Examining Expert Representation Recommendations 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 remains an open problem. Indeed, there are very few tools that support problem solvers to even transform their problem between representations, let alone recommend suitable representations. Previously, we built an intelligent representational system recommender, Robin, within our *rep2rep* framework. In this paper, we present a study that examines how human experts choose representations in order to establish a benchmark for evaluating such tools. We asked high school mathematics teachers to produce representational system recommendations which we could compare against Robin's output. We found the teachers updated their recommendations based on the problem and student profile, but were inconsistent with each other. Still, where the teachers do tend towards agreement, we find Robin produces scores that are in line with their responses. The inconsistency between the teachers highlights a need for more training and support in representational system selection, and presents an opportunity for AI tools to improve pedagogy.**

***Index Terms*—representations, heterogeneous reasoning, problem solving, artificial intelligence, intelligent tutoring systems**

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]. This motivates the need for new AI tools: representational system recommenders that can consider both the problem being solved, and the person solving the problem.

One major application for such an AI tool is in education [5]. 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 [6], [7]. There is evidence that proper use of RSs in mathematical problem solving can improve learning [8], [9]. 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 are able to use these data to evaluate an AI tool, called Robin, that we developed in prior work as part of the *rep2rep* project [10], [11]. In essence, our contributions are two-fold. First, 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. Second, we compare the teachers' recommendations against those made by Robin.

We found that teachers do take into consideration 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. With disagreement between the teachers, comparisons with Robin are limited, but we emphasise the need for such AI tools to help bring consistency and clarity for both teachers and their students when selecting RSs for problem solving.

This paper presents our study and results in the following order. In Section II we present our hypotheses for our experiment, 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

The experiment we present is two-fold. First, 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' of an RS. Second, we compare the scores produced by Robin, the *rep2rep* framework implementation [10], with the teachers' responses; their informational suitability rankings are the benchmark against which we compare our framework.[1].

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.

[1]While the *rep2rep* framework has cognitive properties [12], adding this to Robin is ongoing work. We only evaluate Robin for informational suitability, but explore whether teachers consider the students' cognitive profile.

We break down our high level goals into four hypotheses. Our first hypothesis can be stated as follows:

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 subsequent two hypotheses), but they are also 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 rank of RSs.

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.

The recommendation should also vary based on the cognitive abilities of the student for which they are making the recommendation:

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

Finally, we make a direct comparison with the *rep2rep* framework implementation, Robin. We anticipate that the framework is producing scores comparable to those assigned by experts. Thus, our fourth hypothesis is:

H4. The Robin implementation of the *rep2rep* framework produces scores that are correlated with the teachers' responses, when considering the informational suitability of the same problem and RS.

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. The overall flow of each experiment session was:

1) Introduce and explain the experiment;
2) Train the teachers in each RS;
3) Phase one: evaluate each RS for each problem *without* considering any personas;
4) Phase two: evaluate each RS for each problem *with* consideration to the personas; and
5) Debrief, interview, and the demographics survey.

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 which can be 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 can be added together for the probability of $A \cup B$. See Fig. 1(a).

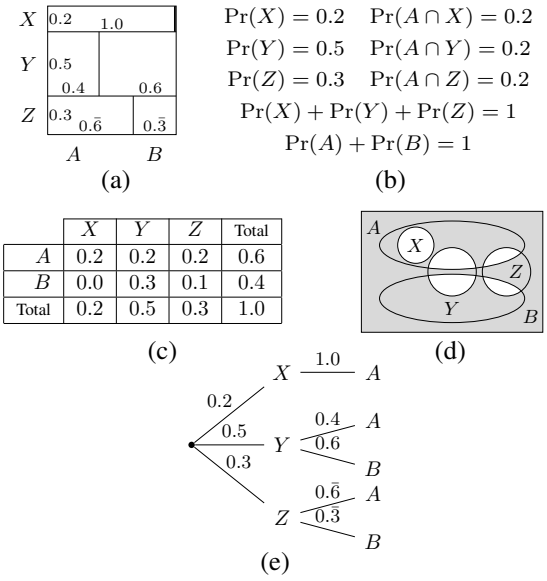


Fig. 1. (a) A probability tree, (b) a contingency table, (c) an area diagram, (d) an Euler diagram, and (e) five Bayesian algebra equations. Representations are (near) equivalent in describing events A, B, X, Y, and Z.

Bayesian Algebra is standard algebraic notation, augmented with two probability functions $\Pr(\cdot)$ and $\Pr(\cdot \mid \cdot)$, conditional probability laws, and Bayes' Theorem. See Fig. 1(b).

