Examining Experts' Recommendations of Representational Systems for Problem Solving

Aaron Stockdill*[†], Gem Stapleton* Daniel Raggi*, Mateja Jamnik*, Grecia Garcia Garcia[†], Peter C.-H. Cheng[†]
*University of Cambridge, Cambridge, UK
{aaron.stockdill, ges55, daniel.raggi, mateja.jamnik}@cl.cam.ac.uk
[†]University of Sussex, Brighton, UK
{a.a.stockdill, g.garcia-garcia, p.c.h.cheng}@sussex.ac.uk

***Abstract*—Pólya and others recognised that an appropriate representation of a problem is key for enabling us to solve it. But choosing the right representation is a problem that novice problem solvers find difficult, so must turn to experts for guidance. In this paper, we present a study that examines how human experts recommend representations. We asked high school mathematics teachers to order representational systems based on their suitability generally, and with respect to a student profile. We found the teachers updated their recommendations based on the problem and student profile, but were inconsistent with each other. This inconsistency highlights a need for more training and support in representational system selection.**

***Index Terms*—representations, experts, problem solving**

I. Introduction

Problems appear everywhere, from everyday activities to advanced mathematics. One general problem solving principle is to formulate and transform the problem into a new representational system [1], [2], potentially providing new inferences 'for free' [3]. But reformulating problems is a challenge, particularly for non-experts [4]. Experts need to guide them, but are experts consistent in their guidance? Does this guidance reflect the problem, and the person solving it?

In this paper, we focus on mathematics education, specifically high school level probability problems, for which diverse representational systems (RSs) are a core part of the curriculum [5], [6]. There is evidence that proper use of RSs in mathematical problem solving can improve learning [7], [8]. This area is also of particular interest since there is a ready supply of experts – specifically mathematics teachers – from whom we can gain insight into which RSs are suitable for which problems and for which students. We have collected data from mathematics teachers, and so compiled and analysed a dataset of recommendations of RSs regarding which should be used for certain probability problems for particular student profiles.

We found that teachers consider both the problem, and the student. However, the resulting recommendations are inconsistent: for the same problem and student profile, the teachers' recommendations can, in some cases, vary greatly.

This paper is organised as follows. Section II presents our hypotheses, while Section III details the experimental design. Section IV provides details of our participants, and how the experiment was conducted. We provide a quantitative analysis of the participants' responses in Section V, and a qualitative analysis in Section VI. We discuss some limitations of our experiment in Section VII, before concluding in Section VIII.

II. Hypotheses

We aim to determine whether experts, specifically secondary school mathematics teachers, produce similar RS recommendations. These recommendations should consider both the problem being solved and the cognitive profile of the person – in this case, a student – doing the solving. We ask them also to consider their recommendations in the general case, with no cognitive profile in mind. That is, they should consider what the representational system is capable of expressing, not the elegance with which it expresses it; this is the 'informational suitability (IS)' of an RS.

We break down our high level goals into three hypotheses.

H1. From the teachers' individual responses it is possible to produce an overall ranking of RSs for each problem and cognitive context.

That is, their responses should be at least partially consistent with each other – they are all starting from the same problem and cognitive situation (see the next two hypotheses), and they are working within the same curriculum with a related cohort of students mostly educated in that same curriculum. Thus we would expect that the teachers' responses would be sufficiently similar that we can extract some RS ranking.

We expect that the teachers' recommendations would change based on the situation, too. Thus, we hypothesise that:

H2. The teachers' aggregate RS recommendations change based on the problem that they are considering.

Finally, the recommendation should also vary based on the cognitive abilities of the student they are helping:

H3. The teachers' aggregate RS recommendations change based on the cognitive context (with IS only [no person in mind], a low-ability student, or a high-ability student) that they are considering.

Aaron Stockdill was supported by the Hamilton Cambridge International PhD Scholarship. This work was supported by the EPSRC grants EP/R030650/1, EP/T019603/1, EP/R030642/1, and EP/T019034/1.

978-1-6654-4214-5/22/$31.00 ©2022 IEEE

III. Design

We designed the experiment in the context of New Zealand mathematics students, aged 15–18. We chose the domain of probability as there are a wide variety of potential RSs, and the problems cover a range of difficulties.

A. Representational systems

We selected five diverse RSs for this study: Area Diagrams, Bayesian Algebra, Contingency Tables, Euler Diagrams, and Probability Trees. Each is obviously distinct – there is no confusion to which RS a particular representation belongs.

Area Diagrams use a unit square partitioned into regions with horizontal and vertical lines, where the area of a region with edges labelled by events X and Y represents the probability of $X \cap Y$; areas of disjoint regions for events A and B sum to the probability of $A \cup B$.

Bayesian Algebra is standard algebraic notation, augmented with two probability functions $\Pr(\cdot)$ and $\Pr(\cdot \mid \cdot)$, conditional probability laws, and Bayes' Theorem.

Contingency Tables use a grid of cells where the sum of all the values in the table must be 1. The value in a cell in row X and column Y contains the probability of $X \cap Y$.

Euler Diagrams represent events as contours (circles) and the overlapping regions represent their conjunction. This RS cannot represent the magnitude of most probabilities, so is unsuitable for any of our problems. That is, we considered *non-proportional* Euler Diagrams.

Probability Trees represent events as nodes in a rooted tree, and the (directed) edges are labelled with conditional probabilities. Multiplying along branches computes conjunction, while adding between branches computes disjunction. Edge length and order are not meaningful.

Examples of each are in Appendix A. Area Diagrams were included because this RS is *not* commonly taught in New Zealand; we wished to see what effect an unfamiliar RS has on the teachers' responses.

B. Cognitive contexts

To evaluate the RSs, the teachers need to consider the *cognitive context* that the RS will be used in. For this study, we use three contexts: IS (i.e., without any student in mind), a low-ability student context, and a high-ability student context. We expect the teachers to adjust their responses based on the cognitive contexts, addressing H3.

For the IS context, teachers were not given any persona to consider when scoring the RSs. For the contexts involving students, we provided two *personas*: Student A, a low-ability 15-year-old, and Student B, a high-ability 17-year-old. They were chosen to be sufficiently distinct for teachers to update their recommendation, if they choose. The precise wording of the personas is in Appendix B.

