Journal of Logic, Language and Information (2021) 30:557–600
https://doi.org/10.1007/s10849-021-09331-0

Evaluating Free Rides and Observational Advantages in Set Visualizations

Andrew Blake[1] · **Gem Stapleton**[2] · **Peter Rodgers**[3] · **Anestis Touloumis**[1]

Accepted: 13 March 2021 / Published online: 15 April 2021

Abstract

Free rides and observational advantages occur in visualizations when they reveal facts that must be inferred from an alternative representation. Understanding whether these concepts correspond to cognitive advantages is important: do they facilitate information extraction, saving the 'deductive cost' of making inferences? This paper presents the first evaluations of free rides and observational advantages in visualizations of sets compared to text. We found that, for Euler and linear diagrams, free rides and observational advantages yielded significant improvements in task performance. For Venn diagrams, whilst their observational advantages yielded significant performance benefits over text, this was not universally true for free rides. The consequences are two-fold: more research is needed to establish when free rides are beneficial, and the results suggest that observational advantages better explain the cognitive advantages of diagrams over text. A take-away message is that visualizations with observational advantages are likely to be cognitively advantageous over competing representations.

Keywords Linear diagrams · Venn diagrams · Euler diagrams · Free rides · Observational advantages

This research was partially funded by a Leverhulme Trust Research Project Grant RPG-2016-082 for the project entitled Accessible Reasoning with Diagrams.

✉ Gem Stapleton
ges55@cam.ac.uk

Andrew Blake
a.l.blake@brighton.ac.uk

Peter Rodgers
p.j.rodgers@kent.ac.uk

Anestis Touloumis
a.touloumis@brighton.ac.uk

[1] University of Brighton, Brighton, UK

[2] University of Cambridge, Cambridge, UK

[3] University of Kent, Canterbury, UK

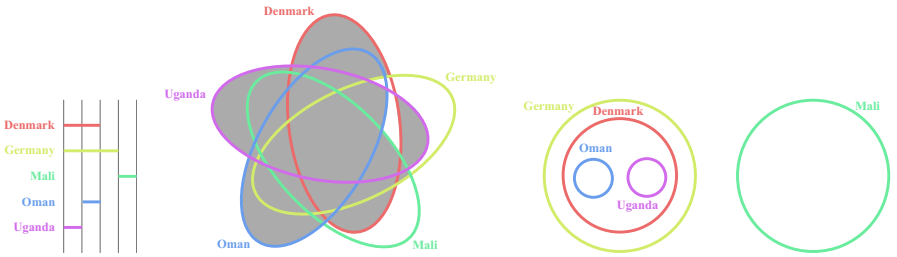


Fig. 1 Diagrams obtained from translating textual statements: (left to right) linear, Venn, and Euler

1 Introduction

This paper sets out to empirically test the belief that free rides (Shimojima 2015) and observational advantages (Stapleton et al. 2017) are features of visual modes of communication that aid cognition. Free rides occur when a given representation of information *is translated into another* and the resulting representation makes explicit some facts that must be derived (inferred) from the original. Such explicit facts are precisely the free rides. The generalisation of free rides to observational advantages removes the requirement for a translation. Instead, it allows the comparison of competing representations that are semantically equivalent: if one representation makes a fact explicit that must be inferred from another representation then that fact is an observational advantage of the former over the latter. In this paper, we study the potential cognitive benefits of free rides and observational advantages in the context of representations in textual form and diagrammatic form. The specific research questions we address are:

(RQ1) does using text *alongside* (semantically equivalent) diagrams that are derived from the text, lead to significant performance benefits over *just* using text when identifying information that is conveyed by free rides?
(RQ2) do diagrams lead to significant performance benefits over using text when identifying information that is conveyed by observational advantages?

RQ1 is designed to suggest whether free rides, where one representation of information needs to be translated into another to reveal facts that would otherwise have to be inferred, lead to cognitive advantages. RQ2 addresses the newer idea of an observational advantage, where there is no requirement for such a translation and, thus, no expectation that a user will be viewing multiple representations of information in different notations. Answering these questions is important for the design of visual modes of communication: if free rides and observational advantages yield demonstrable performance benefits then we should favour visualization methods that exhibit them as compared to competing representations. This paper specifically focuses on information about sets, visualized by linear diagrams, Venn diagrams, and Euler diagrams as seen in Fig. 1.

We now give a simple example. Suppose we have information about people who have visited various countries:

– Everyone who visited Denmark visited Germany

– No one visited both Germany and Mali
– Everyone who visited Oman visited Denmark
– Everyone who visited Uganda visited Denmark
– No one visited Uganda and Oman.

From these statements, which correspond to subset (*Everyone...*) and disjointness (*No one...*) relations between sets, various facts can be inferred which are not explicitly stated. These inferences include *Everyone who visited Uganda visited Germany* and *No one visited both Mali and Uganda.* By translating the originally given five textually-expressed facts into the Euler diagram in Fig. 1, we make these two inferred facts explicit and, so, they are examples of free rides from the diagram. This is because, in the first case, the translation necessarily places the Uganda circle inside the Germany circle. In the second case, the Mali and Uganda circles necessarily do not overlap. Indeed, this Euler diagram also makes *additional* derivable facts explicit, such as *No one visited both Mali and Oman*, and therefore has many free rides. All of these free rides are also observational advantages, due to this being a more general concept.

Whilst diagrams can make some information explicit through free rides, there is a lack of empirical evidence that using text and diagrams *in combination* improves task performance (RQ1). For instance, providing two representations of information may increase the time taken to perform tasks without bringing any accuracy improvements. The first study on which we report presents participants with either just text, or text in combination with a diagram, allowing us to suggest whether the addition of a diagram, obtained by translating the text, more effectively supports the identification of information which *must be inferred from the text* but *is a free ride in the diagram*. Our hypothesis, in line with the widely held view, is that free rides do bring significant performance benefits. The second study in this paper addresses RQ2, by presenting participants with either just text or just diagrams. Answering RQ2 will suggest whether observational advantages yield significant performance benefits.

Focusing on the visualizations of sets is of particular importance because there are enormous amounts of set-based data available in a wide variety of application areas (Alsallakh et al. 2014). Reflecting this abundance of data, the research community is actively devising numerous methods for visualizing it. Set visualization techniques often exploit closed curves (or variations thereof) (Collins et al. 2009; Chow and Ruskey 2005; Meulemans et al. 2013; Riche and Dwyer 2010; Simonetto et al. 2009) or lines for representing sets (Alper et al. 2011; Cheng 2011; Gottfried 2015; Rodgers et al. 2015). This paper therefore focuses on such methods by evaluating Venn diagrams and Euler diagrams both of which use closed curves (Stapleton et al. 2011; Wilkinson 2012; Venn 1880), see the middle and right of Fig. 1 for Venn and Euler diagrams, as well as linear diagrams (which use line segments) (Rodgers et al. 2015), see the left of Fig. 1.

In linear diagrams, if one set-line occupies only x-coordinates that another set-line occupies then this asserts a subset relationship. For instance, in Fig. 1 the set-line for Uganda occurs only where Germany also occurs: everyone who visited Uganda visited Germany. Non-overlapping lines corresponds to set disjointness so, here, one can see that Mali and Uganda are disjoint. By contrast, the Venn diagram in the middle uses shading to assert set emptiness: the region inside both (the curve for) Mali and

(the curve for) Uganda is entirely shaded, so the corresponding sets are disjoint. For subset relations, such as (informally) 'Uganda is subset of Germany', we see that the non-shaded region inside Uganda is entirely within Germany. Whilst Euler and Venn diagrams both exploit closed curves to represent sets, they have very different means of expressing information about those sets. In contrast with the use of an additional syntactic element – namely shading – in Venn diagrams, Euler diagrams use spatial relationships between curves. Our empirical studies will, therefore, test the role of free rides and observational advantages in explaining cognitive efficacy in three diagrammatic notations that convey information about sets in fundamentally different ways. This allows us to provide general insights, not results that are specific to one type of diagrammatic notation.

The paper is structured as follows. Section 2 covers free rides and observational advantages in the context of inference and the (potential) cognitive benefits of diagrams. We further illustrate free rides and observational advantages, in relation to observable information and meaning-carrying relationships in Sect. 3. Section 4 covers the methodology adopted for our study, including details of how the textual and diagrammatic stimuli were generated. It covers the tasks and training given to the participants, describes our data collection method and overviews the statistical methods employed. Section 5 presents our first study, evaluating the role of free rides, addressing RQ1. Section 6 covers our second study, evaluating the role of observational advantages, addressing RQ2. We conclude in Sect. 7. The supplementary material includes all of the experimental stimuli, the collected data, and the statistical models and output used for the analysis and is available from various urls that will be provided at appropriate places in the paper.

2 Free Rides, Observational Advantages and Inference

Free rides were introduced by Shimojima, with the idea dating back to his 1996 thesis (Shimojima 1996) so it is far from new. In his much more recent book, published in 2015, Shimojima states the following (Shimojima 2015):

> **Potential for Free Ride in Inference** Expressing a set of information in diagrams can result in the expression of other, consequential information. This enables us to skip certain mental deductive steps and to substitute them with the task of comprehending the consequences from the diagrams.

He goes on to say, when discussing the extraction of derived meanings from text:

> ... we usually go through a rather lengthy process of (i) interpreting each of the ... sentences in the text, (ii) integrating the individual pieces of information thus obtained, and (iii) drawing a conclusion appropriate to the question.

and contrasts this with the case of diagrams[1]:

> [information] can be derived ... given the various constraints holding on [the position of syntactic elements in] diagrams and the situations they represent

[1] Specifically, Shimojima is contrasting text and so-called position diagrams, but the point is a general one.

An entire chapter of his book is devoted to the *potential* for free rides in inference problems. In that chapter, he discusses the occurrence of free rides in both Venn and Euler diagrams, which we explain in the next section. At the heart of all the discussion is how a representation in textual form can be translated into a diagram to reveal 'hidden' information and the potential this has for facilitating inference. This insight is reflected in the first study in this paper, where participants will be presented with textual statements alongside a semantically equivalent diagram, obtained by translating the original text.

Shimojima claims that free rides are "advantageous for the purpose of making efficient inferences" (Shimojima 2015), where we take efficient to mean faster. He goes on to talk about the cost of inferences:

> A free ride saves one the cost of a deductive inference to [a] valid consequence, but not the cost of recognizing and interpreting the source type that is automatically realized in one's diagram.

Here, cost could be taken to mean either time savings or accuracy improvements. No evidence has yet been provided that free rides which occur in visualizations of sets do indeed yield significant performance improvements. This paper sets out to address this assumption by empirically evaluating free rides in a variety of diagrammatic representations of sets, as compared to natural language. In this empirical study, natural language forms the 'original' notation which is then translated into diagrammatic form.

Building on Shimojima's novel idea of a free ride, more recent work has generalised this notion to that of an observational advantage (Stapleton et al. 2017). In this generalised case, there is no requirement for a translation from one notation into another that then reveals facts that would otherwise need to be inferred. Instead, the concept of an observational advantage allows us to compare the 'advantages' of one representation of information over another, semantically equivalent representation: if one representation explicitly represents a fact that must be inferred from the other then that fact is an observational advantage. This paper sets out to address the assumption that observational advantages lead to cognitive advantages by empirically evaluating their occurence in a variety of diagrammatic representations of sets, as compared to natural language.

3 Free Rides and Observational Advantages in Linear, Venn and Euler Diagrams

We are focusing on visualizing information about sets, using three types of diagrams. Here, we informally illustrate free rides and observational advantages evident in these diagram types by comparing them to set-theoretic sentences such as $X \subseteq Y$ and $X \cap Y = \emptyset$ (although in our empirical studies, as seen in the introduction, constrained natural language expressions of the form *Everyone who visited X visited Y* and *No one visited both* X *and Y* are used for the purposes of accessibility).

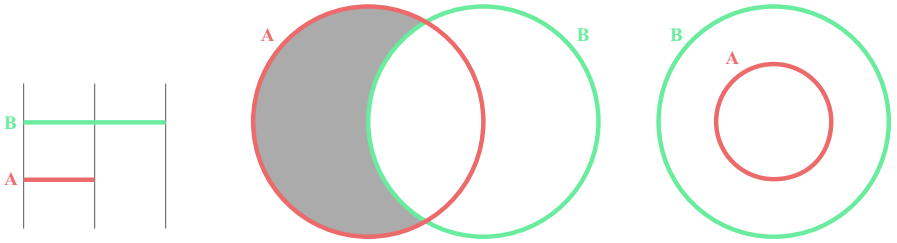


Fig. 2 Expressing $A \subseteq B$ using diagrams

3.1 Meaning Carrying Relationships

To understand free rides and observational advantages more precisely, we need to consider the idea of a meaning-carrying relationship. Taking a set-theoretic sentence such as $A \subseteq B$, there is a unique meaning-carrier: the symbol A is written to the left of $\subseteq$ and B to the right. It is from this meaning-carrying relationship that the sentence $A \subseteq B$ conveys the information that A is a subset of B. More precisely, a meaning carrying relationship is a relation on the syntactic items in a statement that carries semantics and evaluates to either true or false (Stapleton et al. 2017). In our example, either it is true that A is a subset of B or it is not. In general, set-theoretic sentences have unique meaning-carriers. By extension, our constrained natural language expressions are also taken to have one meaning carrier and the sentence either makes a true statement or a false one.