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 $X \cap Y$. Using these rules, missing values can be computed. See Fig. 1(c).

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 specifically considered *non-proportional* Euler Diagrams. See Fig. 1(d).

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. While the edges between the nodes are meaningful, their length, order, and position are not. See Fig. 1(e).

The choice to include Area Diagrams was motivated because we were aware that this RS is *not* commonly taught in New Zealand. While we do not actively use this fact in our analysis, we wish to see what effect an unfamiliar RS has on the teachers' responses. We provided training for the teachers in all RSs, regardless of familiarity, before starting the main tasks.

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: informational suitability (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 informational suitability context, participants were not given any persona to consider when scoring the RSs. For the contexts involving students, we provided two *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.

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 responses for this problem were not used in the analysis; the teachers were *not* made aware that their responses would be discarded for this problem.

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.

For convenience, we categorise the final four as 'easy' or 'hard', and 'verbal' or 'formulaic'. We use the abbreviations E and H, and V and F, respectively, to name problems 2 through 4 as EV, EF, HV, and HF.

D. Training

Because the teachers may not all be familiar with the RSs used in the study – or may have a different understanding of the RS to what we intend – 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. All five training documents are in Appendix B.

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 promoted as 'better' than the others. The examples were a representation belonging to that RS, and a short textual description of the representation. Participants 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.

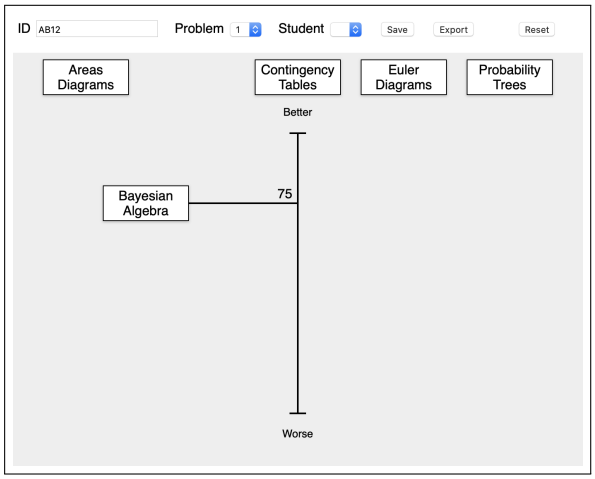


Fig. 2. A screenshot of the response form participants used to arrange RSs according to the informational and cognitive suitability.

E. Tasks

The experiment was divided into two phases: in phase one, the teachers were to assess the 'informational suitability' of the RSs for each problem; in phase two, they were to assess the suitability of the RSs for each problem for a specific student persona. For each problem and cognitive context, the participants were asked to arrange the RSs on the online response form shown in Fig. 2. The participants entered their identification code, the problem, and cognitive context, then dragged the labels of the RSs onto the central scale, 0 to 100. The draggable boxes all begin in the top row, can be dragged anywhere, and may overlap freely. The boxes have a horizontal line that always connects to the central scale; the horizontal position has no meaning, and participants were informed as such. When they were happy with their response, they clicked the 'Save' button, then 'Reset' to return the labels to the top row. Through the relative vertical positioning of the RSs, we will be able to address all three of our hypotheses.

For phase one on informational suitability – that is, the RSs' suitability when considering *only* the problem, and not who

might solve it – we presented the teachers with the problem statement, requested that they read the problem (and to *not* solve the problem), asked if they had any questions, then asked:

> Thinking in the general case, how informationally suitable do you think each representation would be? How well it captures the important parts of the problem, and how well it can be used to solve the problem.[2]

They then proceeded to arrange the RS labels on the response form. We used these responses to address H4.

After all problems had been presented to the teachers, and their responses saved, we moved onto phase two wherein we asked the teachers to consider not just the informational suitability of the problem, 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 first arranged them based on the following prompt:

> Please arrange the representations based on how suitable each is for Student A to solve the problem.

Once their responses had been saved for Student A, we immediately did the same problem for Student B. They completed all problems as before, and then we moved onto the debrief and questioning.

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?

We followed interesting discussion points as they arose.

Participants 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

The participants of this study were high school mathematics teachers in New Zealand. We advertised the study by directly reaching out to the heads of faculty of high schools in Canterbury. We recruited 10 teachers in total (3 male, 7 female) from five separate schools; nine teachers returned usable quantitative data – one teacher returned data that had somehow been corrupted. The participants' teaching experience ranged from two-and-a-half to sixteen years, and all had been mathematics teachers for the entirety of their teaching career. All have a bachelors degree and a postgraduate diploma in Teaching; the major of the degree was varied. One participant has a doctoral degree in Statistics, while one has a masters degree in Computer Science. One teacher was studying for a masters degree in Specialist Teaching at the time of the experiment. All had taught courses that included probability content within the past two years.