C. Problems

We selected five typical probability problems to address H2:

1) 1% of the population has a disease. A test is reliable 98% if you have the disease and 97% if you do not have the disease. Assuming the test comes out positive, what is the probability of having the disease?
2) One quarter of all animals are birds. Two thirds of all birds can fly. Half of all flying animals are birds. Birds have feathers. If X is an animal, what is the probability that it's not a bird and it cannot fly?
3) Let A, B be events, and $\Pr(A) = 0.2$. We also have that $\Pr(B \mid A) = 0.75$ and $\Pr(A \mid B) = 0.5$. Calculate $\Pr(\bar{A} \cap \bar{B})$.
4) There are two lightbulb manufacturers in town. One of them is known to produce defective lightbulbs 30% of the time, whereas for the other one the percentage is 80%. You do not know which one is which. You pick one to buy a lightbulb from, and it turns out to be defective. The same manufacturer gives you a replacement. What is the probability that this one is also defective?
5) Let S, T, U be events. We have that $\Pr(S) = 0.5$. We also have that $\Pr(T \mid S) = \Pr(U \mid S) = 0.1$, and that $\Pr(T \mid \bar{S}) = \Pr(U \mid \bar{S}) = 0.2$. We assume that T and U are independent with respect to S, that is $\Pr(T \cap U \mid S) = \Pr(T \mid S) \times \Pr(U \mid S)$. Calculate $\Pr(U \mid T)$.

The first problem about medical testing was used as practice, and always presented first; the rest were counterbalanced. The responses for the first problem were discarded; the teachers were *not* made aware of this.

Problems 2 and 3 are 'equivalent' – these contain the same information and goal. Similarly, problems 4 and 5 are 'equivalent'. The information content of the problem, and the solution paths in each RS, would be identical for each pair. The teachers were not informed of this until debriefing.

We categorise the final four as 'easy' or 'hard', and 'verbal' or 'formulaic'. We use the abbreviations E, H, V, and F, respectively, so problems 2 through 5 are EV, EF, HV, and HF.

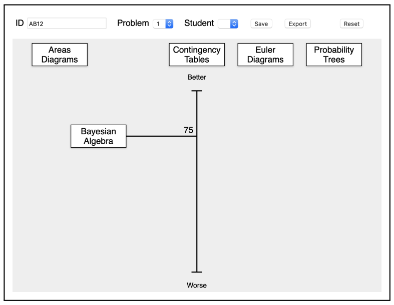
D. Training

Because the teachers may not be familiar with our chosen RSs – or understand the RSs differently – we provided training on each. This consisted of going over a one-page PDF document with the teachers; the RSs were introduced in a counterbalanced order. The training documents are in Appendix D.

The training document for each RS contained a brief description of the RS, along with four examples. The training resources were kept uniform in what they described, and their length: each described how the RS encoded the underlying probability concepts such as events, 'and', and 'or', as well as general syntactic rules. This ensured that no particular RS was promoted as 'better' than the others. The examples were a representation belonging to that RS, and a short textual description of the representation. The teachers were asked to explain how the text described the representation, and then answer some brief questions about extracting information from the representation. The correct answers were then given.

E. Tasks

The experiment was divided into two phases: in phase one, the teachers assessed the IS of the RSs for each problem; in phase two, they assessed the suitability of the RSs for each problem for a specific student persona. For each problem and cognitive context, the teachers were asked to arrange the RSs on the online response form shown in Fig. 1. The teachers entered their identification code, the problem, and cognitive context, then dragged the labels of the RSs onto the central scale, 0 to 100. The labels all begin in the top row, can be dragged anywhere, and may overlap. The boxes have a horizontal line that connects to the central scale; the horizontal position has no meaning, which the teachers were told. When they were happy with their response, they clicked the 'Save' button, then 'Reset' to return the labels to the top row.



ID AB12 Problem 1 Student

Fig. 1. A screenshot of the response form teachers used to score RSs.

For phase one – RSs suitability when considering *only* the problem, not who might solve it – we presented the teachers with the problem statement, requested that they read the problem (and to *not* solve it), asked if they had any questions, then asked them how *informationally suitable* each RS would be. They then arranged the RS labels on the web form.

After all problems had been presented to the teachers, we entered phase two: we asked the teachers to consider not just the IS of an RS, but also how appropriate they would be for students via the personas. The teachers then saw the same problems in the same order, but for each they arranged them based on how suitable the RS is for each persona. Once their responses had been saved for Student A, they immediately did the same problem for Student B. They completed all problems.

F. *Interview*

Following the two experimental phases, we conducted a semi-structured interview guided by four questions:

1) Did you find this task difficult or easy, and how confident are you in your answers?
2) How familiar were you with each representational system before we started, on a scale from 1 to 10?
3) Which representational systems do you use while teaching, and which are your ‘go-to’?
4) When answering our questions, what were the key factors in making your decision?

The teachers were also given a short survey to collect demographic information: education, years of teaching experience, recently taught courses, and the school at which they work.

IV. Participants and procedure

Due to the COVID-19 pandemic, the experiment was conducted via Zoom videoconference. The participants were high school mathematics teachers in New Zealand. We advertised by directly reaching out to the heads of faculty of high schools in Canterbury. We recruited 10 teachers (3 male, 7 female) from five schools; nine teachers returned usable quantitative data – one set of data was corrupted. The participants' teaching experience ranged from two-and-a-half to sixteen years; all have been mathematics teachers for all of their teaching career. All have a bachelors degree, and a postgraduate diploma in Teaching; the degree major varied. One participant has a PhD in Statistics, while one has an MSc in Computer Science. One teacher was studying for a Masters of Specialist Teaching. All had taught courses that included probability within two years.

All participants were rewarded with an NZ$20 gift voucher.

This study received ethics approval from the University of Cambridge Department of Computer Science and Technology.

A. *Introducing the experiment*

To open, we motivated this experiment: to understand how teachers consider solving problems, both generally and for students. We explained terms like ‘informational suitability’.