Our three diagram types can also express $A \subseteq B$, seen in Fig. 2. In the linear diagram (left), the set-line for A only occupies x-coordinates shared with the set-line for B: this meaning-carrying relationship expresses $A \subseteq B$. In the Venn diagram (middle), the non-shaded region inside A is entirely within B and in the Euler diagram the curve A is entirely inside the curve B: again, these meaning-carrying relationships express $A \subseteq B$.

Linear, Venn and Euler diagrams, unlike symbolic set-theory, typically have *multiple* meaning-carrying relationships. For instance, in Fig. 1, the relationship between any pair of curves in the Euler diagram is a meaning carrier. Here, the inclusion of the Oman circle inside the Germany circle is a meaning-carrier, as is the non-overlapping nature of Denmark and Mali. The meaning-carrying relationships evident in linear and Venn diagrams are, in fact, in a direct correspondence to those we see in Euler diagrams. For example, in Fig. 1, we can see that the line in the linear diagram for Oman shares all its x-coordinates with the line for Germany, asserting that Oman is a subset of Germany. In the Venn diagram, the equivalent meaning-carrier is a little more subtle, perhaps: the non-shaded region inside Oman is also inside Germany. For disjointness information, the linear diagram ensures that the lines for Denmark and Mali do not overlap whereas, in the Venn diagram, the region inside both curves is entirely shaded.

3.2 Observation

Now we have intuitively introduced the idea of a meaning-carrier, we can consider what it means to be able to *observe* information from a representation: given a representation of information, R, a statement that is directly obtained from a meaning-carrying relationship is *observable* from R (Stapleton et al. 2017). In the simple example seen in Fig. 2, we can observe $A \subseteq B$ from each of the three diagrams from their previously stated meaning-carriers. In Fig. 1, again from the diagrammatic meaning-carriers just given we can observe $Oman \subseteq Germany$ and $Denmark \cap Mali = \emptyset$.

3.3 Free Rides

Putting all this together, we can more precisely state what is meant by free ride, although we refer the reader to Stapleton et al. (2017) and Shimojima (2015) for more complete descriptions: a *free ride* from one representation of information, say R_1, given a semantically equivalent representation, R_2 that is derived by translating R_1, is a statement that is observable from R_1 but not from R_2[2]. Thus, the two statements just given, $Oman \subseteq Germany$ and $Denmark \cap Mali = \emptyset$, are examples of free rides from each of the diagrams[3] in Fig. 1 when they are presented alongside an alternative set-theoretic representation:

1. $Denmark \subseteq Germany$
2. $Germany \cap Mali = \emptyset$
3. $Oman \subseteq Denmark$
4. $Uganda \subseteq Denmark$
5. $Uganda \cap Oman = \emptyset$.

These five statements are the set-theoretic versions of the five textual statements given in introduction from which the three diagrams were derived. Thus, by extension, we have the fact that the textual statements

1. Everyone who visited Oman visited Germany, and
2. No one visited both Denmark and Mali

are free rides.

3.4 Observational Advantages

Using meaning carriers and observation, we are also able to give a more precise definition of an observational advantage: an *observational advantage* from one representation of information, say R_1, given a semantically equivalent representation, R_2, is a statement that is observable from R_1 but not from R_2. Thus, the two statements just

[2] The description of a free ride just given is in fact too weak but it is sufficient for understanding the contributions of this paper. There are stronger conditions that R_1 and R_2 must satisfy. As stated above, there is no requirement for a defined translation in the case of observational advantages. In addition, the stronger conditions placed on R_1 and R_2 for free rides do not arise in the case of observational advantages. In particular, observational advantages *only* require the semantic equivalence of R_1 and R_2.

[3] See (Takemura 2019) for a related account of free rides in Venn diagrams.

given, $Oman \subseteq Germany$ and $Denmark \cap Mali = \emptyset$, are examples of observational advantages from each of the diagrams in Fig. 1; there is no assumption that these diagrams were derived by translating the set-theoretic representation or, therefore, that they are presented alongside that representation.

3.5 Summary

The meaning-carrying relationships in linear, Venn and Euler diagrams are essentially equivalent, even though these diagram types use very different syntactic conventions to represent information. From meaning-carriers, we can identify free rides, and observational advantages, in the context of alternative representations of information. On this basis, we can readily compare equivalent linear, Venn and Euler diagrams to constrained natural language statements about sets, to determine whether free rides and observational advantages bring about the cognitive benefits alluded to by Shimojima's prior work.

4 Methods

To address our research questions, two empirical studies were conducted that measured task performance in terms of accuracy and time. This section describes the approach adopted to collect performance data, including the information being presented in textual form and visualized by diagrams, the tasks participants were asked to perform, the method used for data collection and the statistical methods employed for its analysis.

The first study presented participants with either textual statements or a diagram alongside the textual statements. For the second study, the diagrams were presented in isolation (so not in combination with text) and compared to the textual statements. In the studies, participants were asked to perform 20 tasks, the details of which are provided in what follows; these 20 tasks were presented in the *performance phase* of the study which was preceded by a *training phase*. Each task was a multiple choice question with five options, exactly one of which was the correct answer. Two options related to subset-style statements and two were disjointness-style statements. The fifth option was always 'none of the above'. For associated study materials, see https://www.cs.kent.ac.uk/people/staff/pjr/freerides/paper.html and https://www.cs.kent.ac.uk/people/staff/pjr/observationaladvantages/paper.html.

4.1 Generating Set Relationships and Corresponding Textual Stimuli

For the studies, we needed to generate textual statements that would be used as task stimuli and that would be translated into diagrams. It was essential that the textual statements yielded diagrams that exhibited free rides (and, therefore, observational advantages). Moreover, to avoid ceiling and floor effects, the information contained in the statements should require cognitive effort to understand without being overly complex. This meant that a reasonable number of sets needed to be used in the statements. For instance, using just three sets would lead to very few diagrams that exhibited free

Everyone who visited 1 visited 3
Everyone who visited 1 visited 4
Everyone who visited 4 visited 5
Everyone who visited 5 visited 3
Everyone who visited 2 visited 3

Fig. 3 Five randomly generated textual statements, prior to set names being assigned

Everyone who visited Denmark visited France
Everyone who visited Denmark visited Jamaica
Everyone who visited Jamaica visited Mali
Everyone who visited Mali visited France
Everyone who visited Vietnam visited France

Fig. 4 Five randomly generated textual statements, after set names were assigned

rides and observational advantages and the information would be simple to interpret. Informal experimentation suggested that using five sets led to controlled variability in the diagrams and that sufficient free rides could be generated. The first pilot study that we conducted supported our belief that using five sets would lead to cognitive effort being required by the participants, but without causing undue hinderance to performance. That is, there was no obvious ceiling or floor effect; descriptive statistics obtained from the pilot study data, in the first study, are given in Sect. 5.

To limit the complexity of the statements, we included information about subset and disjointness relationships between pairs of sets only. So, the textual statements were of the form:

1. Subset: *Everyone who visited A visited B*.
2. Disjointness: *No one visited both A and B*.

Each task was based on five such statements that were randomly generated in the order in which they were to be presented to participants. In each case, the statement type was randomly generated (i.e. subset or disjoint), then the sets A and B were randomly selected from the five sets involved in the task, at this point simply called sets 1 to 5; their names were determined later. This gave us information about five sets in the form of subset and disjointness relationships. An example of five randomly generated statements can be seen in Fig. 3. In this case, all statements concern subset relations.

Each collection of five statements was required to conform to the following characteristics:

1. the two sets A and B involved in a statement were never the same set; this ruled out trivially true assertions in the subset case and empty sets in the disjointness case,
2. if $A \subseteq B$ was asserted then $B \subseteq A$ was not asserted; this prevented sets being equal,
3. if $A \subseteq B$ or $B \subseteq A$ was asserted then $A \cap B = \emptyset$ was not asserted; this also prevented sets being empty,
4. for $1 < i \leq 5$, the i^{th} statement could never be inferred from the preceding statements; this meant each statement contained new information that was not already given by the statements generated (and written down in the question) before it.
5. the five statements had to give rise to at least one free ride and, therefore, observational advantage.

The first four requirements were to avoid the potential for confusion amongst participants. The last requirement was essential for the purposes of the study.

There were also requirements that had to be met by the 20 collections of five statements. Firstly, no two sets of five statements were isomorphic (i.e. the informational content was never the same up to the chosen set names). Including isomorphic sets of statements could impact task performance due to increased participant familiarity with the essential structure of the information conveyed. Further requirements arose because we also needed two *categories of task*: 10 tasks were about identifying subset statements were true (i.e. participants would need to identify that *Everyone who visited A visited B* was necessarily true) and the remaining 10 tasks were about disjointness statements (i.e. participants would need to identify that *No one visited both A and B* was necessarily true). Therefore, we needed 10 sets of statements that had a subset free ride and 10 sets of statements that had a disjointness free ride.

Once a set of 20 collections of five statements was generated that had the requisite properties, names were assigned to the sets. Country names were used, with no two names starting with the same letter. From a set of 26 such names, five were randomly selected for each task. They were allocated to sets 1 to 5, used in the statement generation, so that they first appeared in the five statements in alphabetical order. Figure 4 shows the result of assigning five randomly selected set names to the statements in Fig. 3.

4.2 Creating Diagrams

The 20 sets of statements generated for the studies were used to create linear, Venn and Euler diagrams. Where possible, the layout features were kept consistent across notations but due to their syntactic properties some differences are inherent. Linear diagrams were drawn with with straight line segments, Venn diagrams with ellipses, and Euler diagrams with circles. All three diagrammatic notations employed the following common layout features:

1. Colour: each set was assigned a unique colour, with a set of five colours generated by colorbrewer (Harrower and Brewer 2003). The colours were chosen to ensure they were visually distinguishable and suitable for categorical data. These colours were then used in the linear, Venn and Euler diagrams to colour the lines, ellipses, and circles respectively. It is known that the use of colour in this way improves the effectiveness of Euler diagrams (and therefore Venn diagrams) (Blake et al. 2016). For linear diagrams, using colour in this way does not significantly reduce (or improve) performance as compared to using monochrome (Rodgers et al. 2015). The five colours were assigned in a fixed order and then, for each set of statements, allocated to the sets in alphabetical order.
2. Font: each set name was written in times roman font, size 12, matching the textual statements. The name was assigned the same colour as its associated set and, thus, line, ellipse, or circle.
3. Line thickness: each line, ellipse and circle was 3.85 pixels wide.

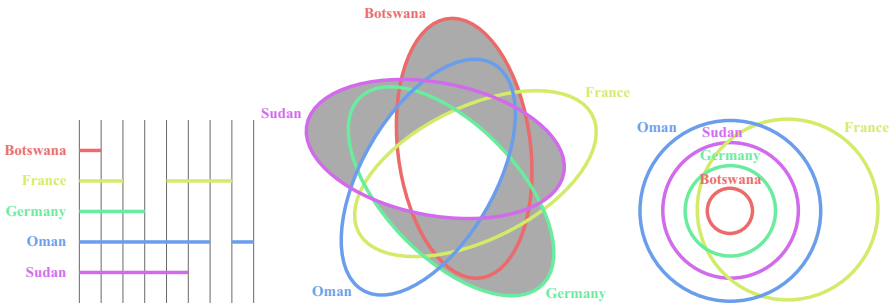


Fig. 5 Drawing diagrams for the studies

An example is given in Fig. 5. It can be seen that the colours assigned to each set are the same across notations, that the fonts match and take the same colours as their associated line or curve, and that the line thickness are the same.

4.2.1 Linear Diagram Layout Features

The linear diagrams were drawn using the layout guidelines in Rodgers et al. (2015). Each set-line was drawn horizontally with few line breaks. Vertical grid lines were used to mark the start and end of overlaps; an overlap corresponds to a particular set intersection, such as the rightmost overlap in the linear diagram of Fig. 1 which represents the set of people who visited Denmark, Germany and Uganda but not Mali or Oman. The grid lines can also be seen in Fig. 5. The sets were ordered alphabetically from top to bottom, which meant that the colours always appeared in the same top-to-bottom order, as is evident by comparing Figs. 1 and 5.

4.2.2 Venn Diagram Layout Features

Each Venn diagram comprised five ellipses and had a symmetric layout. A fixed shade of grey was used to indicate the emptiness of sets. The set names were assigned to the ellipses alphabetically in a clockwise direction starting from the top of the diagram. This meant, given that colours are assigned to set names in alphabetic order, that each Venn diagram used in the 20 performance phase questions differed from the others only by the regions which were shaded and the names of the sets, as is evident by comparing Figs. 1 and 5.