All the teachers 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

During the introduction we motivated the purpose of this experiment to understand how teachers consider solving problems, both in general and for students – the participants were instructed to *not* solve the problems we gave them. We also explained terms such as 'representational system' and 'informational suitability'.

B. Training

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

- They were unfamiliar with Area Diagrams (but one had seen eikosograms before, which are related [13]).
- They were familiar with Contingency Tables under the name 'two-way tables'.
- They were familiar with Euler Diagrams under the name 'Venn diagrams'; this is the name used by the NCEA standards specification documents.[3]

This training period lasted about 30 minutes.

C. Representational systems without cognitive context

We then began phase one. We presented the teacher with each of the five problems, and asked them to read and understand each problem but *not to solve it*. The teacher was asked if they understood the problem; every response was affirmative. We then asked them to position the labels of the RSs in the web interface based on their informational suitability. The problems were presented such that the medical testing problem was always given first, as a 'practice' (although this was not disclosed), and then in a counterbalanced manner with the restriction that the pairs of 'equivalent' problems never follow each other.

D. Representational systems with cognitive context

After completing the evaluation task for each problem only for informational suitability, 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 participant 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. Each of the problems were then presented in the same order as the previous phase, but this time the teachers were asked to evaluate the RSs for first the lower ability student, then immediately after for the higher ability student.

[2]This clarifying fragment acted as a prompt; we explained informational suitability when introducing the participants to the experiment.

[3]What are called Venn diagrams are in fact Euler diagrams.

