each

SKO presents materials to the learner didactically and then solicits a verbal response from the learner. The didactic presentation is made by an animated talking agent with the ability to present text, still images, movie clips, and sounds. ATL currently has 17 avatars to choose from. Learner responses can take many forms including fill in the blank, multiple choice, and matching, but the heart of the approach is verbal responses in the form of self-reflection.



What have you learned about Subsets?

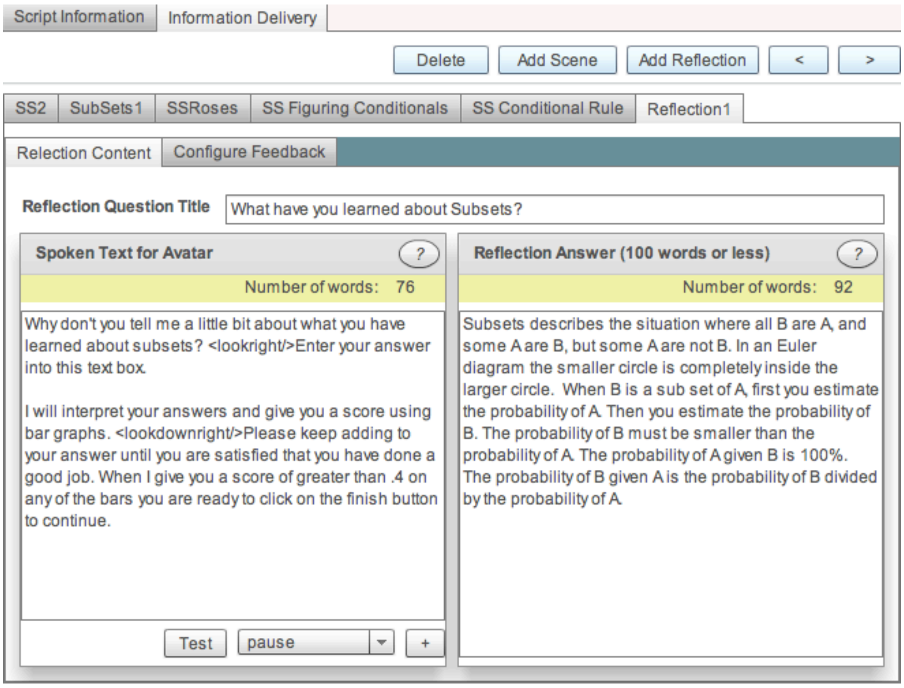
Type Your Answer Below:

When B is a subest of A all B are A. However, some A are B and some A are not B.

Figure 3. Interacting with AutoTutor Lite using a dialogue box (from http://www.x-in-y.com/sko/html/ATL.html ?GAE=skodev2010&guid=5288e86f-da84-430b-a096-1dec09a95d7e#).

Figure 3 is a screen shot from an ATL tutorial that shows an animated agent that has just asked the learner the question orally, and with screen text, “What have you learned about subsets?”. The learner has responded by typing in the textbox, “When B is a subset of A all B are A. However, some A are B and some A are not B.”. The bar graph shows the number of “turns” or sentences on the X axis, in this case two bars scoring the two sentences for total coverage. The total semantic similarity score is represented by the height of the bars on the Y axis. Attempts are turns or sentences triggered by a period (.) or return. In this case, there have been two turns. After the first turn the learner received a score of about 0.18 and after the second turn the total score increased to about 0.22. This indicates that the learner’s response captures some of the system’s expectations, but not all of them. As the learner continues to add text he or she will receive additional feedback.