4.2.3 Euler Diagram Layout Features

Each Euler diagram was drawn using circles of a range of sizes. Circles are known to be a cognitively effective shape (Blake et al. 2016). Set names (labels) were positioned so that they were near the outside of their associated circle. Labels did not obfuscate each other. Where possible, the labels did not overlap with a circle. The regions formed by the circles did not have overly small areas.

4.3 Tasks and Training

As stated above, each question was multiple choice with five options, exactly one of which was correct. One option was 'None of the above'. The other four options always included two 'Everyone ...' statements and two 'No one ...' statements.

Ten of the 20 tasks required the identification of subset-style statements. The remaining ten tasks corresponded to disjointness-style statements. The 20 sets of five statements were thus divided into two sets of ten. In some cases, a set of five statements exhibited only subset free rides and observational advatnages ('Everyone ...' statements) and so were assigned to the 'subset' task type. Similarly, in other cases only disjointness free rides and observational advantages ('No one ...' statements) were present and so the five statements were assigned to the 'disjointness' task type. The remaining sets of five statements were randomly divided between the two categories whilst ensuring ten tasks in each.

Given the allocation of task types to sets of statements, we had to choose a statement to be the correct answer. In each case, one of the statements in the appropriate category was randomly selected. The sets involved in the other three options were randomly chosen whilst ensuring that the information in the associated option could not be inferred from the original five statements (i.e. the incorrect options were not *necessarily true*).

Having identified the correct answer and three incorrect options for each question (as well as a further incorrect option, namely 'None of the above'), we paid particular attention to the order in which the options were presented. To control the variability between subset and disjointness tasks, each task type had 2 correct answers as option one, 3 correct answers as option two, 3 correct answers as option three, and 2 correct answers as option four. The remaining four (incorrect) options were randomly ordered around the correct answer except that 'None of the above' always appeared last. A screenshot from the first study, addressing RQ1, can be seen in Fig. 6 where the correct answer is option 1.

In order for participants to be able to perform the tasks, initial training was provided. This comprised a series of four tasks. The first task used just three sets and the correct answer was a disjointness-style statement. The second training task used four sets and was in the subset category. The final two training tasks used five sets, making them similar to the tasks used in the study, one for subset and one for disjointess. The screenshot in Fig. 7 shows how a training question was presented in the first study, using text and an Euler diagram representation, in the case of three sets. Figure 8 shows the corresponding explanation given after the participant submitted their answer. Training was similar for the other groups in the first study, differing due to the nature of the syntax in the representation. For the second study, the training removed reference to the textual statements when participants were exposed to one of the diagrammatic treatments, as in Figs. 9 and 10. The 20 performance phase questions were presented similarly to Fig. 6. The training material and all of the performance material can be found in the supplementary files.

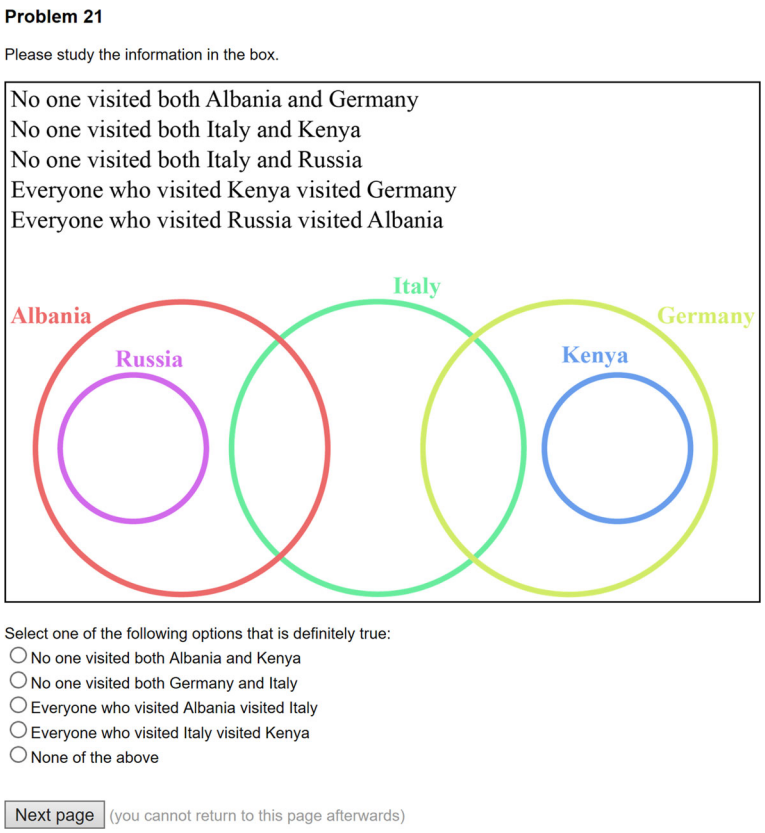


Fig. 6 Euler diagram question: the correct answer is option 1

4.4 Data Collection Method

Both studies adopted a between group study design with four groups. The first study, for RQ1, included: text-only, linear diagrams with text, Venn diagrams with text, and Euler diagrams with text. The second study, which ran at a later time, for RQ2 included: text-only, linear diagrams, Venn diagrams and Euler diagrams. For each study pre-screening was used, described in the relevant sections below.

Participants were randomly assigned to one group. Prolific Academic was used to crowdsource participants from the general population. It is recognised that in some crowdsourced studies, participants do not always give questions their full attention, or have difficulties with the language used, and this is hard to control (Chen et al. 2011). We call such participants *inattentive*. A common technique to identify inattentive participants is to include questions that are trivial to answer. An example, from the linear diagrams with text group, is in Fig. 11. It can be seen that the participant was told it was an attention checking question, the first four options used country names that did not appear in the textual statements or diagram, and the last option instructed them to choose that one. The presentation for the questions designed to identify inattentive

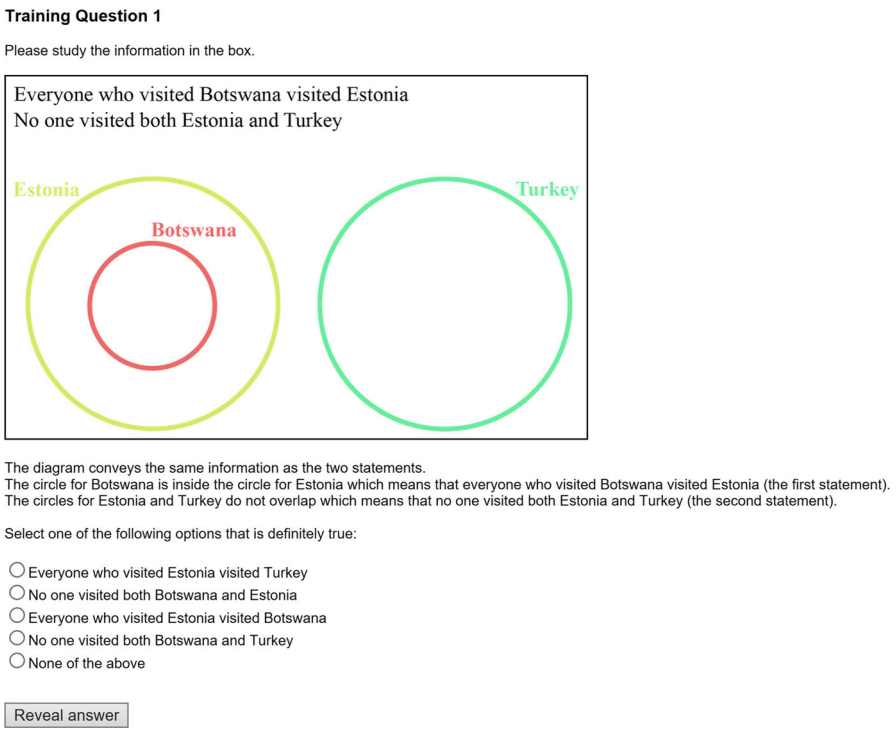


Fig. 7 Euler diagram with text: training, initial task presentation

participants was different in the second study, with details given later. Our studies each included two attention checking questions, for each group, and they always appeared as the 7th and 14th questions in the performance phase. A participant was classified as inattentive if they answered either of these two questions incorrectly; their data were not analyzed.

In both the training phase and the performance phase, each question was displayed on a separate page. Participants could not return to pages and subsequent pages were not revealed until the previous answer was submitted. Unlike the training questions, which were presented in the same order for all participants, in the performance phase the questions were randomly ordered. Participants were instructed to maintain their concentration on the study and to answer questions without delay, unless a question explicitly said otherwise (these were the attention checking questions). The full information provided to participants is given in the supplementary material.

4.5 Statistical Methods

Recall that we are collecting accuracy and time data as indicators of performance. Accuracy is viewed as more important than time: one representation of information is judged to be more effective than another if users can perform tasks significantly more accurately with it or, if no significant accuracy difference exists, performance is sig-

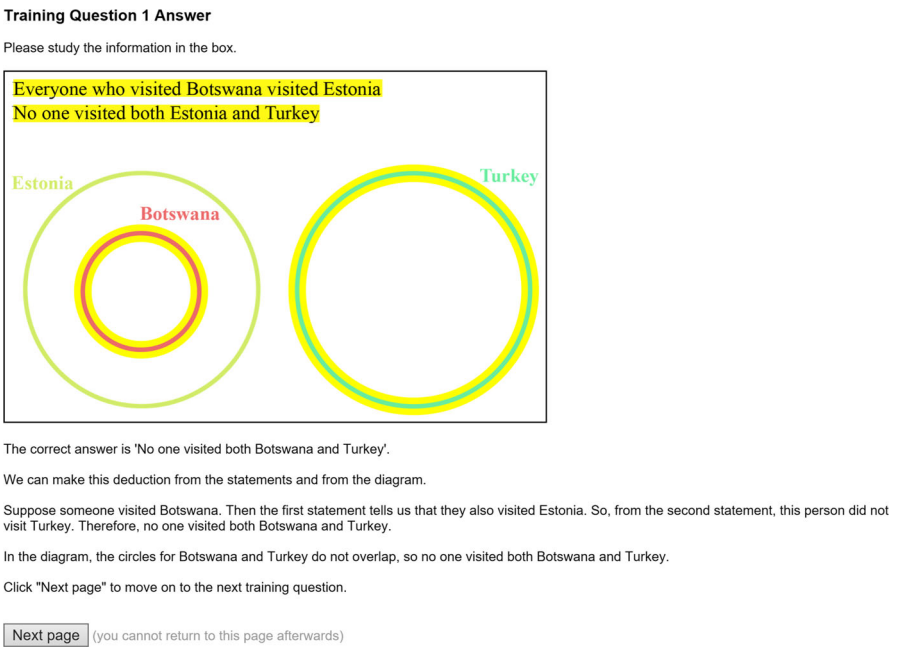


Fig. 8 Euler diagram with text: training, explanation

nificantly quicker. For each study, we employed two generalized estimating equations models (Liang and Zeger 1986) to analyse the *accuracy* data. An ANOVA calculation was not appropriate as the data violated the normality assumption. The non-parametric version of ANOVA, Kruskal-Wallis, was also not appropriate, as the responses for each individual are correlated, and so not independent.

For the *time* data, for each study we used two generalized estimation models (Liang and Zeger 1986) that allowed us to estimate whether the time taken to provide answers was significantly different. Again, alternative statistical tests, such as ANOVA, were not appropriate as assumptions were violated by the data.

Full details of the models and the statistical output can be found in the Supplementary Material. Whilst we view accuracy as the most important indicator of performance differences, all analysis that was performed is reported in the paper. Throughout, results are declared significant if $p \leq 0.05$.

5 Free Rides Study: Diagrams as a Support for Text

This section addresses RQ1, which is broken down in to the following more specific questions:

RQ1a: Do the free rides exhibited by the considered diagram types, when presented in combination with text, bring about significant task performance benefits over text alone?