B. *Training*

The teachers were then given training as stated in Section III-D. They consistently made three remarks:

- They were unfamiliar with Area Diagrams (but one had seen eikosograms before, which are related [9]).
- They knew Contingency Tables as ‘two-way tables’.
- They knew Euler Diagrams as ‘Venn diagrams’; this name is used by the NCEA standards documents for Euler diagrams.

This training period lasted about 30 minutes.

C. *Representational systems without cognitive context*

We presented each problem, and asked the teacher to read it but *not to solve it*. The teacher was asked if they understood the problem; all said yes. We then asked them to place the RSs labels in the web form based on their IS.

D. *Representational systems with cognitive context*

After completing the evaluation task for each problem only for IS, we presented the teachers with a PDF containing the personas of the two students. They were asked to read the personas, and we asked if they had any questions. One teacher asked whether either student would be allowed a calculator when solving these problems, and we confirmed with yes. Another queried how strictly the low-ability student would not use knowledge from other areas of mathematics, and we confirmed that they had basic knowledge, but they would not use skills beyond basic arithmetic without prompting. Responses were recorded using the same interface.

E. *Debrief and questions*

To end, we asked the teachers our four follow-up questions, and any questions that arose during the conversation. We also invited them to complete a demographics survey. Finally, we debriefed the teachers on some details of the experiment, notably that the questions came in ‘pairs’ of the same problem – no participant acknowledged noticing this.

V. Quantitative analysis

To make sense of the teachers' responses, we break down the data by problem and cognitive context. For each $\langle \text{problem}, \text{cognitive context} \rangle$ pair, we consider all of the teachers' responses. In this section, we explore two examples: the lightbulbs-equivalent problem with a high-ability student persona ($\langle \text{5/HF}, \text{high-ability} \rangle$), which shows clear groupings; and the birds problem when considered without any persona ($\langle \text{2/HV}, \text{no persona} \rangle$), which does not show clear groupings. All statistical test results are included in Appendix C.

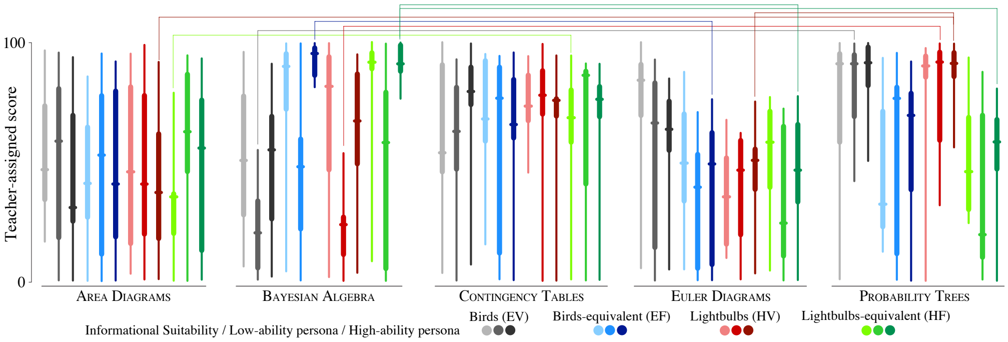


Fig. 2. The teachers' assigned scores for each RS, as box plots, colour-coded for each problem and cognitive context (legend at the bottom). Connecting lines (top) indicate between which RSs we find significant differences in rankings, when considering each $\langle \text{problem}, \text{cognitive context} \rangle$ pair.

A. *Lightbulbs-equivalent problem for high-ability persona*

The lightbulbs-equivalent problem is number 5/HF. We asked the teachers to consider each RS, and evaluate them based on their suitability for the high-ability student persona.

We plot the data in Fig. 2, focusing on the dark green box plots (right-most within each RS). We note two things: first, the Bayesian Algebra distribution is visually much tighter than the others – indicating more agreement between the teachers – and also much higher – the teachers believed this representational system to be generally more suitable for this problem, and this student persona. This gives us what appears to be a 'winner': for this problem and this student persona, the teachers would consistently recommend Bayesian Algebra. This seems to support H1, in that the teachers have consistently identified a representation to recommend. It then appears the teachers would suggest (after Bayesian Algebra) to use Contingency Tables, followed by Probability Trees, then Euler Diagrams. Second we see, in particular, Area Diagrams scores are spread out – the teachers do *not* agree with each other. This was the teachers' least familiar RS.

Due to few scores, and their significant non-normal distribution, we converted the teachers' responses (integers from 0 to 100) to ranks for each RS, preserving ties. Using these ranks, we performed a Friedman test between the mean ranking of each RS. For the problem and cognitive context described above ($\langle$ 5/HF, high-ability $\rangle$), we find there is a significant difference between the RS rankings ($Q = 20.50$, $p = 0.0004 < 0.05$). Post-hoc Wilcoxon signed-rank tests between every pair of RSs reveal two significant differences after Bonferroni correction: between Bayesian Algebra and Euler Diagrams ($W = 0$, $p = 0.004 < 0.005$) and between Bayesian Algebra and Probability Trees ($W = 0$, $p = 0.004 < 0.005$). Fig. 2 marks both with a connecting line. Thus we have evidence that the teachers would recommend Bayesian Algebra over Euler Diagrams and Probability Trees. While not a comprehensive ranking, we have extracted a ranking from the teachers' responses: tentative evidence for H1.

B. *Birds problem without any persona*

The birds problem is number 2/EV. As before, we plot the teachers' responses in Fig. 2, now focusing on the light grey box plots (left-most for each RS). This time, any patterns are less clear. The RSs' scores are spread across the scale, with none clearly being better or worse than the others. We might state that Euler Diagrams has higher scores, but this is not conclusive. This time, there is no apparent 'better' RS.

After the transformation from scores to ranks, we performed a Friedman test to determine if there is a significant difference between the RS rankings. No significant difference was found ($Q = 7.75$, $p = 0.101$), matching our visual intuition. Thus in this case, we have no evidence supporting H1, that the teachers were able to agree on the IS of each RS for the birds problem. We made the assumption that the teachers are working from a similar situation, knowledge, and experience; there may be individual differences unaccounted for.