Figure 4 is a screenshot of the authoring tools for ATL for the SKO described above. The box in the lower left corner of the figure shows the text spoken by the avatar, “Why don’t you tell me a little about what you have learned about subsets…”. The commands in angle brackets control the actions of the avatar. In this case, <lookright/> makes the avatar “look” at the textbox, and <lookdownright/> makes the avatar look down and right at the histogram. The authoring tools for ATL provide 32 such commands including shake head, eyes wide, confused, and flirty. The box in the lower right shows the expectations for this question in the form of a reflection answer, in this case, “Subsets describe the situation where all Bare A and some A are B, but…”. This was the expectation text used in the present study using AutoTutor Lite. However, subsequent experience suggests that it is better to strip out small words such as “the” and “of,” and to avoid redundancy. Another limitation of this study is that we did not provide verbal feedback to participants. Rather, we simply asked participants to, “keep adding to your answer until you are satisfied that you have done a good job. When I give you a score of greater than .4 on any of the bars you are ready to click on the next arrow button to continue.”.



Reflection Question Title: What have you learned about Subsets?

Spoken Text for Avatar

Number of words: 76

Why don't you tell me a little bit about what you have learned about subsets? <lookright/>Enter your answer into this text box.

I will interpret your answers and give you a score using bar graphs. <lookdownright/>Please keep adding to your answer until you are satisfied that you have done a good job. When I give you a score of greater than .4 on any of the bars you are ready to click on the finish button to continue.

Reflection Answer (100 words or less)

Number of words: 92

Subsets describes the situation where all B are A, and some A are B, but some A are not B. In an Euler diagram the smaller circle is completely inside the larger circle. When B is a sub set of A, first you estimate the probability of A. Then you estimate the probability of B. The probability of B must be smaller than the probability of A. The probability of A given B is 100%. The probability of B given A is the probability of B divided by the probability of A.

Figure 4. AutoTutor Lite authoring tools including expectations and reflection prompt.

ATL scores the learner’s responses with the aid of Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). ATL allows the author to choose from several semantic spaces such as College LSA. LSA can produce association values comparing one word to another word, one word to a larger text, or one text to another text in the semantic space. Cutoff scores can be set for these association values, including for weight criteria, minimum association strength, minimum item weight, and the criteria minimum rank, which is the product of weight and strength.

Methods

In a randomized block design, 137 undergraduate participants at Miami University were randomly assigned to one of three conditions, with 46 participants assigned to the AutoTutor Lite group, 48 assigned to the Web-based Euler diagram tutorial group (the same one used in Experiment 2), and 43 assigned to the control group. As in the other two Web-based experiments, participants were run individually or in groups of 2–4 in the laboratory, with each participant seated at a separate workstation.

After providing their informed consent, participants were seated at separate workstations. Each of the three tutorials was titled Conditional Probability Estimation, including the control tutorial. The AutoTutor Lite and static tutorials showed participants how to make conditional probability estimates with the aid of an Euler diagram (see Figure 2). Both of these groups received a worked example. Participants in the AutoTutor Lite group had to explain different relationships among sets by entering their responses into a textbox (see Figure 3). When they completed the experiment participants were thanked and debriefed.

Results

With respect to internal consistency and semantic coherence on conditional probability problems, participants performed universally poorly on problems describing subsets and overlapping sets – including those with necessary conditions, sufficient conditions, and statistical causal relationships. Participants did fairly well on problems describing mutually exclusive sets, identical sets, and independent sets. The approximate percentage of problems on which participants demonstrated semantic coherence (with inconsistent patterns in parentheses) were 10% for overlapping sets (70% inconsistent), 7% for subsets (79% inconsistent), 7% necessary and sufficient conditions (87% inconsistent), 5% causal relationship (75% inconsistent), 75% for mutually exclusive sets (19% inconsistent), 71% identical sets (15% inconsistent), and 57% Independent Sets (15% inconsistent).