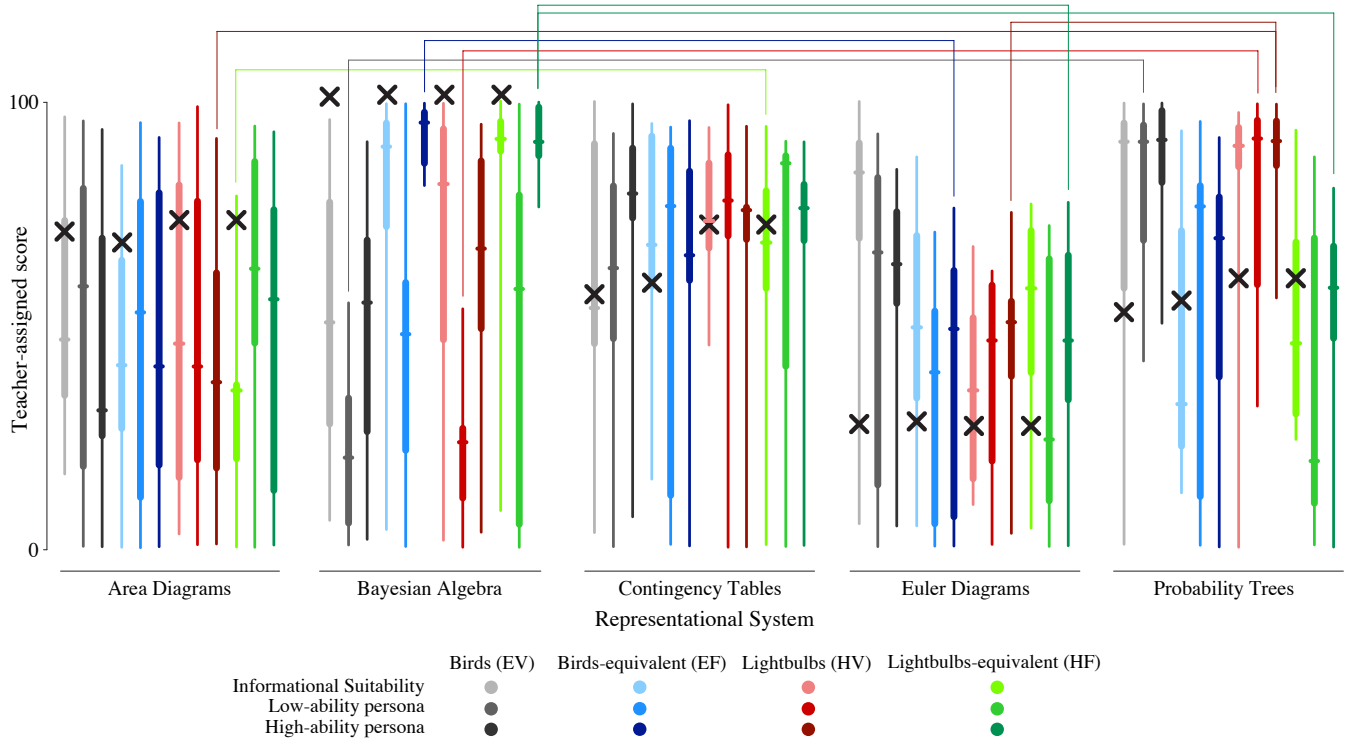


Fig. 3. The distributions of the teachers' assigned scores for each RS, as box-and-whisker plots, colour-coded for each problem and cognitive context (legend at the bottom). Connecting lines (top) indicate between which RSs we find significant differences in rankings, when considering each ⟨problem, cognitive context⟩ pair. The black crosses are Robin's scores, scaled to the same $[0, 100]$ interval as the teachers' scores.

E. Debrief and questions

To end, we asked the participants 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 participants 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 ⟨ problem, cognitive context ⟩ pair, we consider all of the responses from the participants. In this section, we explore two representative examples: the problem that is equivalent to the lightbulbs problem when considering high-ability student persona (⟨ 5/HF, high-ability ⟩), which shows clear groupings in the responses; and the birds problem when considered without any persona (⟨ 2/HV, no persona ⟩), which does not show clear groupings. Tables of all statistical test results are included in Appendix A.

A. Lightbulbs-equivalent problem for high-ability persona

The lightbulbs-equivalent problem is number 5/HF, in Section III-C. We asked the teachers to consider each RS, and in this case they evaluated them based on their suitability for the high-ability Year 13 student persona.

To better understand the teachers' responses, we first plot the data in Fig. 3, focusing on the dark green box-and-whisker plots (the right-most column within each RS). We note two things: first, the Bayesian Algebra distribution is visually much tighter than the others – indicating more agreement between the participants – and also much higher – the participants believed this representational system to be generally more suitable for this problem, and this student persona. Second we see, in particular, Area Diagrams scores are spread out – the participants do *not* agree with each other. The first observation 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. As we mentioned, the responses for Area Diagrams are too spread to make a clear statement: for this RS, the teachers were inconsistent with each other. We note that this was the RS with which our participants were least familiar.

Following this visual inspection, we performed a more formal evaluation. Due to few scores, and their significant non-normal distribution we used non-parametric statistics. The teachers' responses were integers from 0 to 100; we converted these to ranks for each RS, preserving ties. Using these ranks, we performed a Friedman test between the mean rankings of each

RS. For the problem and cognitive context described above (⟨ 5/HF, high-ability ⟩), 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 (for ten comparisons): 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. 3 marks both with a connecting line. Thus we can state that we have evidence that the participants would recommend BAYESIAN ALGEBRA over EULER DIAGRAMS and PROBABILITY TREES. While not a comprehensive ranking, we have extracted a ranking from the teachers' responses, giving tentative evidence for H1.

B. Birds problem without any persona

The birds problem is number 2/EV, in Section III-C. As before, we plot the teachers' responses in Fig. 3, now focusing on the light grey box-and-whisker plot (left-most column for each RS). This time, any patterns are much less clear. All the RSs' scores are spread across the scale, with none clearly being better or worse than the others. We might generously state that EULER DIAGRAMS has scores that are typically higher than the others, but this is far from conclusive. We also notice a slight separation on CONTINGENCY TABLES, but the cause or meaning is difficult to determine. Unlike last time, there is no apparent 'better' RS.

After performing the same transformation from scores to ranks, we can perform a Friedman test to determine if there exists a significant difference between the rankings of the RSs. No significant difference was found ($Q = 7.75$, $p = 0.101$), which matches our visual intuition. Thus in this case, we have no evidence supporting H1, that the teachers were able to agree on the informational suitability 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 ⟨ problem, cognitive context ⟩ pairs are representative of the results from all twelve pairs. Fig. 3 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 – ⟨ 2/EV, high-ability persona ⟩, ⟨ 3/EF, low-ability persona ⟩, and ⟨ 4/HV, informational suitability ⟩ – 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 statistical test results are in Appendix A.