C. *Other combinations*

These two $\langle$ problem, cognitive context $\rangle$ pairs are representative of the results from all twelve pairs. Fig. 2 shows connecting lines between all pairs of RSs between which a post-hoc Wilcoxon signed-rank test indicate a significant difference in the teachers' rankings ($p < 0.005$). In three further cases – $\langle$ 2/EV, high-ability persona $\rangle$, $\langle$ 3/EF, low-ability persona $\rangle$, and $\langle$ 4/HV, IS $\rangle$ – the Friedman tests find a significant difference between the teachers' rankings, but post-hoc tests failed to determine between which RSs the difference occurred. Full results are in Appendix C.

Based on summarised results, in one quarter of cases there is no evidence of a difference between each RS. In another quarter, we found evidence that there might be a difference in rankings between the RSs, but post-hoc tests were not sensitive enough to determine the difference. But we note that there is no consistency in the problems or cognitive contexts in which we determine significant differences: both the problem and the cognitive context seem to be influencing the result. There is some consistency in three cases: in the birds variants, problems 2/EV and 3/EF, for IS (in that there *were no significant preferences*); in the 'contextual' problems for low-ability learners (for favouring Probability Trees over Bayesian Algebra); and in the 'equivalent' problems for high ability learners (for favouring Bayesian Algebra over Euler Diagrams). However, three cases of significant differences matching out of the twelve cases (and each case has ten possible pairings) suggests that each case is being treated differently by the teachers. Thus, for H2 and H3 we have evidence to suggest that the teachers were considering both the problem and cognitive context in their evaluation of each RS.

VI. Qualitative analysis

In the debriefing interview we asked the four questions from Section III-F. We asked the teachers how familiar they had been with each RS prior to the training we provided. The teachers were universally confident with Probability Trees, Contingency Tables, and Euler Diagrams; two thirds were comfortable with Bayesian Algebra, but the rest had only memories of having learned it before; none had seen Area Diagrams before the study, but one third still felt they would confidently be able to use the RS even before our training.

More than half of teachers initially answered that their responses were based on 'gut instinct' rather than external factors; further discussion revealed influences from the curriculum towards Probability Trees and Contingency Tables.

The teachers also commented on a lack of training, in particular with respect to what one referred to as 'rich task problem solving': using contexts, representations, and discussions to improve mathematics learning [10]. This highlights a need for teacher training and resources that allow for using more diverse RSs, which could improve learning opportunities for students.

We asked the teachers if they had any 'go-to' RSs for probability. All responded either Probability Trees or Contingency Tables, half pointing out that these are encouraged by the assessment standards. We notice these RSs were favoured by our participants. Many said they were reluctant to use Euler Diagrams because they felt they were 'too hard' for students.

Based on these responses, we suspect that personal preference and curriculum were factors in our participants' responses. We cannot directly untangle the link between curriculum and preference: the teachers all work within the New Zealand mathematics curriculum, so are most familiar with (and have most experience with) the mandated RSs. A similar experiment on a different cohort of teachers from multiple curricula could identify if this influences the teachers' responses.

The 'equivalent' problems may also have influenced our participants: we had 'contextual' problems (birds, 2/EV, and lightbulbs, 3/HV) and 'context-less' problems (3/EF and 5/HF) using letters as variables and a probability function. These 'context-less' variants might have encouraged Bayesian Algebra, which also uses letters as variables and a probability function. Indeed, we see this in Fig. 2: every situation where Bayesian Algebra is preferred is 'context-less'. In future, we suggest using 'equivalent' problems that retain context to avoid the Bayesian Algebra bias, but to use a *different* context.

We also cannot discount the possibility that this RS recommendation task was difficult, even for experienced teachers well-versed in the subject matter. As part of the debriefing interview, we asked the teachers to self-assess how difficult they found the evaluation task. Responses were split to extremes: just under half responded that it was difficult, with the rest responding that it was easy; there was no obvious relationship between this response and years of experience. This binary split on a self-assessment question suggests more work is needed to determine what makes this task simple or difficult; or, we need to find what assumptions some of the teachers might be making that caused the task to be easier or more difficult.

Overall, we find that the teachers are only partially able to produce a consistent recommendation of RSs. There are some trends, but our participants did not consistently agree with each other – against our initial hypotheses. The inconsistency, and the teachers mentioning a lack of training in re-representation, indicates that while teachers have an interest in learning about teaching with multiple representations, this need is not being met; in turn, this means that students may not be exposed to the diversity of representations they could be.

VII. Limitations and threats to validity

Number of participants While nine participants provides useful information, the power of the study is limited. We provide interesting preliminary data, but both researchers and practitioners would benefit from a larger study.

Population homogeneity We recruited our participants from a limited set of schools in a geographically restricted area. Participants were self-selected, likely knew each other professionally, and shared an interest in representations in education; this reduces the diversity of our participants. This study was restricted to probability problems.

Mismatch of problems The initial problems and RSs selection was based on the English mathematics curriculum, but due to the COVID-19 pandemic the study was adapted for New Zealand. While some changes could be made quickly – e.g., translating the GCSE/A-levels personas to the NZQA framework – we could not make others because we did not identify them ahead of time. One difference is the order and age at which different RSs are introduced by each curriculum. For example, in England, Euler Diagrams[1] are introduced at Key Stages 3 and 4 [6], for students aged 11 to 14; in New Zealand, Euler Diagrams are introduced at NCEA Level 3 [5], for students aged 17 to 18. This likely caused the teachers to associate Euler Diagrams with difficult materials.

VIII. Conclusions

This study has provided valuable information about how teachers evaluate RSs. We have found, contrary to H1, they are not as consistent as we expect: teachers often fail to agree with each other on the suitability of a particular RS. But they *are* reacting to the situation in which they are making a recommendation: the teachers' responses support H2 (that the problem is a factor in their evaluation) and H3 (that the cognitive context is a factor in their evaluation). Further studies are needed to determine the influence of these factors – and potentially others – on the final recommendation. But we have demonstrated a need, and desire, for more training on diverse representation use. This may allow teachers and students to unlock their problem solving potential.

[1] Both curricula refer to Euler Diagrams as Venn diagrams.