Table 3
Mean Inconsistent Responses per Problem by Problem Type (SD in Parentheses)

Problem type	AutoTutor Lite (*N* = 46)	Euler Web (*N* = 48)	Control (*N* = 43)
Identical sets	.11 (.19)	.11 (.20)	.14 (.23)
Mutually exclusive sets	.21 (.27)	.13 (.20)	.19 (.20)
Subsets	.67 (.25)	.78 (.22)	.79 (.18)
Overlapping sets	.53 (.31)	.53 (.24)	.70 (.24)
Independent sets	.24 (.30)	.19 (.27)	.30 (.27)
Causal statistical	.76 (.33)	.78 (.31)	.86 (.27)
Causal necessary or sufficient	.62 (.30)	.64 (.27)	.74 (.21)

Turning to the efficacy of the interventions, Table 3 shows the mean internal inconsistency for each problem type in each condition. For internal inconsistency, a low score corresponds with good performance. As can be seen in Table 3, both AutoTutor Lite and the Web-based Euler diagram intervention significantly reduced inconsistent responses on overlapping sets problems, *F*(2, 134) = 6.09, *p* = .003, η^2 = .098, Tukey's HSD Alpha = .05. AutoTutor Lite also significantly reduced internal inconsistency on problems depicting subsets, *F*(2, 134) = 3.92, *p* = .022, η^2 = .055, Tukey's HSD Alpha = .05. On problems depicting identical sets there were no differences among groups with respect to internal consistency, *F*(2, 134) < 1. The tutorial groups did not differ significantly with respect to internal consistency on mutually exclusive sets problems, *F*(2, 134) = 1.36, *p* = .26. There were no significant differences among groups with respect to inconsistent responses on problems depicting independent sets, *F*(2, 134) = 1.90, *p* = .15. On problems depicting a causal relationship, there was a non-significant trend for internal consistency, *F*(2, 134) = 2.53, *p* = .084. Finally, on problems depicting a statistical causal relationship with neither sufficient nor necessary connections between events (Crisp & Fenney, 2009) there were no differences among groups with respect to internal consistency, *F*(2, 134) = 1.31, *p* = .27.

Table 4
Mean Semantic Coherence per Problem by Problem Type (SD in Parentheses)

Problem type	AutoTutor Lite (*N* = 46)	Euler Web (*N* = 48)	Control (*N* = 43)
Identical sets	.71 (.33)	.71 (.32)	.71 (.33)
Mutually exclusive sets	.68 (.35)	.78 (.28)	.76 (.25)
Subsets	.04 (.01)	.05 (.10)	.07 (.02)
Overlapping sets	.10 (.16)	.10 (.14)	.10 (.16)
Independent sets	.61 (.06)	.58 (.39)	.57 (.37)
Causal statistical	.04 (.14)	.05 (.15)	.01 (.07)
Causal necessary or sufficient	.04 (.08)	.10 (.14)	.05 (.11)

Table 4 shows the mean semantic coherence for each problem type in each condition. For semantic coherence, a high score indicates good performance. AutoTutor Lite did not improve semantic coherence for any problem type. On problems depicting identical sets there were no differences among groups with respect to semantic coherence, *F*(2, 134) < 1.

The tutorial groups did not differ significantly with respect to semantic coherence on mutually exclusive sets problems, *F*(2, 134) = 1.61, *p* = .20. For subsets problems condition did not affect semantic coherence, *F*(2, 134) = 1.42, *p* = .25. As can be seen in Table 3, the tutorials had no effect on semantic coherence on problems depicting overlapping sets, *F*(2, 134) < 1. The interventions had no significant effect on semantic coherence on problems depicting independent sets, *F*(2, 134) < 1. On problems depicting a causal relationship, the Euler diagram Web-based tutorial increased semantic coherence significantly more than the AutoTutor Lite tutorial, *F*(2, 134) = 3.72, *p* = .027, η^2 = .053. On problems depicting a statistical causal relationship with neither sufficient nor necessary connections between events (Crisp & Fenney, 2009) there were no differences among groups with respect to semantic coherence, *F*(2, 134) = 1.19, *p* = .31.

Discussion

As predicted, problems depicting overlapping sets and subsets were more difficult than problems with identical sets, mutually exclusive sets, or independent sets. The AutoTutor Lite tutorial was effective in significantly reducing internal inconsistency on both of these most difficult conditional probability problems. However, neither Web-based tutorial was effective in increasing semantic coherence on any problem type. Participants responded differently to problems depicting a causal relationship, be it a causal relationship with necessary and sufficient conditions, or a statistical relationship such as the relationship between increased taxation and decreased consumption. However, performance on those problems was generally worse than on other kinds of problems depicting overlapping sets – with or without the aid of a Web-based intervention.