Based on the summarised results, we see that 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 our 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

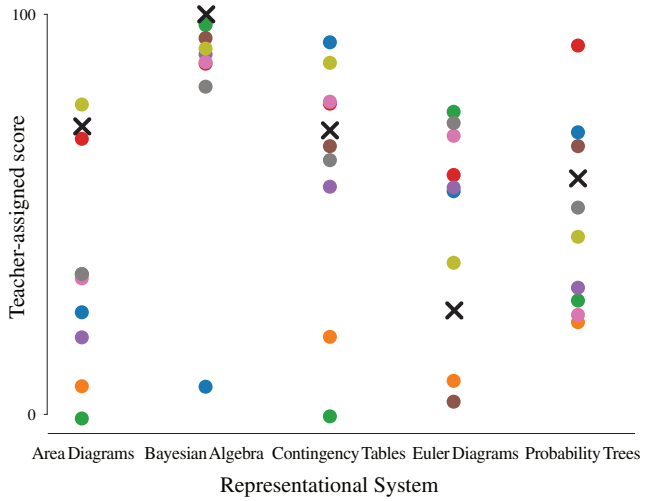


Fig. 4. The teachers' assigned scores for each RS, when asked to consider their informational suitability when solving the problem equivalent to the lightbulbs problem, 5/HF. The black crosses are Robin's output, scaled to $[0, 100]$. Each colour is an individual participant

the problem and the cognitive context seem to be having an influence on the result. There is some limited matching in three cases: in the birds variants, problems 2/EV and 3/EF, for informational suitability (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 participants. 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.

D. Comparisons with the rep2rep framework

H4 contrasts the results of the *rep2rep* framework implementation, Robin, with the responses of the participants. We restrict our analysis to the informational suitability ratings. From these four cases – one for each problem – we can only discern one pattern from the teachers responses: the CONTINGENCY TABLES would be more informationally suitable than AREA DIAGRAMS for the problem equivalent to the lightbulbs problem, 5/HF, (read from Fig. 3, only one connecting line occurs between 'light' box-and-whisker plots, in this case light green). Posing the same lightbulbs-equivalent problem to Robin,[4] we get the following ordering:

Bayes 8.6 Areas 6.2 Contingency 6.1 Trees 5.1 Euler 2.2

We see that the algorithm was *not* able to clearly separate AREA DIAGRAMS (6.2) from CONTINGENCY TABLES (6.1).

[4]We use the input based on BAYESIAN ALGEBRA, rather than natural language, as we find this has more specific, structured information for the algorithm to work with.

Using the median value of the teachers' responses, shown in Fig. 3, light green box-and-whisker plots (third column from the right within each RS), we can extract the following ordering, best to worst:

Bayes Contingency Euler Trees Areas

which agrees on the ordering of BAYESIAN ALGEBRA, CONTINGENCY TABLES, and PROBABILITY TREES relative to one another, and on the relative ordering of BAYESIAN ALGEBRA, CONTINGENCY TABLES, and EULER DIAGRAMS. The placement of AREA DIAGRAMS and EULER DIAGRAMS differ significantly between the two orderings. To assess the rankings' similarity, we use Kendall's rank correlation coefficient (τ) to determine the correlation of the teachers with our framework [14]. Running Kendall's τ-B test on the two orderings returns a non-significant result ($\tau = 0.40$, $p = 0.48$), failing to find evidence of correlation.

Performing the same τ-B analysis between Robin and each teacher individually, across all the problems, finds no instance of significant correlation.[5] So we have failed to find evidence of correlation between Robin's scores and the teachers' responses, both individually and in aggregate. This is not surprising, given the teachers do not seem to correlate with each other.[6]

More informally, we can inspect how Robin's output compares to the responses of the participants by plotting the scores alongside the participants. We scale the scores from the interval $[0, \max(\text{scores})]$ – where max is computed per problem – to $[0, 100]$, and plot these as black crosses in Fig. 3. For clarity, we extract just the lightbulbs-equivalent problem, 5/HF, into Fig. 4; in this figure, each colour is an individual participant. Visually inspecting the placement of the crosses, and the general groupings of the points from the participants' responses, we see that the crosses generally follow the 'mass' of the points.

Of the four cases where we can compare Robin's scores against the participants' responses, we would argue that in three plots Robin's scores fall within the distribution of the teachers' scores: 3/EF, 4/HV, and 5/HF. (blue, red, and green box-and-whisker plots, respectively).[7] While the participants' responses for the birds problem (2/EV) are too spread out and Robin's score for EULER DIAGRAMS too different, the framework's scores for the remaining three problems visually follow the same groupings. So to address H4, while we cannot state with any statistical confidence that our framework produces scores that agree with the participants' responses, there are indications that the framework is evaluating RSs in a manner comparable to that of our expert participants. A larger study may allow this analysis to be formalised.