A version of this paper without appendices has been published at VL/HCC 2022.

References

[1] G. Pólya, *How to Solve It: A New Aspect of Mathematical Method*, ser. Princeton Science Library. Princeton University Press, 1957.

[2] R. Cox, "Representation construction, externalised cognition and individual differences," *Learning and Instruction*, vol. 9, no. 4, pp. 343–363, 1999.

[3] A. Shimojima, "The graphic-linguistic distinction," in *Thinking with Diagrams*, A. F. Blackwell, Ed. Springer, 2001, pp. 5–27.

[4] Y. Uesaka, E. Manalo, and S. Ichikawa, "The effects of perception of efficacy and diagram construction skills on students' spontaneous use of diagrams when solving math word problems," in *Diagrammatic Representation and Inference, Diagrams 2010*, ser. Lecture Notes in Computer Science, A. K. Goel, M. Jamnik, and N. H. Narayanan, Eds. Springer, 2010, pp. 197–211.

[5] New Zealand Qualification Authority, "Achievement Standard AS91585: Apply probability concepts in solving problems," https://www.nzqa.govt.nz/nqfdocs/ncea-resource/achievements/2019/as91585.pdf, November 2016 (Last retrieved 6 Dec 2021).

[6] UK Department for Education, "National curriculum in England: mathematics programmes of study," https://www.gov.uk/government/publications/national-curriculum-in-england-mathematics-programmes-of-study/national-curriculum-in-england-mathematics-programmes-of-study, September 2021 (Last retrieved 6 Dec 2021).

[7] G. A. Goldin, "Representational systems, learning, and problem solving in mathematics," *The Journal of Mathematical Behavior*, vol. 17, no. 2, pp. 137–165, 1998.

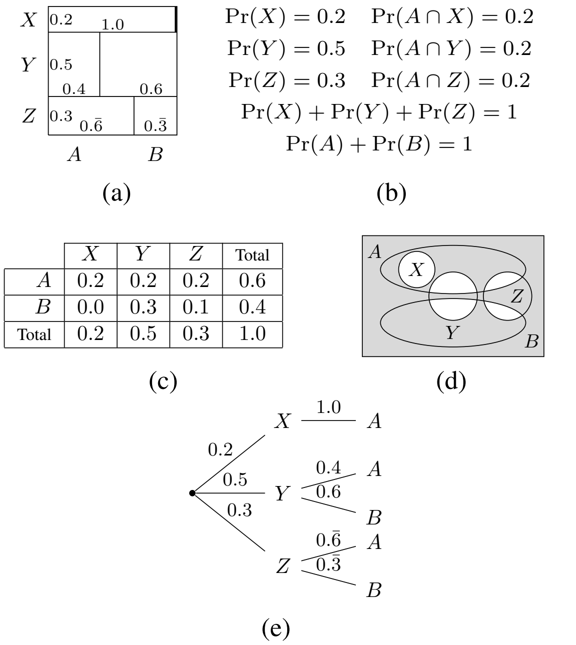
[8] S. Ainsworth, *The Educational Value of Multiple-representations when Learning Complex Scientific Concepts*. Springer, 2008, vol. 3, ch. 9, pp. 191–208.

[9] R. W. Oldford and W. H. Cherry, "Picturing probability: the poverty of venn diagrams, the richness of eikosograms," Retrieved from http://www.stats.uwaterloo.ca/~rwoldfor/papers/venn/eikosograms/paperpdf.pdf, 2006.

[10] J. Piggott, "Rich tasks and contexts," Online article (accessed 6 Dec. 2020). https://nrich.maths.org/5662, 2008.

Appendix A
Examples of Representational Systems

Figure 3 gives examples of all five RSs we considered in this experiment.



(a)

$$\Pr(X) = 0.2 \quad \Pr(A \cap X) = 0.2$$
$$\Pr(Y) = 0.5 \quad \Pr(A \cap Y) = 0.2$$
$$\Pr(Z) = 0.3 \quad \Pr(A \cap Z) = 0.2$$
$$\Pr(X) + \Pr(Y) + \Pr(Z) = 1$$
$$\Pr(A) + \Pr(B) = 1$$

(b)

	X	Y	Z	Total
A	0.2	0.2	0.2	0.6
B	0.0	0.3	0.1	0.4
Total	0.2	0.5	0.3	1.0

(c)

(d)

(e)

Fig. 3. (a) A probability tree, (b) a contingency table, (c) an area diagram, (d) an Euler diagram, and (e) five Bayesian algebra equations. Representations are (near) equivalent in describing events A, B, X, Y, and Z.

Appendix B
Student personas

Student A is 15 years old, and in Year 11. They are able to add and subtract well but are less confident with multiplication and division. They can perform one or two steps independently if they have seen them done before, but problems that require more steps to solve will leave them unable to start. They cannot use knowledge from other areas of mathematics; they only use skills they have learned in probability to solve probability problems.

Student B is 17 years old, and in Year 13. They are confident with addition, subtraction, multiplication, and division. They can solve problems that require many steps and are willing to try steps they have not explicitly seen demonstrated before. The student is able to combine knowledge from across mathematics to solve their current problem.

Appendix C
Summary tables of teacher's responses statistics

The following tables summarise the analysis of the responses from the teachers who participated in the evaluation. These tables form part of the analysis in Section V.

In the Friedman test tables, the final column contains an asterisk if the p-value is below 0.05, indicating we should pursue post-hoc tests. If a context exhibits significant differences, Wilcoxon post-hoc tests were conducted. In the Wilcoxon test tables, the final column contains an asterisk if the p-value is below 0.005, which is the significance threshold after Bonferroni correction.