General discussion

Assessing semantic coherence and inconsistency by asking participants to estimate *P*(A), *P*(B), *P*(A|B), and *P*(B|A) opens new avenues for researching questions about rationality, judgment, and Bayesian inference (Wolfe & Reyna, 2010b). For example, one can investigate under which conditions people correctly understand the relationship between breast cancer and BRCA (breast cancer genetic) mutations as one of overlapping sets – i.e., some women with breast cancer have BRCA mutations, and others do not, while some women with BRCA mutations have breast cancer and others do not (Reyna et al., 2001). Moreover, this approach can also help uncover cases where the relationship between BRCA mutations and breast cancer is misunderstood as one of subsets, (e.g., all BRCA mutations lead to cancer but some cancers have other causes) identical sets, (e.g., all BRCA mutations lead to breast cancer and all breast cancer is caused by BRCA mutations) or even independent sets (e.g., genetic testing for BRCA mutations is completely uninformative about breast cancer risk). Fisher and Wolfe (2011) have developed and made available software tools in the form of spreadsheet formulae that make it easy for researchers to reduce a very large pattern of responses to just a handful of meaningful patterns. Such tools are available both for joint probability problems (see Wolfe & Reyna, 2010a) and conditional probability estimates (see Fisher & Wolfe, 2011).

AutoTutor Lite is undergoing a period of rapid development and already there have been important advances since we conducted Experiment 3 such as the creation of the domain specific semantic processing portal (Hu, Dai, & Starnes, 2011). In this experiment we encouraged participants to elaborate their understanding of complex nested hierarchical relationships, such as subsets, with the instructions to “keep adding to your answer until you are satisfied that you have done a good job. When I give you a score of greater than .4 on any of the bars you are ready to click on the next arrow button to continue.”. This resulted in significant improvements over the Web-based tutor using Euler diagrams in reducing inconsistency, but in many respects the AutoTutor Lite tutorial was not better than this simpler tutor. To achieve larger and more pervasive gains, it will be necessary to make the interactions between learners and AutoTutor Lite more like those between students and human tutors (Chi et al., 2001). In future research, we plan to develop AutoTutor Lite tutorials that verbally respond to natural language input from learners in ways that are context sensitive and appropriate to the goals of promoting deep learning and comprehension.

The features and capabilities of AutoTutor Lite suggest that it shows great promise in helping people understand the issues associated with conditional probability estimation. Developing Web-based Intelligent Tutoring Systems capable of conversing with learners in their natural language, and responding appropriately to what learners say in different contexts, has the potential to be a transformative technology. However, this technology will only achieve its potential through careful and systematic empirical research. The central thrust of our future work will be developing tutorials created with AutoTutor Lite to help women decide about testing for genetic breast cancer risk, and assessing the efficacy of these tutorials in randomized, controlled experiments.

Collectively, these studies suggest that in making conditional probability estimates, people are highly sensitive to semantic content, particularly the relationship among sets depicted in the problems. Semantic coherence is a very high standard for internal consistency. In the absence of any intervention, semantic coherence was high on problems depicting identical sets, with rates ranging from .81 in Experiment 1, .66 in Experiment 2, and .71 in Experiment 3. Comparable results were found and independent sets, with rates ranging from .71 in Experiment 1, .65 in Experiment 2, and .57 in Experiment 3. Semantic coherence was dramatically lower for problems depicting subsets and overlapping sets in all three experiments with the rates ranging from .03 to .12 across both problem types and all three studies. Problems depicting mutually exclusive sets were consistently in the middle. Patterns of inconsistent responding in the absence of any intervention were comparable. These differences cannot be attributed to random guessing or rounding error since we get comparable results even when a rounding parameter is set very loosely (Fisher & Wolfe, 2011). Rather, both problems of processing and problems of representation account for the difficulty with the more difficult problems with nested sets.