[5] 32 τ tests, all with $p > 0.05$. In one case for the birds problem (2/EV) we have $p = 0.083$, an *almost* significant correlation, but $\tau = -0.8$ meaning the teacher is *negatively* correlated with Robin.

[6] 144 τ tests between all pairs of teachers, all with $p > 0.05$. We have 15 cases (about 10%) where $p < 0.1$, 12 of which are positively correlated, three negatively. This is likely due to chance, and is a Type-I error.

[7] The non-normal nature of the data prevents us from verifying this using parametric statistical tests such as t-tests.

VI. QUALITATIVE ANALYSIS

The inconsistency of the participants prevented us from making more extensive comparisons between the scores they assign and those computed by the *rep2rep* framework. We propose two possible causes for this inconsistency: there is an underlying factor we did not control for; or the task is difficult, even for experts. Let us consider each in turn.

Before the study, we identified three factors that might influence the participants' responses that we could not control:

- external motivation for using one RSs over another;
- preferring some RSs over others; and
- experience as a teacher, and in using particular RSs.

During the debriefing interview we asked the participants four questions, which we stated in Section III. We asked the participants how familiar they had been with each RS prior to the training we provided. The participants 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.

While more than half of teachers initially answered that their responses were primarily based on 'gut instinct' rather than external factors, further discussion revealed influences from the curriculum towards PROBABILITY TREES and CONTINGENCY TABLES, including in the standard assessments.

Participants 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 [15]. This highlights a need for frameworks that allow for using more diverse RSs: by reducing the barrier to using multiple RSs, our *rep2rep* framework could support improved learning opportunities, for both students and teachers.

We asked participants if they had any 'go-to' RSs when teaching probability. All responded with either PROBABILITY TREES or CONTINGENCY TABLES, with half pointing out that these are encouraged by the assessment standards, as mentioned above. We observe these RSs (particularly PROBABILITY TREES) to be favoured by our participants. EULER DIAGRAMS are only introduced at Year 13 in the New Zealand curriculum – students typically aged 17 to 18 – and many teachers acknowledged they were reluctant to use them purely because they felt they were 'too hard' for students.

> The kids do find [EULER DIAGRAMS] quite challenging, so you tend to only use those when you really need to, or when they're indicated as a 'good way' to solve that particular problem [by resource materials]. [...] They're the last resort.

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 participants 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 working with a different curriculum, or with 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 participants to favour BAYESIAN ALGEBRA, which also uses letters as variables and a probability function. Indeed, we see this in Fig. 3: every situation where BAYESIAN ALGEBRA is preferred is a 'context-less' problem.[8] In future studies, we suggest using 'equivalent' problems that retain a 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 participants 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 participants 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 general trends, but our participants did not consistently agree with each other – against our initial hypotheses. The inconsistency, and the participants' explicit mention of 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. This reinforces a need for tools such as those from the *rep2rep* project that are able to support heterogeneous reasoning [16].

VII. LIMITATIONS AND THREATS TO VALIDITY

We identify three limitations and threats to the validity of this study, which we were unable to address in the study design.

Number of participants While nine participants provides useful information, the lower number means the power of the study is limited. The responses from this study provide interesting preliminary data, but the fields of information representation and education would both benefit from a larger version of this study: does the disagreement we have found continue to be present over larger groups of participants? Are there regularities in the disagreements that we were unable to discover? Can a more complete recommendation be extracted from a larger pool of participants?

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 potentially reduces the diversity of our participants, and their responses. The scope of this study was also restricted to probability problems. While these restrictions reduce the variability in our sample, they restrict the generality of our results.

Mismatch of problems A design limitation of this study is that the initial problems and RSs selection was based on the mathematics curriculum in England, but due to the COVID-19 pandemic the study was updated to run in New Zealand. While some changes could be made quickly – such as translating the English GCSE/A-levels student personas to the NZQA framework – we could not make others because we did not identify them ahead of time. One notable change is the order and age at which different RSs are introduced by each curriculum. For example, in England, EULER DIAGRAMS[9] are introduced at 'Key Stages' 3 and 4 [7], aimed at students aged 11 to 14; in New Zealand, EULER DIAGRAMS are introduced at NCEA Level 3 [6], aimed at students aged 17 to 18. Such a large difference is surprising, and may account for why our New Zealand-based participants were reluctant to recommend EULER DIAGRAMS: they associate them with advanced mathematics content.