A. *Birds problem*

Friedman tests

Context	Friedman Q	p	
No persona	5.03	0.284	
Low ability	11.29	0.024	*
High ability	11.01	0.027	*

Wilcoxon tests, low ability

Representational Systems		Wilcoxon W	p	
Areas	Bayes	10.5	0.164	
Areas	Contingency	15.5	0.719	
Areas	Euler	14.0	0.570	
Areas	Trees	5.0	0.067	
Bayes	Contingency	9.0	0.129	
Bayes	Euler	8.5	0.129	
Bayes	Trees	0.0	0.004	*
Contingency	Euler	17.5	0.944	
Contingency	Trees	7.0	0.121	
Euler	Trees	10.0	0.164	

Wilcoxon tests, high ability

Representational Systems		Wilcoxon W	p
Areas	Bayes	21.5	0.910
Areas	Contingency	5.0	0.039
Areas	Euler	16.0	0.496
Areas	Trees	2.0	0.012
Bayes	Contingency	8.5	0.129
Bayes	Euler	14.5	0.618
Bayes	Trees	6.5	0.074
Contingency	Euler	13.5	0.301
Contingency	Trees	9.0	0.389
Euler	Trees	6.0	0.055

B. *Birds-equivalent problem*

Friedman tests

Context	Friedman Q	p	
No persona	7.75	0.101	
Low ability	9.98	0.041	*
High ability	18.18	0.001	*

Wilcoxon tests, low ability

Representational Systems		Wilcoxon W	p
Areas	Bayes	10.5	0.285
Areas	Contingency	11.0	0.203
Areas	Euler	4.5	0.102
Areas	Trees	15.5	0.722
Bayes	Contingency	8.0	0.098
Bayes	Euler	13.5	0.518
Bayes	Trees	11.5	0.359
Contingency	Euler	3.5	0.020
Contingency	Trees	6.5	0.105
Euler	Trees	7.0	0.119

Wilcoxon tests, high ability

Representational Systems		Wilcoxon W	p	
Areas	Bayes	2.0	0.012	
Areas	Contingency	2.0	0.012	
Areas	Euler	15.0	0.669	
Areas	Trees	10.5	0.286	
Bayes	Contingency	10.5	0.164	
Bayes	Euler	0.0	0.004	*
Bayes	Trees	2.5	0.012	
Contingency	Euler	0.0	0.011	
Contingency	Trees	7.5	0.136	
Euler	Trees	10.0	0.164	

C. *Lightbulbs problem*

Friedman tests

Context	Friedman Q	p	
No persona	12.02	0.017	*
Low ability	22.01	0.000	*
High ability	20.83	0.000	*

Wilcoxon tests, no persona

Representational Systems		Wilcoxon W	p
Areas	Bayes	13.0	0.478
Areas	Contingency	9.0	0.203
Areas	Euler	7.0	0.120
Areas	Trees	3.0	0.035
Bayes	Contingency	15.0	0.670
Bayes	Euler	6.5	0.055
Bayes	Trees	9.5	0.231
Contingency	Euler	2.5	0.012
Contingency	Trees	5.5	0.143
Euler	Trees	4.5	0.039

Wilcoxon tests, low ability

Representational Systems		Wilcoxon W	p	
Areas	Bayes	3.0	0.031	
Areas	Contingency	4.0	0.048	
Areas	Euler	22.0	1.00	
Areas	Trees	4.5	0.039	
Bayes	Contingency	0.0	0.011	
Bayes	Euler	3.0	0.020	
Bayes	Trees	0.0	0.004	*
Contingency	Euler	7.0	0.074	
Contingency	Trees	9.5	0.222	
Euler	Trees	2.0	0.012	

Wilcoxon tests, high ability

Representational Systems		Wilcoxon W	p	
Areas	Bayes	6.5	0.074	
Areas	Contingency	3.0	0.034	
Areas	Euler	21.5	1.00	
Areas	Trees	0.0	0.004	*
Bayes	Contingency	21.5	1.00	
Bayes	Euler	5.0	0.039	
Bayes	Trees	2.5	0.020	
Contingency	Euler	7.0	0.074	
Contingency	Trees	2.5	0.028	
Euler	Trees	0.0	0.004	*

D. *Lightbulbs-equivalent problem*

Friedman tests

Context	Friedman Q	p	
No persona	16.02	0.003	*
Low ability	7.43	0.115	
High ability	20.50	0.000	*

Wilcoxon tests, no persona

Representational Systems		Wilcoxon W	p	
Areas	Bayes	1.0	0.008	
Areas	Contingency	0.0	0.004	*
Areas	Euler	12.5	0.301	
Areas	Trees	5.0	0.039	
Bayes	Contingency	9.0	0.129	
Bayes	Euler	4.0	0.027	
Bayes	Trees	8.0	0.098	
Contingency	Euler	8.5	0.098	
Contingency	Trees	11.0	0.319	
Euler	Trees	17.0	0.570	

Wilcoxon tests, high ability

Representational Systems		Wilcoxon W	p	
Areas	Bayes	1.5	0.012	
Areas	Contingency	4.5	0.055	
Areas	Euler	16.0	0.774	
Areas	Trees	17.5	0.943	
Bayes	Contingency	3.5	0.034	
Bayes	Euler	0.0	0.004	*
Bayes	Trees	0.0	0.004	*
Contingency	Euler	3.0	0.034	
Contingency	Trees	2.5	0.028	
Euler	Trees	14.5	0.608	

Appendix D
Representational system training resources

The following pages are direct copies of the training material given to participants during our experiment.

The documents are included verbatim from the study; errors present here were also present in versions shown to participants. In particular, the Area Diagrams information sheet incorrectly states in the second example that ‘three of the five even numbers are prime' – three of the five *odd* numbers are prime, not even. A few participants did pick up on this, and correctly inferred the mistake. Many did not pick up on our error: we believe they implicitly understood the intended meaning.

The example also required the participants to have general knowledge about integers and playing cards; they all had no problem understanding the examples as given.

Representation – Area diagrams

Summary

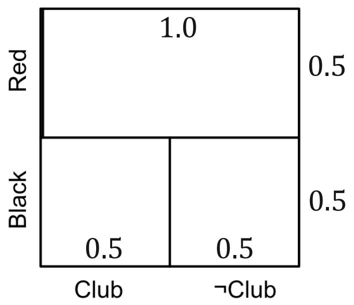
An area diagram is a unit square representing all possible outcomes, with labels for events, their split length representing the probability of each event. Labels might use “not” (¬)

The area enclosed by lines represents the probability X *and* Y together, where X and Y are the edge labels. Areas can be added together to find A *or* B, where A and B are areas.