Unlike joint probabilities, there are no first-order fallacies of conditional probability estimation (Wolfe & Reyna, 2010a). However, the paradigm used in these studies permits us to assess internal inconsistency as a minimal standard for rationality. Here our theoretically-motivated interventions focusing on denominator neglect and addressing both processing and representation consistently reduced fallacious internal consistency. On problems depicting overlapping sets, in Experiment 2 the Euler diagrams significantly reduced internal inconsistency from .75 to .62; and in Experiment 3 the AutoTutor Lite tutorial significantly reduced internal inconsistency on overlapping sets problems from .70 to .53. In Experiment 1 the analogy and 2 × 2 table produced a non-significant reduction in internal inconsistency from .52 to .39. A similar pattern of results was found on problems depicting subsets. In Experiment 2 the Euler diagrams significantly reduced internal inconsistency from .83 to .73; and in Experiment 3 the AutoTutor Lite tutorial significantly reduced internal inconsistency on subsets problems from .79 to .67. In Experiment 1 the analogy and 2 ✕ 2 table produced a non-significant reduction in internal inconsistency on subsets problems from .60 to .54. Although far from ideal, these data suggest that performance is somewhat malleable with even brief interventions yielding consistent improvements.

Conditional probability estimation poses serious challenges above and beyond those faced in making joint probability estimates (Wolfe & Reyna, 2010b). Although difficulties in reasoning with nested sets have long ben recognized (Tversky & Kahneman, 1983), we are only beginning to understand the nuances of denominator neglect (Reyna & Brainerd, 2008). Thus it is not surprising that the Web-based interventions had only a limited effect on the very high standard of semantic coherence – particularly on problems depicting subsets and overlapping sets. To illustrate, the rates of semantic coherence with an intervention ranging from a high of only .15 on subsets problems with the 2 ✕ 2 Table and analogy in Experiment 1 to a low of only .04 on subsets problems with ATL in Experiment 3. With respect to the high standard of semantic coherence, performance deviates sharply from the dictates of Bayesian rationality. Turning to larger issues of rationality, the portrait emerging from this research is of a cognitive system that is generally adaptive (the glass is half full) but also prone to inconsistency and other systematic deviations from normative performance (the glass is half empty).

Author note

The project described was supported by Award Number R21CA149796 from the National Cancer Institute, U.S.A. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. We thank the National Cancer Institute for its support.

We also wish to thank Kate Bassolino, Elizabeth Cedillos, Eric Cooke, Elizabeth Crawford, Isabelle Damas Vannucchi, Ben Nelligan, Kelly Papenfus, Jessica Reigrut, Carly Schaeffer, Jon Turpin, Colin Widmer, Triana Williams, and Mandy Withrow for capable assistance with data collection.

References

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*, 241–254.

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, *Educational Researcher, 13,* 4–16.

Brainerd, C. J., & Reyna, V. F. (1990). Inclusion illusions: Fuzzy-trace theory and perceptual salience effects in cognitive development. *Developmental Review, 10*, 365–403.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471–533. doi: 10.1207/s15516709cog2504_1

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 15*, 145–182. doi: 10.1207 /s15516709cog0502_2

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477. doi: 10.1016/0364-0213(94)90016-7

Crisp, A. K., & Feeney, A. (2009). Causal conjunction fallacies: The roles of causal strength and mental resources. *Quarterly Journal of Experimental Psychology, 63*, 2320–2337.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*, 378–395.

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Journal of Personality and Social Psychology. 91*, 797–813.

Fisher, C. R., & Wolfe, C. R. (2011). Assessing semantic coherence in conditional probability estimates. *Behavior Research Methods*, 43, 999–1002. doi: 10.3758/s13428-011-0099-3

Graesser, A., & McNamara, D. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist, 45*, 234–244.

Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 53–65). Cambridge, UK: Cambridge University Press.

Holyoak. K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62*, 135–163.