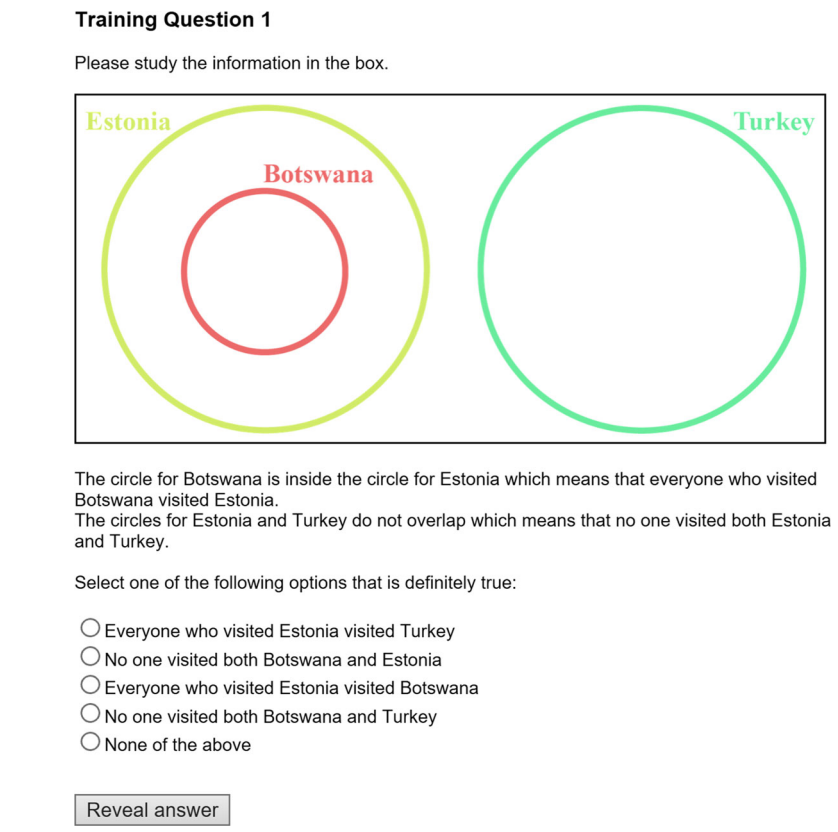


Fig. 9 Euler diagram: training, initial task presentation

RQ1b: Do the free rides exhibited by any one diagram type lead to significant task performance benefits over the other diagram types?
RQ1c: Do tasks concerning subsets lead to significantly different task performance compared to those concerning disjointness?

In reporting on our data collection and results, we will refer to the treatments as

- T for 'text only',
- L&T for linear diagrams and text,
- V&T for Venn diagrams and text, and
- E&T for Euler diagrams and text.

The online version of this study, where it is possible to select the group in which to take part (unlike the actual study where participants were randomly assigned to a group) can be found here: https://www.cs.kent.ac.uk/people/staff/pjr/freerides/. An example of the stimuli for one study task can be seen in Figs. 12, 13, 14 and 15, with the associated options in Fig. 16. The correct answer is option 3, so this is a disjointness-style task.

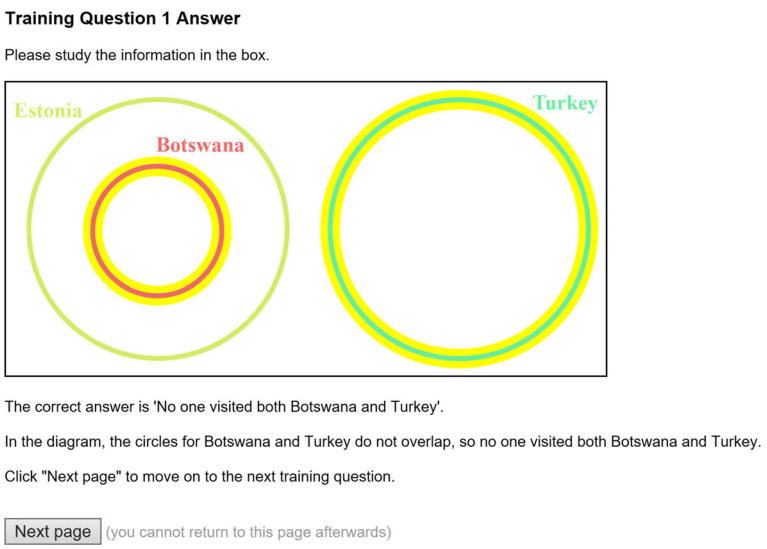


Fig. 10 Euler diagram: training, explanation

When running a pilot study[4], we pre-screened participants using the following criteria:

1. they had to have a Prolific approval rate of 97% or higher, and
2. they had to have completed at least 5 studies on Prolific previously.

This left a pool of 25, 313 potential participants, out of 58, 462, so over half were disqualified. Each participant was allowed a maximum of 45 minutes to complete the study, with an expected completion time of 20 minutes, and was randomly allocated to one of the four groups. They were each paid £2.61, reflecting the time we expected it to take to complete the study.

A total of 41 people began the pilot study. Of these, five were classified as inattentive, two timed-out after 45 minutes and a further four (all in the Venn group) withdrew before completion. This left data from 30 participants. Prolific indicated that it took participants on average 23 minutes to complete the study which includes all time spent on training, performing the tasks and supplying demographic information. For the pilot, the overall accuracy rate and the average (mean) time in seconds to answer each question (in seconds) are given in Table 1, alongside a breakdown for each group; the low number of participants in the Venn group is likely due to a combination of the random allocation and the four withdrawals. The overall data do not indicate a ceiling or floor effect and the range of accuracy rates and mean times suggest that there may be differences across the groups.

Whilst no problems were identified in the study materials presented to the participants, the average time taken to complete the entire study was longer than the expected

[4] After running an initial pilot and main study, errors were found in the study materials, primarily for the text-only treatment. As such, the errors were rectified and this paper reports, therefore, on a second pilot and second main study. No participant who took part in the flawed initial studies was allowed to take part in the subsequent studies reported on here.

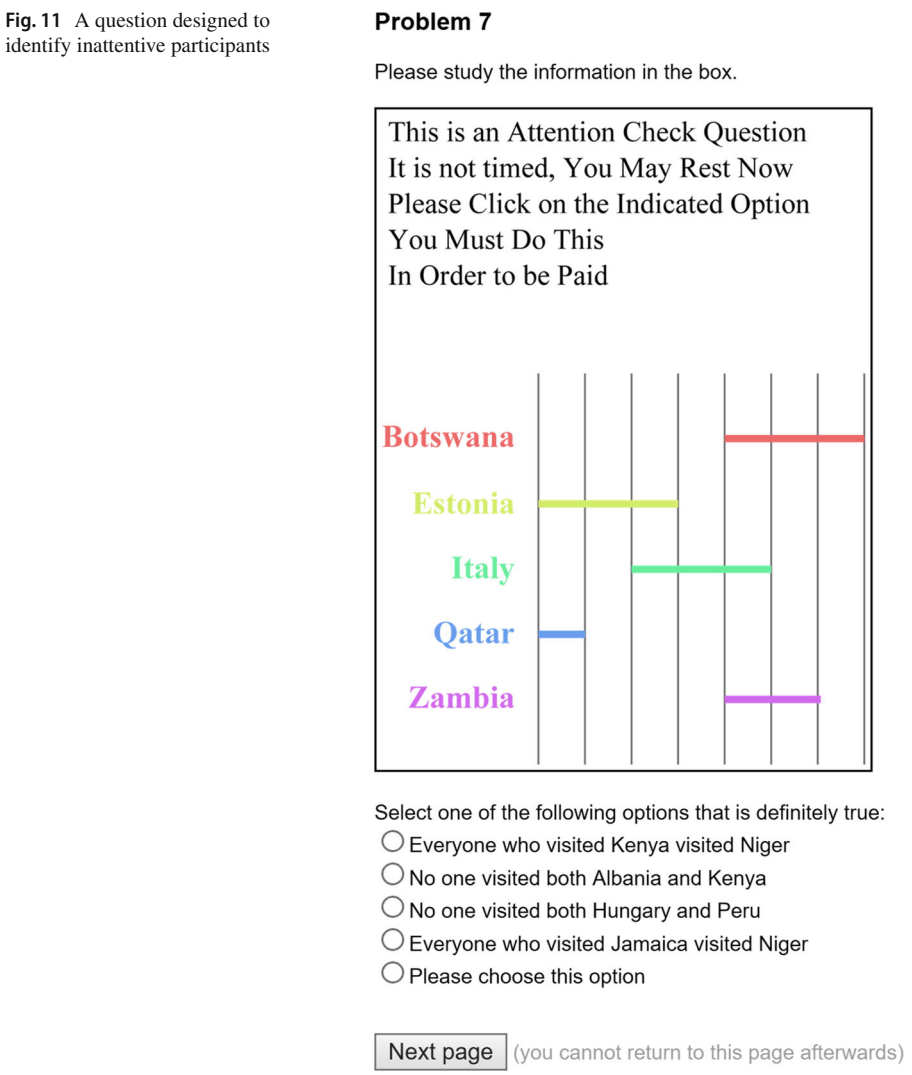


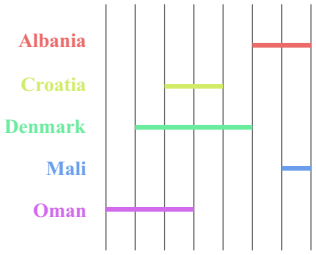
Fig. 11 A question designed to identify inattentive participants

Fig. 12 T group stimulus

No one visited both Albania and Croatia
No one visited both Albania and Denmark
No one visited both Albania and Oman
Everyone who visited Croatia visited Denmark
Everyone who visited Mali visited Albania

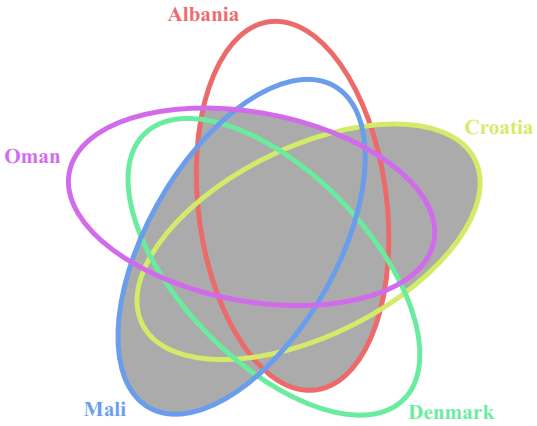
20 minutes and two people timed out after 45 minutes. We revised our cap of 45 minutes to be 60 minutes and also the expected time was increased from 20 minutes to 30 minutes. In light of these changes, we also increased payment for taking part to £3.92. In addition, due to the number of people failing to complete the study (either by being inattentive, timing out, or withdrawing) we strengthened the pre-screening criteria:

1. participants had to have a Prolific approval rate of 98% or higher,



No one visited both Albania and Croatia
No one visited both Albania and Denmark
No one visited both Albania and Oman
Everyone who visited Croatia visited Denmark
Everyone who visited Mali visited Albania

Fig. 13 L&T group stimulus



No one visited both Albania and Croatia
No one visited both Albania and Denmark
No one visited both Albania and Oman
Everyone who visited Croatia visited Denmark
Everyone who visited Mali visited Albania

Fig. 14 V&T group stimulus

Table 1 Summary of the pilot data

Group	No. of participants	Accuracy (%)	Mean time
Overall	30	84.67	43.04
T	9	61.67	49.91
L&T	6	94.17	41.40
V&T	1	100	68.14
E&T	14	94.29	37.53

No one visited both Albania and Croatia
No one visited both Albania and Denmark
No one visited both Albania and Oman
Everyone who visited Croatia visited Denmark
Everyone who visited Mali visited Albania

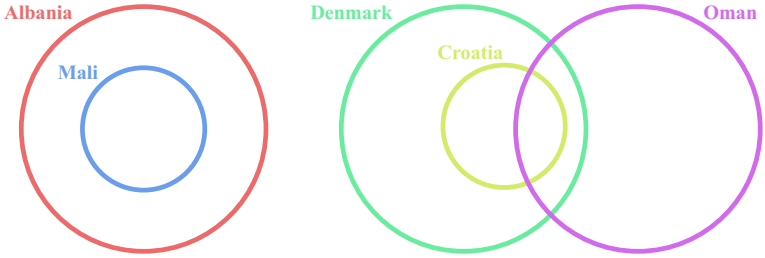


Fig. 15 E&T group stimulus

Fig. 16 Options for the task

Everyone who visited Croatia visited Oman
No one visited both Croatia and Denmark
No one visited both Mali and Oman
Everyone who visited Denmark visited Oman

2. they had to have completed at least 5 studies on Prolific previously (as in the pilot),
3. no one who participated in the pilot could take part, and
4. English had to be the first language.

Using these criteria, there were 20270 eligible Prolific users. We also indicated that the study was compatible with desktop devices. Lastly, we identified a few minor bugs in the way in which the collected data were recorded which were rectified.

A total of 554 people began the main study. Of these, 16 were classified as inattentive and 125 were recorded as starting the study but did not complete it (either through withdrawing or timing out[5]). There were also 9 participants that Prolific recorded as completing the study but for whom we could not obtain their data. These participants were paid. This left data from 404 participants. The breakdown of these data across the groups is given in Table 2. The 404 participants had the following demographic profile: age: mean 34, range from 18 to 69; gender: 167 M , 234 F, 3 other; the distribution was similar across the four groups.

It is particularly striking that there were a large number of non-completions from the V&T group which could impact the validity of the results for this group. Of the 150 participants who failed to complete the study, at least 61 (40%) were in the V&T group (we had no group information for 13 of the participants who did not complete). Whilst we can only speculate on why so many people failed to complete the V&T tasks, one reason could be that they perceived the tasks to be difficult and chose not to proceed. Thus, the performance data we have for this group may suggest an artificially high accuracy rate and an artificially low time performance. Given this, the

[5] Note that there are various reasons why participants may time-out, such as finding the tasks time consuming or essentially withdrawing through failure to proceed through the tasks. Therefore, from this point, we report on non-completions which includes time-outs and withdrawals.