The order of the events and factors is not meaningful.

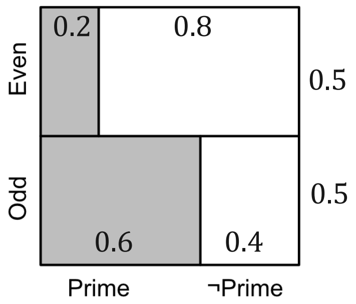
Examples

1.



In a deck of cards, half are red, and half are black. No red cards are clubs. Half the black cards are clubs.

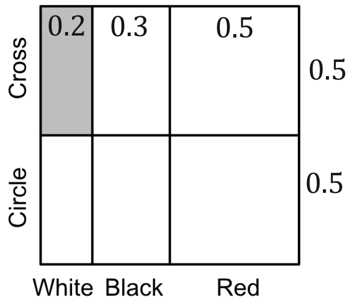
2.



Of the numbers between 1 and 10, half are even, and half are odd. One of the five even numbers is prime. Three of the five even numbers are prime.

Thus 40% numbers are prime.

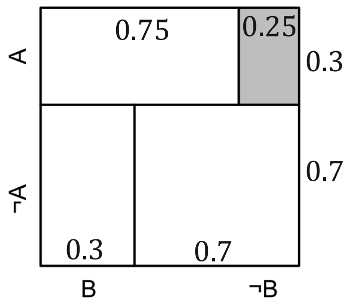
3.



Counters are 20% white, 30% black, and 50% red. On one side they have a cross, and the other they have a circle, with an even chance of being either side.

The probability of a white counter showing a cross is 10%.

4.



The probability of A is 30%. The probability of B given A is 75%, but only 30% given not A.

Thus, the probability of A and not B is 7.5%.

Representation – Bayesian algebra

Summary

Bayesian algebra consists of numbers, letters, and words, which are combined using standard mathematical operations ($+$, $-$, $\times$, $\div$) and probability functions $\mathrm{P}(x)$ and $\mathrm{P}(x|y)$ which map events numbers between 0 and 1. Symbols or “and”, “or”, and “not” (∩, ∪,¬) are used to combine events.

Progress is made by rewriting equations through applying operations, simplifying equations, and rearranging terms.

The size and absolute position of equations have no meaning.

Examples

1.
$$\begin{aligned} \mathrm{P}(\mathrm{red}) &= 0.5 \\ \mathrm{P}(\mathrm{club}) &= 0.25 \\ \mathrm{P}(\mathrm{club} \mid \mathrm{red}) &= 0 \end{aligned}$$

In a deck of cards, half are red, and one quarter are clubs. If the card is red then it cannot be a club.

2.
$$\begin{aligned} \mathrm{P}(E) &= 0.5 \\ \mathrm{P}(P \cap E) &= 0.1 \\ \mathrm{P}(P \mid E) &= \frac{\mathrm{P}(P \cap E)}{\mathrm{P}(E)} = 0.2 \end{aligned}$$

Let U be the set of integers from 1 to 10. Let E be the event that a number from U is even and let P be the event that a number from U is prime. The probability that a number from U is both prime and even is 0.1. Then the probability that a number in U is prime given that it is even is 0.2.

3.
$$\begin{aligned} \mathrm{P}(M) &= 0.92 \\ \mathrm{P}(N) &= 0.24 \\ \mathrm{P}(M \mid N) &= 0.75 \\ \mathrm{P}(M \cap N) &= \mathrm{P}(M \mid N) \cdot \mathrm{P}(N) \\ &= 0.75 \times 0.24 = 0.18 \end{aligned}$$

The probability of M is 0.92, and N is 0.24. Given N, the probability of M becomes 0.75. Thus the probability of both M and N is 0.18.

4.
$$\begin{aligned} \mathrm{P}(\mathrm{meow} \mid \mathrm{hungry}) &= 90\% \\ \mathrm{P}(\mathrm{hungry}) &= 10\% \\ \mathrm{P}(\mathrm{meow}) &= 15\% \\ \mathrm{P}(\mathrm{hungry} \mid \mathrm{meow}) &= \mathrm{P}(\mathrm{meow} \mid \mathrm{hungry}) \cdot \frac{\mathrm{P}(\mathrm{hungry})}{\mathrm{P}(\mathrm{meow})} \\ &= 60\% \end{aligned}$$

The cat will meow if it is hungry 90% of the time. The cat is hungry 10% of the time, and the cat meows 15% of the time. Thus, the probability that the cat is hungry given that it is meowing is 60%.

Representation – Contingency tables

Summary

A contingency table is a grid where the first row and column are reserved for labels, which (along each axis) are mutually exclusive but together are all possible outcomes. Labels may use the symbol “not” (¬).

The final row and column contain numbers which must be the sum of the numbers in their own (completely filled) row/column. The value in the final cell is always 1.

Inner cells are filled with real values between 0 and 1, and represent the probability of X *and* Y, assuming labels X and Y align with that cell.

The size of the cells has no meaning.

Examples

1.

	Red	Black	Total
Club	0.0	0.25	0.25
¬Club	0.5	0.25	0.75
Total	0.5	0.5	1

From a deck of cards, the probability of being red and a club is 0, red and not a club is 0.5, black and a club is 0.25, and black and not a club is 0.25.

2.

	Even	Odd	Total
Prime	0.1	0.3	0.4
¬Prime	0.4	0.2	0.6
Total	0.5	0.5	1

For the numbers from 1 to 10, the probability of a number being even and prime is 0.1, even and not prime is 0.4, odd and prime is 0.3, and odd and not prime is 0.2.

3.

	X	¬X	Total
Y	0.18	0.22	0.4
¬Y	0.27	0.33	0.6
Total	0.45	0.55	1

The probability of X and Y is 0.18, X and not Y is 0.27, not X and Y is 0.22, and not X and not Y is 0.33.

4.

	Young	Mid	Old	Total
Vote	0.08	0.27	0.25	0.6
¬Vote	0.12	0.23	0.05	0.4
Total	0.2	0.5	0.3	1

From a population, the probability of a citizen being young and voting is 0.08, young and not voting is 0.12, middle aged and voting is 0.27, middle aged and not voting is 0.23, old and voting is 0.25, and old and not voting is 0.05.