VIII. CONCLUSIONS

This study, while not able to directly evaluate the framework as intended, 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 do indicate that H2 (that the problem is a factor in their evaluation) and H3 (that the cognitive context is a factor in their evaluation) may be correct. Further studies are needed to determine the influence of these factors – and potentially others – on the final recommendation. Finally, H4 is weakly supported by the apparent grouping of our framework's scores and the teachers' responses; our *rep2rep* framework is capturing at least some of what influences the teachers' RS recommendations. Together these four hypotheses give some evidence that our framework is grounded and useful: when the teachers make recommendations that are broadly aligned, the *rep2rep* framework makes a comparable RS recommendation.

A version of this paper without appendices has been submitted to 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.

[8] They are also exclusively for the high-ability student persona.

[9] Both curricula refer to EULER DIAGRAMS as Venn diagrams.

[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] B. Grawemeyer, "Evaluation of erst – an external representation selection tutor," in *Diagrammatic Representation and Inference, Diagrams 2006*, ser. Lecture Notes in Computer Science, D. Barker-Plummer, R. Cox, and N. Swoboda, Eds. Springer, 2006, pp. 154–167.

[6] 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).

[7] 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).

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

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

[10] D. Raggi, G. Stapleton, A. Stockdill, M. Jamnik, G. Garcia Garcia, and P. C.-H. Cheng, "How to (re)represent it?" in *IEEE 32nd International Conference on Tools with Artificial Intelligence, ICTAI 2020*, M. Alamaniotis and S. Pan, Eds., 2020, pp. 1224–1232.

[11] D. Raggi, A. Stockdill, M. Jamnik, G. Garcia Garcia, H. E. A. Sutherland, and P. C.-H. Cheng, "Dissecting representations," in *Diagrammatic Representation and Inference, Diagrams 2020*, ser. Lecture Notes in Computer Science, A.-V. Pietarinen, P. Chapman, L. Bosveld-de Smet, V. Giardino, J. Corter, and S. Linker, Eds., vol. 12169. Springer, 2020, pp. 144–152.

[12] P. C.-H. Cheng, G. Garcia Garcia, D. Raggi, A. Stockdill, and M. Jamnik, "Cognitive properties of representations: A framework," in *Diagrammatic Representation and Inference*, A. Basu, G. Stapleton, S. Linker, C. Legg, E. Manalo, and P. Viana, Eds. Springer, 2021, pp. 415–430.

[13] 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.

[14] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

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

[16] J. Barwise and J. Etchemendy, "Heterogeneous logic," in *Logical Reasoning with Diagrams*, G. Allwein and J. Barwise, Eds. Oxford University Press, 1996, ch. 8, pp. 179–200.

APPENDIX A
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 B
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

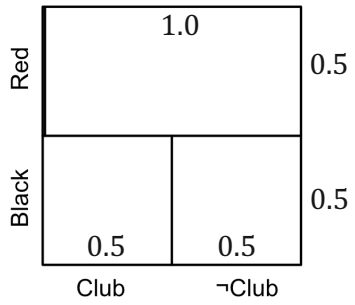
An area diagram is a unit square representing all possible outcomes, with labels for events, their split length representing the probability of each event. Labels might use “not” (¬)

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.

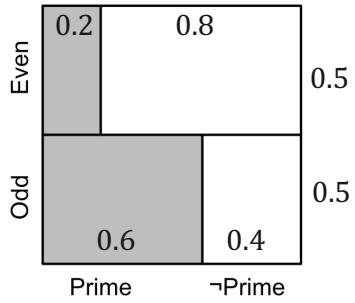
Examples

1.



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

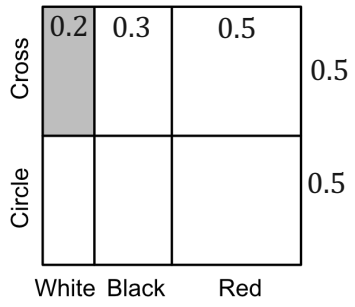
2.



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.

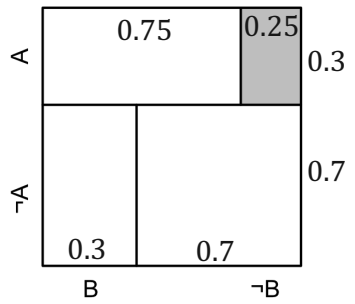
3.