Table 2 Free rides study: summary of the participant distribution for the main study

Group	Completions	Inattentive	Non-completions	No data
Overall	404	16	125	9
T	99	6	28	2
L&T	99	1	19	1
V&T	103	6	54	1
E&T	103	3	11	5
Unknown	–	–	13	–

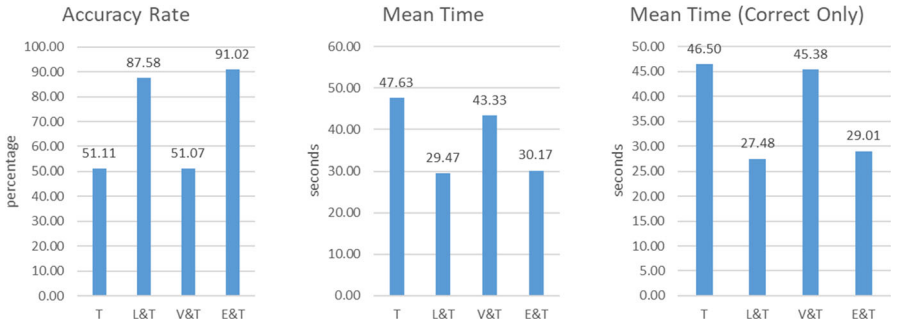


Fig. 17 Free rides study: visualizing the accuracy rate and mean times

actual performance for V&T may be substantially worse than our reported analysis suggests.

5.1 Comparison of Representations

Here we report on the overall comparison between the four treatment groups. The data and associated statistical output can be found at https://www.cs.kent.ac.uk/people/staff/pjr/freerides/paper.html. The accuracy rates and mean times (in seconds) are summarised in Table 3; the final column indicates the mean time taken to provide a correct answer, thus excluding the data where incorrect answers were provided. Whilst the accuracy rates and mean times are in *indicator* of relative performance, it is important to note that the statistical methods employed do not compare these data: the statistical methods that compare means (e.g. ANOVA) do not account for correlated responses from participants and make other assumptions that our data violate. For this reason, the standard graphical plots visualising means and related features of the data, given in Fig. 17, can be misleading when interpreting the results of our statistical analysis. However, they provide a useful insight into the differences between the performances of each group.

Using a GEE based statistical model for the accuracy data, we estimated a 95% confidence interval (CI) for the odds of providing a correct answer with one treatment compared to another. We also computed p-values to determine whether the treatments gave rise to significantly different accuracy performance. The estimated odds of cor-

Table 3 Free rides study: summary of the main study data

Group	No. of participants	Accuracy (%)	Mean time	Mean time (correct only)
Overall	404	70.21	37.63	34.70
T	99	51.11	47.63	46.50
L&T	99	87.58	29.47	27.48
V&T	103	51.07	43.33	45.38
E&T	103	91.02	30.17	29.01

Table 4 Free rides study: overall comparison of treatments by accuracy

Treatments	Odds	CI	p-value	Sig.	Most accurate
L&T versus T	6.74	(4.53, 10.04)	< 0.0001	✓	L&T
L&T versus V&T	6.75	(4.37, 10.44)	< 0.0001	✓	L&T
L&T versus E&T	0.70	(0.42, 1.15)	0.1537	×	–
T versus V&T	1.00	(0.75, 1.33)	0.9906	×	–
T versus E&T	0.10	(0.07, 0.15)	< 0.0001	✓	E&T
V&T versus E&T	0.10	(0.07, 0.16)	< 0.0001	✓	E&T

rectly answering questions with L&T was 6.74 (to 2d.p.) times higher than that of T with a 95% CI of (4.53,10.04) and p-value of < 0.0001 (to 4d.p.). Therefore, L&T supported significantly more accurate task performance than T only. Results for the other pairwise comparisons are given in Table 4, from which we can derive an overall ranking for the treatments:

$$\textit{Accuracy overall ranking: } L\&T = E\&T > V\&T = T.$$

Using a GEE based statistical model for the time data, we estimated a 95% CI for the ratio of the time (measured in seconds) needed to answer a question correctly with one treatment compared to another. The CI and its corresponding p-value allowed us to determine whether two treatments were significantly different. The model estimated that the time needed to answer a question correctly with L&T was 0.62 times (2d.p.) that with T with a 95% CI of (0.55, 0.69) and p-value of < 0.0001. Therefore, linear diagrams with text supported significantly faster task performance than text only. Results for the other pairwise comparisons are given in Table 5, from which we can derive an overall ranking for the treatments:

$$\textit{Time overall ranking: } L\&T = E\&T > V\&T = T.$$

Therefore, our accuracy and time analysis consistently support the superiority of linear diagrams and Euler diagrams, when used as a support for a textual representation, as compared to Venn diagrams alongside text or just text alone. Taking into account both the accuracy and time analysis, we have consistent rankings, from which we

Table 5 Free rides study: overall comparison of treatments by time

Treatments	Ratio	CI	p-value	Sig.	Fastest
L&T versus T	0.62	(0.55, 0.69)	< 0.0001	✓	L&T
L&T versus V&T	0.63	(0.56, 0.70)	< 0.0001	✓	L&T
L&T versus E&T	0.95	(0.85, 1.06)	0.3492	×	–
T versus V&T	1.01	(0.89, 1.15)	0.8373	×	–
T versus E&T	1.54	(1.36, 1.74)	< 0.0001	✓	E&T
V&T versus E&T	1.52	(1.34, 1.71)	< 0.0001	✓	E&T

derive an overall ranking of the four treatments:

overall ranking: L&T = E&T > V&T= T.

The odds and ratios computed for the accuracy and, respectively, time data give insight into the effect size. For instance, the odds of producing a correct answer using linear diagrams alongside text compared to text alone are 6.74, which is practically significant[6] (odds of approximately 1 would indicate no significant difference[7]). From the perspective of time, we would expect correctly answering a question using linear diagrams alongside text to be 0.62 of the time taken (i.e. to take 62% of the time) to provide a correct answer using text alone. Again, this is a practically significant effect size. The other effect sizes, where we saw significant differences, are similar and are evident from the odds and ratios given in the tables. To gain further insight into our data, we will now perform an analysis to compare the treatments for the two different task types.

5.1.1 Subset Comparison Across Treatments

For the subset tasks, the indicative accuracy rates and mean times are in Table 6. As with the overall analysis, we produced a GEE based statistical model for the accuracy data. Results for the pairwise comparisons are given in Table 7, from which we can derive a ranking for the treatments for subset tasks:

$$\textit{Accuracy subset-task ranking: } L\&T = E\&T > T > V\&T.$$

A GEE based statistical model for the time data yielded the results given in Table 8, from which we can derive a ranking for the treatments for the subset tasks:

$$\textit{Time subset-task ranking: } L\&T = E\&T > T = V\&T.$$

[6] This is a subjective interpretation of the result, taking into account the fact that the difference between treatments is significant and the effect size is large. In our view, a large odds ratio such as 6.74 suggests that, in practice, the accuracy improvements brought about by L&T over T are of practical use. By contrast, had the odds ratio been, say, 1.03 then even if there was a significant difference there would not, in practice, be much benefit of using L&T over T from the perspective of accuracy.

[7] Strictly speaking, a confidence interval containing 1 would indicate no significant difference.

Table 6 Free rides study: summary of the main study data for subset tasks

Group	No. of participants	Accuracy (%)	Mean time	Mean time (correct only)
Overall	404	72.53	36.34	34.36
T	99	60.40	42.60	42.84
L&T	99	87.78	28.86	27.25
V&T	103	50.29	43.52	45.40
E&T	103	91.75	30.35	29.46

Table 7 Free rides study: subset comparison of treatments by accuracy

Treatments	Ratio	CI	*p*-value	Sig.	Most accurate
L&T versus T	4.71	(3.03, 7.32)	< 0.0001	✓	L&T
L&T versus V&T	7.09	(4.24, 11.39)	< 0.0001	✓	L&T
L&T versus E&T	0.65	(.038, 1.11)	0.1147	×	–
T versus V&T	1.51	(1.07, 2.12)	0.0180	✓	T
T versus E&T	0.14	(0.09, 0.21)	< 0.0001	✓	E&T
V&T versus E&T	0.09	(0.06, 0.14)	< 0.0001	✓	E&T

Viewing accuracy as the more important indicator of relative performance, we thus obtain the following ranking of treatments for subset tasks:

subset-task ranking: L&T = E&T > T > V&T.

It is notable that using text alone supported significantly more accurate performance than using V&T. As with the overall analysis, the odds and ratios computed for the accuracy and, respectively, time data give insight into the effect size. For the most interesting result here, that text alone was superior to Venn diagrams with text, the odds of providing a correct answer with T are 1.51 that of providing a correct answer with V&T; there was no significant time difference between these two treatments and, so, no time effect size to discuss. Whilst 1.51 is not considered to be a particularly large effect size, it is nevertheless unexpected. Here, the use of Venn diagrams, which were intended to act as a support for the text, have actually damaged participant performance.

5.1.2 Disjointness Comparison Across Treatments

For disjointness tasks, a GEE based statistical model for the accuracy data yielded the results given in Tables 9 and 10 and a GEE based model for the time data yielded the results in Table 11. We obtain these two performance rankings:

$$\textit{Accuracy disjoint-task ranking: } L\&T = E\&T > V\&T > T.$$
$$\textit{Time disjoint-task ranking: } L\&T = E\&T > V\&T = T.$$

Table 8 Free rides study: subset comparison of treatments by time

Treatments	Ratio	CI	p-value	Sig.	Fastest
L&T versus T	0.66	(0.58, 0.74)	< 0.0001	✓	L&T
L&T versus V&T	0.61	(0.54, 0.69)	< 0.0001	✓	L&T
L&T versus E&T	0.93	(0.82, 1.07)	0.1896	×	–
T versus V&T	0.94	(0.82, 1.07)	0.357	×	–
T versus E&T	1.42	(1.25, 1.61)	< 0.0001	✓	E&T
V&T versus E&T	1.51	(1.33, 1.71)	< 0.0001	✓	E&T

Table 9 Free rides study: summary of the main study data for disjointness tasks

Group	No. of participants	Accuracy (%)	Mean time	Mean time (correct only)
Overall	404	67.90	38.92	35.07
T	99	41.82	52.13	51.80
L&T	99	87.37	29.78	27.71
V&T	103	51.84	43.13	45.35
E&T	103	90.29	29.99	28.55

Table 10 Free rides study: disjointness comparison of treatments by accuracy

Treatments	Odds	CI	p-value	Sig.	Most accurate
L&T versus T	9.63	(6.12, 15.14)	< 0.0001	✓	L&T
L&T versus V&T	6.43	(3.93, 10.50)	< 0.0001	✓	L&T
L&T versus E&T	0.74	(0.42, 1.31)	0.3068	×	–
T versus V&T	0.67	(0.48, 0.92)	0.0139	✓	V&T
T versus E&T	0.08	(0.05, 0.12)	< 0.0001	✓	E&T
V&T versus E&T	0.12	(0.07, 0.19)	< 0.0001	✓	E&T

Table 11 Free rides study: disjointness comparison of treatments by time

Treatments	Ratio	CI	p-value	Sig.	Fastest
L&T versus T	0.57	(0.50, 0.65)	< 0.0001	✓	L&T
L&T versus V&T	0.64	(0.56, 0.72)	< 0.0001	✓	L&T
L&T versus E&T	0.97	(0.87, 1.09)	0.6174	×	–
T versus V&T	1.12	(0.97, 1.30)	0.1161	×	–
T versus E&T	1.71	(1.49, 1.96)	< 0.0001	✓	E&T
V&T versus E&T	1.52	(1.34, 1.73)	< 0.0001	✓	E&T

Viewing accuracy as the more important indicator of relative performance, we thus obtain the following ranking of treatments for disjoint tasks:

disjoint-task ranking: L&T = E&T > V&T> T.

Unlike for subset tasks, here we saw more accurate performance using V&T than T. For this result, we found that the odds of providing a correct answer with T are 0.67 that of providing a correct answer with V&T; there was no significant time difference between these two treatments and, so, no time effect size to discuss. This suggests that the odds of providing a correct answer with V&T are 1.49 that of T, which is close to the effect size of T over V&T in the subset case. Thus, unlike for subset tasks, here Venn diagrams have acted as a support for the textual representation but with a relatively small effect size.