Hu, X., Dai, J., & Starnes, D. A., (2011, November). *Domain-specific semantic processing portal*. Paper presented at the conference of the Society for Computers in Psychology, Seattle, Washington. Retrieved from https://sites.google.com/site/scipws

Hu, X., Han, L., & Cai, Z. (2008, November). *Semantic decomposition of student's contributions: An implementation of LCC in AutoTutor Lite*. Paper presented at the conference of the Society for Computers in Psychology, Chicago, Illinois. Abstract retrieved from https://docs.google.com/viewer?a=v&pid=sites&srcid =ZGVmYXVsdGRvbWFpbnxzY2lwd3N8Z3g6NmZhMTE2NWI3OGRiMzBlMw&pli=1

Hu, X., & Martindale, T. (2008). Enhance learning with ITS style interactions between learner and content. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). Retreived from http://ntsa .metapress.com/link.asp?id=x44742m02m176563

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus, and Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Reyna, V. F. (2004). How people make decisions that involve risk. A dual-processes approach. *Current Directions in Psychological Science, 13*, 60–66. doi: 10.1111/j.0963-7214.2004.00275.x

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making, 28*, 850–865. doi: 10.1177/0272989X08327066

Reyna, V. F., & Adam, M. B. (2003). Fuzzy-trace theory, risk, communication, and product labeling in sexually transmitted diseases. *Risk Analysis, 23*, 325–342. doi: 10.1111/1539-6924.00332

Reyna, V. F., & Brainerd, C. J. (1993). Fuzzy memory and mathematics in the classroom. In G. M. Davies & R. H. Logie (Eds.): *Memory in everyday life* (pp. 91–119). Amsterdem: North Holland Press.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*, 1–75. doi: 10.1016/1041-6080(95)90031-4

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences, 18*, 89–107. doi: 10.1016/j.lindif.2007.03.011

Reyna, V. F., Lloyd, F., & Whalen, P. (2001). Genetic testing and medical decision making. *Archives of Internal Medicine, 161*, 2406–408.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy-trace theory account. *Journal of Behavioral Decision Making, 8*, 85–108. doi: 10.1002/bdm.3960080203

Wolfe, C. R. (2006). Presidential address: Cognitive technologies for gist processing. *Behavior Research Methods, 38*, 183–189. doi: 10.3758/BF03192767

Wolfe, C. R., Britt, M. A., Petrovic, M., Albrecht, M., & Kopp, K. (2009). The efficacy of a Web-based counterargument tutor. *Behavior Research Methods, 41*, 691–698. doi: 10.3758/BRM.41.3.691

Wolfe, C. R., & Fisher, C. R. (2010, November). *Semantic coherence in conditional probability estimates: 2x2 tables as pedagogic interventions*. Paper presented at the 51st Annual Meeting of the Psychonomic Society, St. Louis, MO.

Wolfe, C. R., & Reyna, V. F. (2010a). Assessing semantic coherence and logical fallacies in joint probability estimates. *Behavior Research Methods, 42*, 366–372. doi: 10.3758/BRM.42.2.373

Wolfe, C. R., & Reyna, V. F. (2010b). Semantic coherence and fallacies in estimating joint probabilities. *Journal of Behavioral Decision Making, 23*, 203–223. doi:10.1002/bdm.650

Appendix A
Bayesian inference and semantic coherence

Bayes' theorem is the standard coherence benchmark by which conditional probability judgments are evaluated (Barbey, & Sloman, 2007; Kahneman, & Tversky, 1973). According to Bayes' theorem, the probability of an event "B" given some other event "A" is defined by the following formula:

$$P(B \mid A) = \frac{P(B) \times P(A \mid B)}{P(B) \times P(A \mid B) + P(\neg B) \times P(A \mid \neg B)}$$