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" ($\cap$, $\cup$,$\neg$) 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. $$\begin{aligned}\mathrm{P}(\text{red}) &= 0.5\\ \mathrm{P}(\text{club}) &= 0.25\\ \mathrm{P}(\text{club} \mid \text{red}) &= 0\end{aligned}$$

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

   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

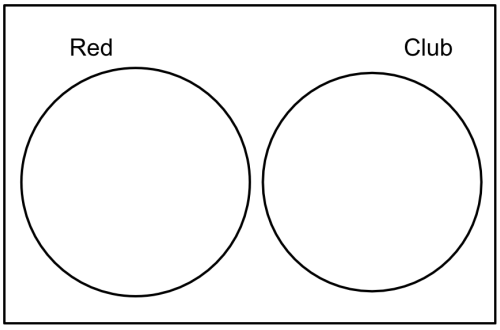
Summary

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

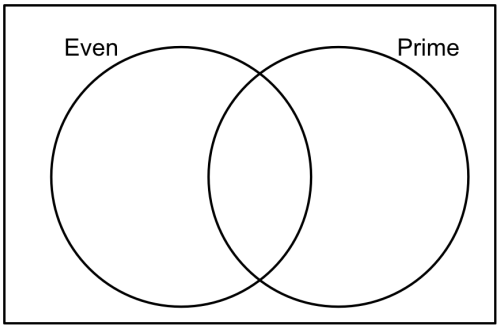
The region inside the curve represents events occuring. Regions inside two curves represent X and Y occuring simultaneously. Regions that do not overlap are disjoint.

The size or shape of the curves are not meaningful.

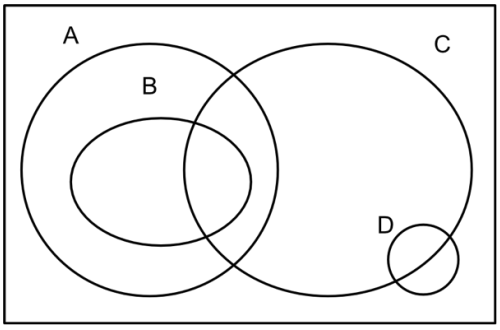
Examples

1. 

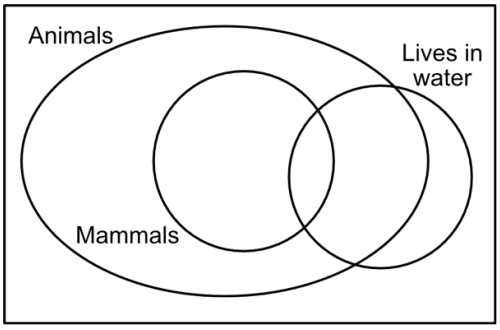
Some cards are red. Some cards are clubs. No card is a red club.

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

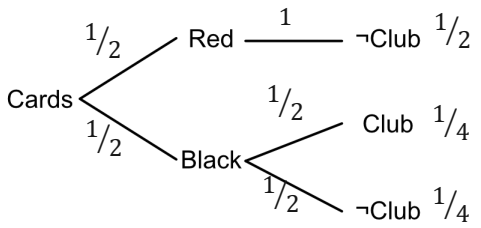
Probability trees consists of events and branches. Events sometimes use a “not” symbol (¬). Each event has exactly one “prevous” event, except for the first event which has no previous. Branches are labelled with the probability of the next event occuring given that the previous event has occurred. The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

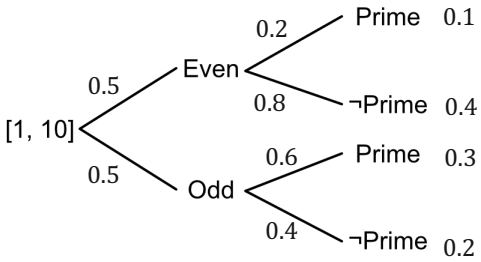
1.



$$P(club) = 1/4$$

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.

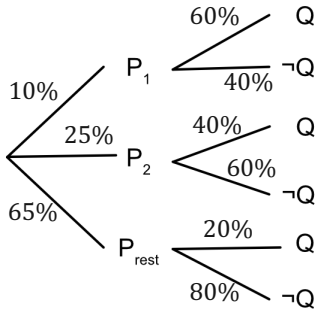
2.



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

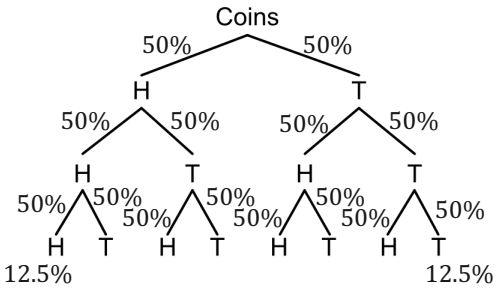
For the numbers from 1 to 10, half of the numbers are even. One of the five even numbers is prime. Three of the five even numbers are prime. The total probability of a number between 1 and 10 being prime is 0.4.

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%.