5.1.3 Summary

The overall analysis and that for the two task types allows us to suggest answers to RQ1a and RQ1b. Concerning RQ1a, our study suggests that Euler and linear diagrams, when used as a support for text, consistently supported significantly better task performance than using text alone. By contrast, using Venn diagrams brought about no significant performance benefit overall. By performing a task-level analysis, we revealed that presenting participants with Venn diagrams alongside text assisted with disjointness tasks but led to significantly worse performance for subset tasks: Venn diagrams are damaging to users' task performance in this case. Thus, in answer to RQ1a, our data support the efficacy of Euler and linear diagrams but shed doubt over the efficacy of Venn diagrams, at least for the standard set-theoretic tasks included in our study. In particular, these results call into question the adequacy of the theory of free rides as a stand-alone predictor of cognitive benefit when solving inference problems.

Focusing on RQ1b, we have found evidence that different diagram types, yet with precisely the same free rides, can give rise to significantly different performance benefits. Euler and linear diagrams consistently outperformed Venn diagrams both overall and for the two task types.

5.2 Comparison of Task Types with Treatments

Here we address RQ1c, which asked whether the two types of task yield significantly different performances. We do this by considering each treatment in turn, with the statistical results given in Tables 12 and 13. In summary, for accuracy, there was no difference between the task types for L&T, E&T and V&T but for T the subset tasks were performed significantly more accurately. Regarding time, there was no difference between the task types for L&T and V&T. With T, subset tasks were performed significantly faster. This suggests a tendency for subset tasks to be easier than disjointness tasks, when using just Text. For E&T, disjointness tasks were performed significantly faster. Hence, in answer to RQ1c, our study suggests that the task type can have a significant impact on task performance. Whilst free rides have the potential for facilitating

Table 12 Free rides study: comparison of subset versus disjoint tasks by treatment w.r.t. accuracy

Treatments	Odds	CI	p-value	Sig.	Most accurate
T	2.12	(1.72, 2.62)	< 0.0001	✓	Subset
L&T	1.04	(0.74, 1.45)	0.8287	×	–
E&T	1.20	(0.84, 1.70)	0.3224	×	–
V&T	0.94	(0.74, 1.19)	0.6096	×	–

Table 13 Free rides study: comparison of subset versus disjointness tasks by treatment w.r.t. time

Treatments	Ratio	CI	p-value	Sig.	Fastest
T	0.87	(0.79, 0.95)	0.0024	✓	Subset
L&T	1.00	(0.96, 1.05)	0.9151	×	–
E&T	1.05	(1.01, 1.09)	0.0179	✓	Disjoint
V&T	1.04	(0.97, 1.11)	0.2435	×	–

inference, as argued by Shimojima, this study illustrates the importance of having a deeper understanding of the role free rides play and their cognitive benefits: different types of free rides within a notation (e.g. Euler diagrams) can lead to differences in performance.

5.3 Free Rides Study: Interpretation

Here we seek to explain why Euler and linear diagrams acted as a support for the textual representation of information whereas the Venn diagrams were sometimes damaging. We must be mindful of the caveat that the V&T data may present an overly supportive view of that treatment. Given the high number of non-completions from the V&T group, we examined the comments provided by the participants. Of the 103 who completed the V&T questions, 32 made comments that indicated the study was hard, with one person saying "[the Venn] diagrams made things much more difficult, had to focus on descriptions." In the T group, 18 out of the 99 participants commented on the study being difficult, whereas in the E&T group only 5 out of 103 participants made comments that could be taken to imply it was difficult (e.g. "harder than expected" and "had to think"). In the L&T group, 3 out of 99 participants made comments indicating difficultly. By contrast, the E&T and L&T groups were slightly more pronounced in indicating that the study was fun (11 in E&T, 12 in L&T) compared to 8 in each of the V&T and T groups. These data support the speculation that people withdrew from the V&T group because they perceived the tasks as difficult.

We now consider the nature of making inferences from the textual representation, when there is no diagrammatic support. Suppose we wish to infer the subset-style statement *Everyone who visited Uganda visited Germany* from the following:

Everyone who visited Denmark visited Germany
No one visited both Germany and Mali

Everyone who visited Oman visited Denmark
Everyone who visited Uganda visited Denmark
No one visited both Uganda and Oman

We need to identify the two statements – *Everyone who visited Denmark visited Germany* and *Everyone who visited Uganda visited Denmark* – from which *Everyone who visited Uganda visited Germany* follows and then use the transitivity of subset to make the deduction. All of the subset-style tasks are of this nature, although they may involve more than two statements from which the deduction must be made.

By contrast, reasoning about disjointness may be conceptually more difficult. For instance, to infer *No one visited both Mali and Uganda* from the textual representation just given, both subset and disjointness statements are required:

1. Everyone who visited Uganda visited Denmark.
2. Everyone who visited Denmark visited Germany.
3. *Therefore,* everyone who visited Uganda visited Germany.
4. No one visited both Germany and Mali.
5. *Therefore,* no one visited both Mali and Uganda

Whilst this example needs three of the original statements to make the deduction, other disjointness-style tasks only require one subset and one disjointness statement.

In general a 'No one...' deduction requires the human reasoner to understand both kinds of statement and how to make inferences from them. By contrast 'Everyone ... ' statements are deduced using just the transitivity of the subset relation. This leads us to:

Speculation 1: using the transitivity of subset and a single statement type (i.e. subset) when performing subset tasks using textual statements is cognitively easier than performing disjoint tasks, which require both statement types (i.e. subset and disjoint).

Hence, speculation 1 could be a reason why participants performed subset tasks significantly more accurately and significantly faster than disjointness tasks when using just the textual representation.

Why, then, were Venn diagrams a hinderance in the case of subset tasks but an assistance in the case of disjointness tasks? Our analysis for RQ1c supports the cognitive similarity of these tasks when undertaken by the V&T group: they do not perform significantly better or worse for either task type. Thus, the difference in performance between the V&T group and the T group is posited to be due to the increased difficulty of disjointness reasoning in the case of text.

This alone, though, is not sufficient to explain the difference in performance, since the V&T group had access to the textual representation as well as the diagram. Participants could use the Venn diagram to reinforce their answer to the question based on their understanding of the text, use the diagram without the text, or not use the diagram at all. We found no difference in time performance between V&T and T for either task type, which could suggest the diagrams were not used predominantly to reinforce understanding (we anticipate that a substantial reliance on both representations would lead to an increase in time taken). If participants did not make significant use of the

Venn diagrams then we would not expect them to reduce accuracy performance as compared to Text. Thus, the subset-style task data suggest that participants did utilise the Venn diagrams to answer the questions. To perform a subset-style task, they would need to identify the non-shaded region inside the 'subset' curve and determine whether this region was inside the 'superset' curve.

To illustrate, in Fig. 1, to deduce that *Everyone who visisted Uganda visited Germany*, participants would need to identify the non-shaded region inside Uganda and then establish that this was inside the Germany curve. The task of identifying disjointness is similar: to deduce *No one visited both Mali and Uganda*, one must find the region inside both curves and determine whether it is shaded. The theory of figure-ground segregation (Wagemans et al. 2012) suggests that the ellipses themselves will be readily identifiable, but not the regions formed by intersecting ellipse interiors. Our data suggest that, since the region inside any given ellipse is subdivided into 16 smaller regions by the other four ellipses, computing the appropriate non-shaded region necessary for subset tasks is cognitively more difficult than reasoning about the transitivity of subset relations in the case of text, causing errors to arise. This reinforces the notion that participants were reliant on the Venn diagrams at the expense of not utilizing the textual statements and gives rise to:

Speculation 2: identifying a non-shaded region formed from multiple intersecting ellipses in a Venn diagram is cognitively harder than using subset transitivity in textual statements when performing subset tasks.

By contrast, our data suggest that there is an increased difficulty of the reasoning involved in disjointness tasks, when using just text as compared to Venn diagrams alongside text. In turn, this suggests that this style of reasoning is harder than forming the appropriate region and determining that it is shaded. For our tasks, this region was always formed from the intersection of two ellipses but in the case of subset tasks the appropriate non-shaded region may be formed from the intersection of multiple ellipses. Hence:

Speculation 3: identifying one region formed from by two intersecting ellipses in a Venn diagram, and determining whether it is shaded, is cognitively easier than drawing inferences from both subset and disjointness textual statements when performing disjoint tasks.

What, then, about the differences between V&T and the other two diagrammatic groups? First we compare Euler and Venn diagrams, as these both exploit closed curves to convey information, but they do so in fundamentally different ways. Firstly, the Euler diagrams used in our study are drawn with circles whereas the Venn diagrams were drawn with ellipses: it is known that circles are a cognitively superior shape (Blake et al. 2014). Moreover, for our tasks, information is extracted from Euler diagrams by comparing the spatial relationships between a pair of circles. The theory of figure-ground segregation suggests that the circles are readily identifiable as shapes in the diagram.

In the Venn case, participants had to identify a region in the diagram, formed from the ellipses, and then determine (a) its spatial relationship with another ellipse in the subset case, or (b) whether it was entirely shaded in the disjointness case. The

appropriate region must be mentally constructed by identifying the interior of one or more ellipses and grouping together the requisite pieces, suggesting that these regions are not readily identifiable in the same way as circles. In the case of linear diagrams, the tasks also boil down to comparing the spatial relationships between lines, and thus similar reasoning concerning figure-ground segregation can explain their superiority over Venn diagrams. This leas to our next speculation:

Speculation 4: identifying the spatial relationships between geometric shapes that are readily identifiable is cognitively easier than identifying and comparing regions formed by shapes.

This discussion leads to an obvious question: would we still see significant performance differences if the tasks were more complex, involving more than two sets? For instance, suppose we wish to determine whether $A \cap B \subseteq C$ holds, as an inference from a textual representation. Here, just like the Venn diagram case, one would need to identify a region formed by the curves used in an Euler diagram (the task no longer involves the identification of the spatial relationship between a pair of circles) and sub-lines would need to be compared in a linear diagram. A clear take-away point is that diagrams that exploit spatial relationships between readily identifiable graphical elements, such as circles or lines, are more effective visualization tools than those which rely on users having to form 'new' graphical elements (e.g. regions) from those which are visually salient (e.g. ellipses). In turn, this may suggest that for more complex tasks, we may reveal performance differences between E&T and L&T, since there may be different cognitive loads associated with identifying regions in the former and line segments in the latter case.

6 Observational Advantages Study: Diagrams as a Stand-alone Representation

We now focus on whether using diagrams, without accompanying text, reveals any further insight than was obtained from the free rides study. Here, RQ2 is broken down into the following questions:
RQ2a: Do the observational advantages exhibited by the considered diagram types bring about significant task performance benefits over text?
RQ2b: Do the observational advantages exhibited by any one diagram type lead to significant task performance benefits over the other diagram types in the absence of an accompanying textual representation?
RQ2c: Do tasks concerning subsets lead to significantly different task performance compared to those concerning disjointness?

In reporting on our data collection and results, we will refer to the treatments as

- T for 'text only',
- L for linear diagrams,
- V for Venn diagrams, and
- E for Euler diagrams.

Fig. 18 T group stimulus

No one visited both Albania and Croatia
No one visited both Albania and Denmark
No one visited both Albania and Oman
Everyone who visited Croatia visited Denmark
Everyone who visited Mali visited Albania

Fig. 19 L group stimulus

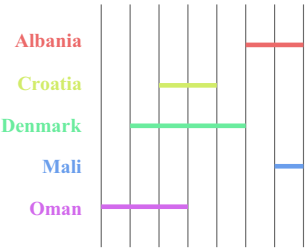
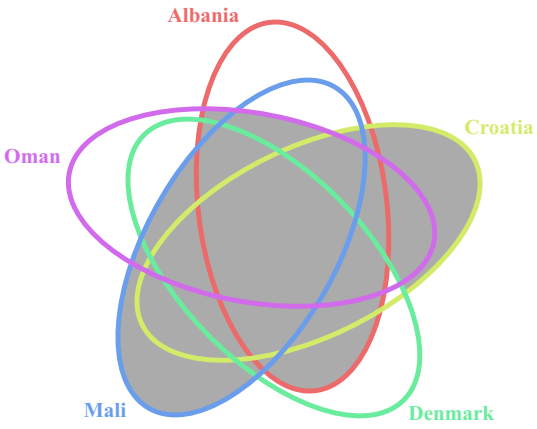


Fig. 20 V group stimulus



The online version of this study, where it is possible to select the group in which to take part (unlike the actual study where participants were randomly assigned to a group) can be found here: https://www.cs.kent.ac.uk/people/staff/pjr/observationaladvantages/. An example of the stimuli for study task, as presented to each group, can be seen in Figs. 18, 19, 20 and 21, with the associated options in Fig. 16; as can be seen, this task for study 2 corresponds to that for Figs. 12, 13, 14 and 15 and highlights the key difference between the two studies regarding the inclusion or otherwise of textual statements with the diagrams.