There are five qualitatively different relationships between two sets and their respective conditional probabilities: identical sets, mutually exclusive sets, subsets, overlapping sets and independent sets. As an example of identical sets, consider a situation in which a beaker of fluid is spilled in a chemistry lab. The following estimates are semantically coherence with respect to identical sets: *P*(Water) = .20, *P*(H_2O) = .20, *P*(Water|H_2O) = 1.00 and *P*(H_2O|Water) = 1.00. Although the estimates for *P*(Water) and *P*(H_2O) may vary, it is necessarily the case that *P*(Water) = *P*(H_2O) because water and H_2O refer to the same substance. Both conditional estimates must equal 1.00 because once Water (or H_2O) is assumed to be true, H_2O (or Water) must be true by definition. In more general terms, semantic coherence for identical sets is defined as *P*(A) = *P*(B) and *P*(A|B) = *P*(B|A) = 1.00 (Fisher & Wolfe, 2011).

As an example of mutually exclusive sets, consider a sports commentator's probability judgments for the winner of a football game: *P*(Steelers) = .40, *P*(Browns) = .60, *P*(Steelers|Browns) = 0 and *P(*Browns|Steelers) = 0. This constellation of estimates is semantically coherent with respect to mutually exclusive sets. By definition, mutually exclusive events cannot occur together and thus implies both conditional probabilities must equal 0 and the sum of the component probabilities cannot exceed 1.00. In more general terms, semantic coherence for mutually exclusive sets is defined as: *P*(A) + *P*(B) ≤ 1.00, *P*(A) > 0, *P*(B) > 0, and *P*(B|A) = *P*(A|B) = 0 (Fisher & Wolfe, 2011). Note that *P*(A) and *P*(B) are constrained to be greater than zero because it's impossible to interpret a constellation of estimates when each estimate equals zero. Recall the example of subsets involving mathematics and Anna's class schedule. In general terms, semantic coherence for subsets (such that A is a subset of B) is defined as 0 < *P*(A) < *P*(B), *P*(B|A) = 1.00, and *P*(A|B) = *P*(A)/*P*(B) (Fisher & Wolfe, 2011).

Independent sets are a special case of overlapping sets in which the occurrence of one event provides no additional information regarding the other event. In other words, the relationship between both events is random. As an example of independent sets, consider the relationship between a coin landing heads up and whether it will snow on a given day. In order to be semantically coherent with respect to independent sets, it must be the case that *P*(Snow) = *P*(Snow|Heads) and *P*(Heads) = *P*(Heads|Snow). In more general terms, semantic coherence for independent sets is defined as *P*(A) = *P*(A|B) > 0 and *P*(B) = *P*(B|A) > 0 (Fisher & Wolfe, 2011).

As an example of overlapping sets, consider the relationship between Cara wearing a jacket and the month of the year. The following estimates are semantically coherent with respect to overlapping sets: *P*(Jacket) = .25, *P*(November) = .20, *P*(Jacket|November) = .75 and *P*(November|Jacket) = .60. In this particular example, the probability of wearing a jacket increased from .25 to .75 when the month was assumed to be November (perhaps because of the lower late autumn temperatures). With overlapping sets such as these, *P*(A) and *P*(B) contain some information about one another, but that information is incomplete or imperfect. Thus, knowing that it is November increases the probability that Cara is wearing a jacket, and knowing that she is wearing a jacket increases the probability that the month in question is November, but neither conditional probability estimate rises to the level of certainty. In general terms, semantic coherence for overlapping sets occurs when *P*(B|A) can be inferred from the other three estimates based on Bayes' theorem, but does not match any of the previously mentioned patterns (Fisher & Wolfe, 2011).

Appendix B
An estimation problem representing subsets

Felix is a finicky pet. He likes to slink around rubbing up against the furniture, but his favorite thing is finding a warm place in the sun to curl up and fall asleep.

Please rate the probability of each of these statements about the story above using a rating scale from 0% (impossible) to 100% (completely certain).

What is the probability that Felix is a cat?

What is the probability that Felix is mammal?

Now suppose for a minute that Felix is a cat. If you make that assumption, what is the probability that Felix is mammal?

Now let’s turn around and suppose for a minute that Felix is mammal. If you start with that assumption, what is the probability that Felix is a cat?