Representation – Euler diagrams

Summary

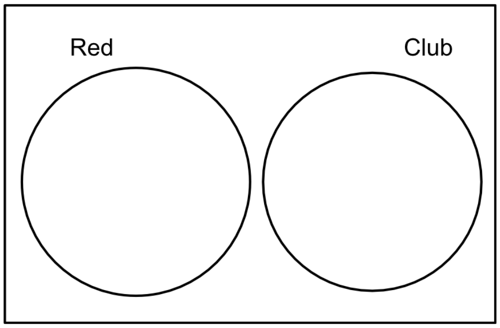
Euler diagrams consist of a “universe” denoted by a rectangle, and ellipses representing events. Events are named with letters or words.

The region inside the curve represents events occuring. Regions inside two curves represent X and Y occuring simultaneously. Regions that do not overlap are disjoint.

The size or shape of the curves are not meaningful.

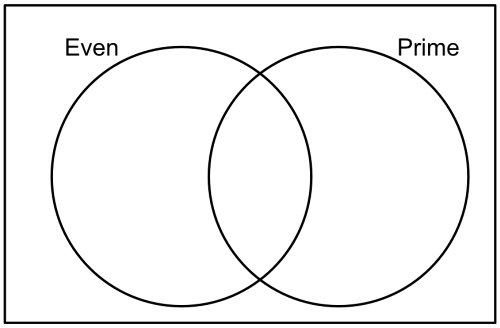
Examples

1.



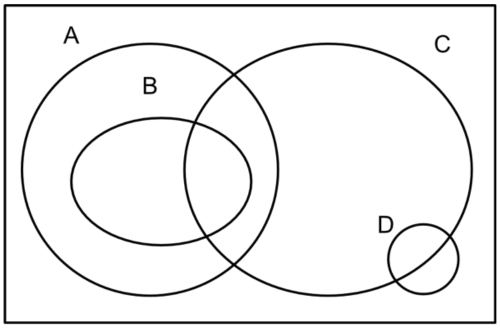
Some cards are red. Some cards are clubs. No card is a red club.

2.



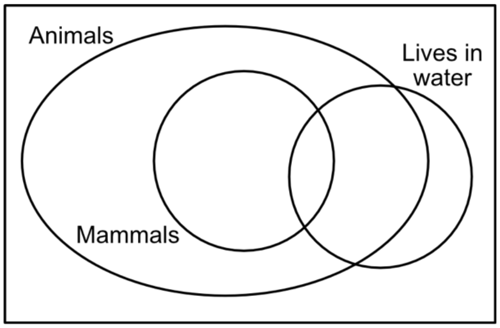
There are even numbers. There are prime numbers. There are even and prime numbers.

3.



Some (but not all) As are Cs, and some (but not all) Cs are As. All Bs are As, and some (but not all) Bs are also Cs. Some (but not all) Ds are Cs, but no D is also an A.

4.



All mammals are animals, but not all animals are mammals. Some mammals live in water, but some do not; some animals live in water, but some do not. Some things that live in water are not animals.

Representation – Probability trees

Summary

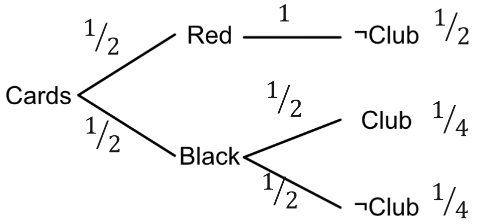
Probability trees consists of events and branches. Events sometimes use a “not” symbol (¬). Each event has exactly one “prevous” event, except for the first event which has no previous. Branches are labelled with the probability of the next event occuring given that the previous event has occurred. The sum of adjacent branches must be 1.

X *and* Y is computed by multiplying along branches; X *or* Y by adding between branches.

Neither the length of branches nor the order of adjacent events is meaningful.

Examples

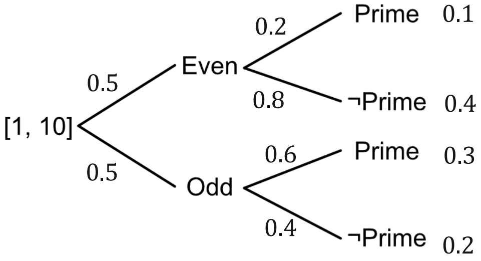
1.



$P(club) = {}^1/_4$

Half of the cards in a deck are red, the other half are black. No red card is a club, but half the black cards are a club. The total probability of getting a club is ¼.

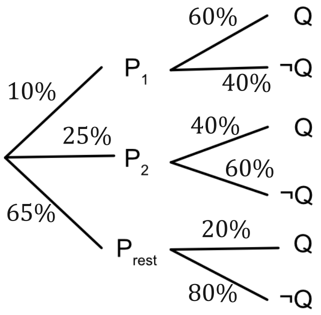
2.



$P(prime) = 0.1 + 0.3 = 0.4$

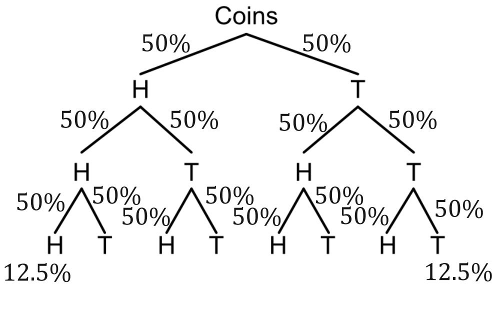
For the numbers from 1 to 10, half of the numbers are even. One of the five even numbers is prime. Three of the five even numbers are prime. The total probability of a number between 1 and 10 being prime is 0.4.

3.



The probability of P_1 is 10%, P_2 is 25%, and the remaining Ps together have probability 65%. If P_1 is true, then Q has probability 60%, whereas given P_2 Q has probability 40%. Otherwise, Q has probability 20%.

4.



$P(HHH \cup TTT) = 12.5\% + 12.5\% = 25\%$

Toss three coins, each with a 50% chance of begin heads or tails. The probability of getting all heads or all tails is 25%.