In this second study, it was necessary to use different questions to identify inattentive participants, since those used in the free rides study relied on the presence of textual statements (above the diagrams) across all four groups; see Fig. 11. Given the absence of textual statements in the diagrams groups, we therefore changed the five multiple choice options presented to participants to indicate that the question was an attention check:

This is an Attention Check Question
It is not timed, You May Rest Now
Please Click on the Indicated Option

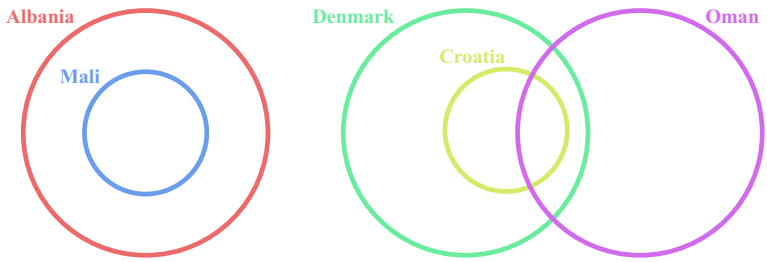


Fig. 21 E group stimulus

You Must Do This In Order to be Paid
Please choose this option

When running a pilot study, we pre-screened participants using the same criteria as the study reported on in Sect. 6, with the additional criterion that nobody had taken part in the first study. Again, we specified that the study was compatible with desktop devices.

This left a pool of 32486 potential participants, out of 75519. Each participant was randomly allocated to one of the four groups. They were each paid £2.75, reflecting the time we expected it to take to complete the study. This was reduced from £3.92 because the average completion time for the first study was lower than had been indicated by the first study pilot study at less than 22 minutes. Based on this, we reverted to an expected completion time of 20 minutes, with a maximum allowed time of 45 minutes.

A total of 20 people began the pilot study. Of these, two were classified as inattentive, none timed-out after 45 minutes and a further seven withdrew before completion (four in the Venn group, two in the Text group and 1 in the Linear group). However, the two participants classified as inattentive both appeared to have paid attention, with 90% and, respectively, 100% accuracy rates. Studying their responses to the questions designed to identify inattentive participants, one of these participants selected "This is an Attention Check Question" the other chose "You Must Do This In Order to be Paid", in each case for both inattentive questions. Therefore, we reworded the options. For one of the inattentive questions, the options became:

Choose this Option This is an Attention Check
Everyone who visited Hungary visited Mali
No one visited both Jamaica and Hungary
Everyone who visited Hungary visited Albania
None of the above

where none of the country names appeared in the associated textual statements (for the T group) or diagram. For the other inattentive question, the changes were similar, again ensuring that none of the country names appeared in the associated textual statements or diagrams.

We then proceeded to run a second pilot study, recruiting 26 participants. Of these, three were classified as inattentive (2 Venn, 1 Euler) and four withdrew before completing (1 Text, 1 Euler, 1 Venn and 1 unknown). Two people timed out after 45 minutes, both in the Venn group and one of these was also classified as inattentive. This left

Table 14 Observational advantages study: summary of the second pilot data

Group	No. of participants	Accuracy (%)	Mean Time
Overall	18	65.56	34.40
T	5	47.00	35.27
L	5	83.00	19.86
V	5	54.00	50.32
E	3	86.67	30.66

data from 18 participants (five Venn, five Text, three Euler, and five Linear). Prolific indicated that it took participants on average 19 minutes to complete the second pilot study which includes all time spent on training, performing the tasks and supplying demographic information. Therefore, the average reward was £8.68 per hour. The overall accuracy rate and the average (mean) time in seconds to answer each question (in seconds) are given in Table 14, alongside a breakdown for each group. Based on these data, the payment was kept at £2.75 for the main study.

In the pilot data, we noted that question 13 had a high completion time, compared to the other questions. However, one participant – in the Venn group – spent over 15 minutes on that question and, so, no change was necessary. A typographical error was also identified and rectified in Q18, where Hungary was misspelled. Lastly, we found an error in the software, where the last option for the inattentive questions had failed to be updated. This was rectified.

Adding the criterion that nobody who had taken part in the pilots to the pre-screening criteria in Prolific left 33026 eligible participants. A total of 549 people began the main study. Of these, 27 were classified as inattentive and a further 98 started the study but we have no record of their completion. Another six participants were recorded as having completed the study by Prolific but for whom we could not obtain their data. These six participants were paid, but are not included in the analysis. In one further case, a participant's data was corrupted, so excluded from the analysis; they were in the linear group. Another participant timed-out after 45 minutes whilst still returning data, but was classified as one of the 27 inattentive. A further five participants timed-out whilst still returning data and, so, their data was included in the analysis. We speculate that these five time-outs arose due to the participants' participation not being finalised correctly in Prolific. We have data from 417 participants (demographics: age: mean 33, range from 18 to 72; gender: 126 M , 290 F, 1 other; the distribution was similar across the four groups). The breakdown of these data across the groups is given in Table 15.

As with the free rides study, we again saw a high number of non-completions. This time, most were in the T group (47 out of 98; 48%), with a substantial number also in the V group (36 out of 98; 37%). For similar reasons to the free rides study, this may mean that our data present an artificially high accuracy rate and low time performance for the T group compared to all other groups. Moreover, the performance of the V group may be artificially high for accuracy and low for time compared to the other two diagrammatic groups. These observations could impact the validity of our results.

Table 15 Observational advantages study: summary of the participant distribution for the main study

Group	Completions	Inattentive	Non-completions	No data
Overall	417	27	98	7
T	103	9	47	–
L	103	3	7	1
V	112	11	36	–
E	99	4	8	–
Unknown	–	–	–	6

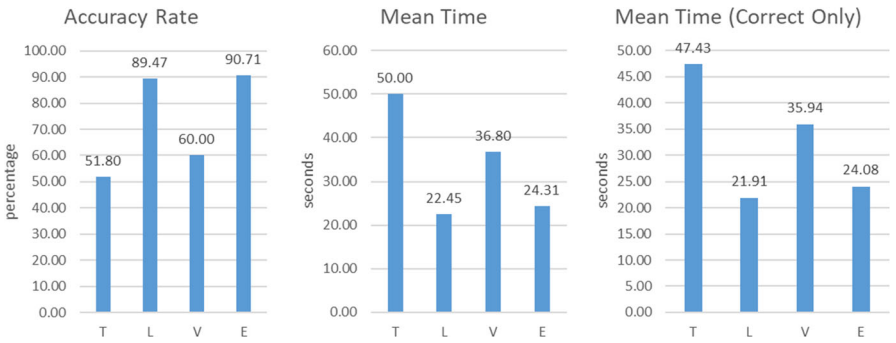


Fig. 22 Observational advantages study: visualizing the accuracy rate and mean times

6.1 Comparison of Representations

For the observational advantages study, the data and associated statistical output can be found at https://www.cs.kent.ac.uk/people/staff/pjr/observationaladvantages/paper.html. The accuracy rates and mean times (in seconds) are summarised in Table 16. We again remind the reader that, whilst the accuracy rates and mean times are in *indicator* of relative performance, the statistical methods employed do not compare these data and these methods are used for the same reasons as given previously. For this reason, the standard graphical plots visualising means and related features of the data, given in Fig. 22, can be misleading when interpreting the results of our statistical analysis. However, they still provide a useful insight into the performances differences.

Before we present our statistical analysis, we make some observations on the accuracy rates and mean times across the two studies. Firstly, the accuracy rates are almost identical for three of the corresponding treatments: the T group recorded a 51.11% accuracy rate in study 1 and a 51.80% rate in study 2, the L group recorded rates of 87.58% and 89.47%, and the E group recorded rates of 91.02% and 90.71%. By contrast, the V&E group recorded a rate of 51.07% in study 1 which increased to 60.00% in study 2. Whilst it is not robust to determine whether this is a statistically significant improvement in accuracy performance, in part since the data were collected at different times, this again indicates that the combination of using Venn diagrams alongside

Table 16 Observational advantages study: summary of the main study data

Group	No. of participants	Accuracy (%)	Mean time	Mean time (correct only)
Overall	417	72.54	33.55	30.17
T	103	51.80	50.00	47.43
L	103	89.47	22.45	21.91
V	112	60.00	36.80	35.94
E	99	90.71	24.31	24.08

Table 17 Observational advantages study: overall comparison of treatments by accuracy

Treatments	Odds	CI	p-value	Sig.	Most accurate
L versus T	7.90	(5.41, 11.56)	< 0.0001	✓	L
L versus V	5.66	(3.76, 8.53)	< 0.0001	✓	L
L versus E	0.87	(0.54, 1.40)	0.5669	×	–
T versus V	0.72	(0.54, 0.95)	0.0211	✓	V
T versus E	0.11	(0.08, 0.16)	< 0.0001	✓	E
V versus E	0.15	(0.10, 0.23)	< 0.0001	✓	E

text is problematic. In addition, there are notable differences in time performance between the studies for the groups including 'diagrammatic' treatments. For instance, the L&T group took, on average, approximately 30% (resp. 25%) longer to provide a (resp. correct) answer than the L group. Similar observations can be made for the other 'diagrammatic' treatments. The accuracy data and the differences in time data may indicate that the textual statements are being used in some way by the participants who were exposed to diagrams alongside text (at the very least, we can speculate that the textual statements were noticed and played some part in the performance of the task) in the free rides study. We now return our attention to the statistical analysis for the observational advantages study.

Using a GEE based statistical model for the accuracy data, we estimated a 95% CI for the odds of providing a correct answer with one treatment compared to another and associated p-values. The estimated odds of correctly answering questions with L was 7.90 (to 2d.p.) times higher than that of T with a 95% CI of (5.41,11.56) and p-value of < 0.0001 (to 4d.p.). Therefore, L supported significantly more accurate task performance than T only. Results for the other pairwise comparisons are given in Table 17, from which we can derive an overall ranking for the treatments:

$$\textit{accuracy overall ranking: } L = E > V > T.$$

Using a GEE based statistical model for the time data, we estimated a 95% CI for the ratio of the time (measured in seconds) needed to answer a question correctly with one treatment compared to another. The model estimated that the time needed to answer a question correctly with L was 0.53 times (2d.p.) that with T with a 95% CI of $(0.47, 0.59)$ and p-value of < 0.0001. Therefore, linear diagrams supported

Table 18 Observational advantages study: overall comparison of treatments by time

Treatments	Ratio	CI	p-value	Sig.	Fastest
L versus T	0.53	(0.47, 0.59)	< 0.0001	✓	L
L versus V	0.63	(0.57, 0.68)	< 0.0001	✓	L
L versus E	0.93	(0.85, 1.02)	0.1393	×	–
T versus V	1.19	(1.07, 1.33)	0.0019	✓	V
T versus E	1.76	(1.57, 1.97)	< 0.0001	✓	E
V versus E	1.48	(1.35, 1.62)	< 0.0001	✓	E

significantly faster task performance than text only. Results for the other pairwise comparisons are given in Table 18, from which we can derive an overall ranking for the treatments:

$$\textit{time overall ranking: } L = E > V > T.$$

Therefore, our accuracy and time analysis consistently support the superiority of linear diagrams and Euler diagrams as compared to Venn diagrams which, in turn are superior to using text alone. Taking into account both the accuracy and time analysis, we have consistent rankings, from which we derive an overall ranking of the four treatments:

$$\textbf{overall ranking: L = E > V > T.}$$

The odds and ratios computed for the accuracy and, respectively, time data give insight into the effect size. For instance, the odds of producing a correct answer using linear diagrams compared to text alone are 7.90, which is a practically significant effect size. From the perspective of time, we would expect correctly answering a question using linear diagrams to be 0.53 of the time taken (i.e. to take 53% of the time) to provide a correct answer using text alone. Again, this is a practically significant effect size. The other effect sizes, where we saw significant differences, are similar and are evident from the odds and ratios given in the tables. It is notable that Venn diagrams performed significantly better than just text, unlike in the free rides study where we did not find significant performance differences overall between these two treatments.

6.1.1 Subset Comparison Across Treatments

For the subset tasks, the indicative accuracy rates and mean times are in Table 19. As with the overall analysis, we produced a GEE based statistical model for the accuracy data. Results for the pairwise comparisons are given in Table 20, from which we can derive a ranking for the treatments for subset tasks:

$$\textit{accuracy subset-task ranking: } L = E > V = T.$$

Table 19 Observational advantages study: summary of the main study data for subset tasks

Group	No. of participants	Accuracy (%)	Mean time	Mean time (correct only)
Overall	417	73.43	32.44	29.48
T	103	55.36	43.43	40.24
L	103	89.42	23.37	22.53
V	112	62.14	37.70	36.23
E	99	88.99	24.49	24.19

Table 20 Observational advantages study: subset comparison of treatments by accuracy

Treatments	Ratio	CI	p-value	Sig.	Most accurate
L versus T	5.15	(3.31, 8.01)	< 0.0001	✓	L
L versus V	6.81	(4.28, 10.85)	< 0.0001	✓	L
L versus E	1.05	(0.62, 1.75)	0.8666	×	–
T versus V	1.32	(0.94, 1.86)	0.1040	×	–
T versus E	0.20	(0.14, 0.31)	< 0.0001	✓	E
V versus E	0.15	(0.10, 0.24)	< 0.0001	✓	E

A GEE based statistical model for the time data yielded the results given in Table 21, from which we can derive a ranking for the treatments for the subset tasks:

$$\textit{Time subset-task ranking: } L = E > V = T.$$

Thus obtain the following ranking of treatments for subset tasks:

Subset-task ranking: L = E > V = T.

As with the overall analysis, the odds and ratios computed for the accuracy and, respectively, time data give insight into the effect size. Perhaps the most interesting result here is that there was no significant performance difference between V and T, unlike for V&T and T where T was significantly more accurate than V&T. This reinforces the conclusion that using Venn diagrams in combination with text hinders performance.

6.1.2 Disjointness Comparison Across Treatments

For the disjointness tasks, the indicative accuracy rates and mean times are in Table 22. As with the overall analysis, we produced a GEE based statistical model for the accuracy data. A GEE based statistical model for the accuracy data yielded the results given in Table 23 and a GEE based model for the time data yielded the results in Table 24. We obtain these two performance rankings:

$$\textit{accuracy disjoint-task ranking: } L = E > V > T.$$

Table 21 Observational advantages study: subset comparison of treatments by time

Treatments	Ratio	CI	p-value	Sig.	Most accurate
L versus T	0.61	(0.55, 0.68)	< 0.0001	✓	L
L versus V	0.64	(0.58, 0.71)	< 0.0001	✓	L
L versus E	0.93	(0.84, 1.03)	0.1611	×	–
T versus V	1.05	(0.94, 1.17)	0.4188		
T versus E	1.52	(1.36, 1.70)	< 0.0001	✓	E
V versus E	1.45	(1.31, 1.61)	< 0.0001	✓	E

Table 22 Observational advantages study: summary of the main study data for disjointness tasks

Group	No. of participants	Accuracy (%)	Mean time	Mean time (correct only)
Overall	417	71.66	34.66	30.88
T	103	64.64	56.58	58.20
L	103	89.52	21.54	21.29
V	112	41.46	35.89	35.70
E	99	92.42	24.12	23.97

Table 23 Observational advantages study: disjointness comparison of treatments by accuracy

Treatments	Odds	CI	p-value	Sig.	Fastest
L versus T	12.06	(7.89, 18.43)	< 0.0001	✓	L
L versus V	4.67	(2.96, 7.36)	< 0.0001	✓	L
L versus E	0.70	(0.41, 1.20)	0.1961	×	–
T versus V	0.39	(0.28, 0.53)	< 0.0001	✓	V
T versus E	0.06	(0.04, 0.09)	< 0.0001	✓	E
V versus E	0.15	(0.09, 0.24)	< 0.0001	✓	E

$$\textit{time disjoint-task ranking: } L = E > V = T.$$

Thus obtain the following ranking of treatments for disjoint tasks:

disjoint-task ranking: L = E > V > T.

6.1.3 Summary

The overall analysis and that for the two task types allows us to suggest answers to RQ2a and RQ2b. Concerning RQ2a, our study suggests that linear, Venn and Euler diagrams consistently supported significantly better task performance than using textual statements. However, Venn diagrams were only superior to textual statements for disjointness tasks, with there being no significant difference for subset tasks. Focusing

Table 24 Observational advantages study: disjointness comparison of treatments by time

Treatments	Ratio	CI	p-value	Sig.	Fastest
L versus T	0.43	(0.37, 0.49)	< 0.0001	✓	L
L versus V	0.62	(0.56, 0.68)	< 0.0001	✓	L
L versus E	0.93	(0.84, 1.03)	0.1446	×	–
T versus V	1.45	(1.26, 1.66)	< 0.0001	✓	V
T versus E	2.18	(1.89, 2.50)	< 0.0001	✓	E
V versus E	1.51	(1.37, 1.66)	< 0.0001	✓	E

Table 25 Observational advantages study: comparison of subset versus disjoint tasks by treatment w.r.t. accuracy

Treatments	Odds	CI	p-value	Sig.	Most accurate
T	2.32	(1.88, 2.86)	< 0.0001	✓	Subset
L	0.99	(0.70, 1.39)	0.9531	×	–
E	0.66	(0.49, 0.89)	0.0062	✓	Disjoint
V	0.68	(0.54, 0.85)	0.0009	✓	Disjoint

on RQ2b, we again found that linear and Euler diagrams are a superior representation as compared to Venn diagrams in this study both overall and for the two task types.

6.2 Observational Advantages Study: Comparison of Task Types

Here we address RQ2c, which asked whether the two types of task yield significantly different performances in the case of observational advantages. We do this by considering each treatment in turn, with the statistical results given in Tables 25 and 26. In summary, for accuracy and time, there was no difference between the task types for L only. For T, we saw that subset tasks were performed significantly more accurately and significantly faster. This is congruent with our results from the free rides study. For Euler and Venn diagrams, disjointness tasks were performed significantly faster, with no accuracy differences. Again, these results are congruent with those from the free rides study. The results for Euler diagrams consistently support the indication that it is easier for people to establish that two circles do not overlap than to determine whether one circle is inside another. But a clear take-away message is that the cognitive load associated with the identification of observational advantages is dependent on the task type. This reinforces our observations from the free rides study that more work is needed to understand the role that free rides and observational advantages have in facilitating inference and their cognitive benefits.

Table 26 Observational advantages study: comparison of subset versus disjoint tasks by treatment w.r.t. time

Treatments	Ratio	CI	p-value	Sig.	Fastest
T	0.72	(0.79, 0.65)	< 0.0001	✓	Subset
L	1.03	(1.00, 1.07)	0.0774	×	–
E	1.03	(0.99, 1.07)	0.1065	×	–
V	0.99	(0.94, 1.06)	0.8681	×	–

Table 27 Rankings of treatments overall and for task types

Category	Free rides study	Observational advantages study
Overall	L&T = E&T > V&T = T	L = E > V > T
Subset	L&T = E&T > T > V&T	L = E > V > T
Disjoint	L&T = E&T > V&T > T	L = E > V > T

Table 28 Rankings of task type by treatment

Treatment	Free rides study	Treatment	Observational advantages study
Text	Subset > disjoint	Text	Subset > disjoint
L&T	Subset = disjoint	L	Subset = disjoint
E&T	Disjoint > subset	E	Disjoint > subset
V&T	Disjoint > subset	V	Subset = disjoint

6.3 Observational Advantages Study: Interpretation

We now seek to explain the results we have found in this study and across the two studies. For ease of reference, the main rankings of treatments across the two studies can be seen in Table 27 and those for the task types are in Table 28.

As with the free rides study, we must be mindful of the high numbers of non-completions, this time in the T and V groups. We saw that 48% (47 out of 98) of the non-completions were in the T group and 37% (36 out of 98) were in the V group. Therefore, our data may present an overly supportive view of these two treatment groups as non-completions could imply that participants withdrew due to their perceived difficulty of the tasks. In this study, these two treatments performed significantly worse than both E and L, so even if our data are overly supportive for T and V, we can be confident that our results ranking L and E more highly than T and V are robust. The results that may be affected concern the ranking of T and V, where overall and for both task types V outperformed T. We noted more non-completions for the T group, however, leading us to posit that the ranking of V and T is also likely to be robust.

The comments provided by participants again support the possibility that the V and T questions were perceived as more difficult: 27 out of the 103 (26%) participants who completed the T questions and 31 out of the 112 (28%) participants who completed the V questions made comments that indicated they found the tasks hard. By contrast,

in the E group, 10 out of 99 (10%) participants commented on the tasks being hard, which reduced to only 2 out of 103 (2%) in the L group. These differences in opinions are supported by our data, which ranked L and E more highly than V which, in turn was ranked more highly than T. The result that Venn diagrams led to significantly worse performance that both L and E is consistent with prior work which evaluated different visualizations of sets with tasks that were similar to those in this paper, but were more varied (Chapman et al. 2014).

As we continue with our discussion, we will make reference to speculations 1 to 4 identified earlier when we interpreted the results of the free rides study. Our first observation is that the results, from the observational advantage study, are more aligned with our expectations: the diagrams groups consistently outperformed the text-only group. We will delve into reasons for these results shortly, with our initial focus on the T group. In particular, the results for T are consistent with the free rides study for the task types: the result that subset tasks were performed significantly more accurately with text was seen in both studies. We would expect consistency across studies as the two sets of T group participants were exposed to exactly the same training and performance tasks. Therefore, this result reinforces speculation 1: the transitivity of subset and the use of single statements types renders the subset tasks more straightforward than the disjointness tasks.

Regarding speculation 2, the results of the observational advantages study are more interesting. We suggested why Venn diagrams may cause problems when performing subset tasks, due to the need to identify and compare regions in the diagrams formed by multiple intersecting ellipses, as compared to disjointness tasks. It is, however, notable that, for subset tasks, Venn diagrams *without accompanying text* do not degrade performance as compared to using just text. Unlike the free rides study, where V&T performed subset tasks significantly worse than T, here we found that V significantly outperformed T. Therefore, this refutes speculation 2 and leads us to instead suggest that the combination of Venn diagrams and textual statements is somehow confusing for these tasks. The reasons for this are unclear, but leads us to posit that, at least in the case of Venn diagrams and text, using multiple modalities to represent information need not be beneficial. Either way, we can make two new speculations based on the results of the observational advantages study, combined with the results of the free rides study:

speculation 5: identifying a non-shaded region formed from multiple intersecting ellipses in a Venn diagram is cognitively easier than using subset transitivity in textual statements when performing subset tasks.

speculation 6: using Venn diagrams alongside text is cognitively harder than using just text or just Venn diagrams when performing subset tasks.

Concerning disjointness tasks, our results are consistent across both studies: the groups exposed to Venn diagrams (with or without text) performed these tasks significantly better than the groups using text alone. Therefore, this result is congruent with, and reinforces, speculation 3. Moreover, since L and E were significantly better than V both overall and for both task types, the observational advantages study also reinforces speculation 4, which focused on the cognitive difference between identify-

ing geometric shapes as opposed to the regions they form. The results of studies 1 and 2 combined, therefore, provide strong evidence for the correctness of speculations 1, 3 and 4 as well as providing new insights, captured in speculations 5 and 6. A key takeaway message here is we have evidence that observational advantages can bring about significant performance benefits in visualizations of sets.

7 Conclusion

This paper has provided the first empirical studies into how free rides and observational advantages help users to understand the information in diagrams used for visualizing sets. Our research has revealed that the role of free rides in explaining the cognitive benefits of diagrams is not clear-cut. Whilst Euler and linear diagrams both provided effective support for the textual representations from which they were derived, Venn diagrams sometimes led to significantly worse task performance. Returning to Shimojima's influential insights, covered in Sect. 2, we can say that just because a diagram expresses other, consequential information, it does not *necessarily* mean that the diagram, when presented alongside a semantically equivalent textual representation, facilitates inference and saves "deductive cost", from either a time or accuracy perspective. It is particularly noteworthy that Shimojima's discussion on the potential cognitive benefits of free rides includes Venn diagrams as an exemplar.

Shimojima's argument is predicated on the assumption that interpreting the diagram is a more straightforward cognitive process than drawing an inference from information expressed in another notation. Our second study, which focused on observational advantages, supports this view where we found all three diagrammatic treatments – linear diagrams, Venn diagrams and Euler diagrams – gave rise to significant performance benefits over using just a textual representation. However, in the case of Venn diagrams, these benefits were not seen for subset tasks, only disjointness tasks. But, by and large, the results from our observational advantages study indicate that when diagrams expresses other, consequential information, they facilitate inference and typically save "deductive cost". That is, the theoretical notion of an observational advantage does indeed lead to cognitive advantages. This insight is important for the exploitation of notations when visualizing data: we should prioritise the choice of representations of information that allow information to be observed, as opposed to inferred.

Our results also indicate that, whilst free rides and observational advantages in effective diagrams can lead to performance improvements, it is also important to use well-designed diagrams suitable for the task at hand. One cannot just rely on the presence of free rides or observational advantages as an argument for the efficacy of a representation. For instance, the consistently poor performance of the Venn diagrams groups, with or without text, as compared to the other diagrammatic groups indicates that other general features of notations are important to identify and evaluate. One such feature is that of well-matchedness (Gurr 1999): if a diagram's meaning carriers resemble the semantics they convey then the diagram is considered to be well-matched. In an Euler diagram, for example, the disjointness of two sets is resembled in the diagram by two circles with disjoint interiors. By contrast, the use of shading to

express set-emptiness and, thus, to indicate the disjointness of two sets is not well-matched. It may well be that the lack of well-matchedness inherent in Venn diagrams could explain the results found in our second study and to some extent the first study.

To summarise, our results show that the free rides and observational advantages present in linear and Euler diagrams aid user understanding. However, the lack of a consistent benefit with Venn diagrams points to the need for future research so as to better understand the features of representations that are indicators of cognitive efficacy. We suggest that a more comprehensive and integrated theory of general features of diagrams that make them effective for cognition should be a major goal of the information visualization community.

References

Alper, B., Henry Riche, N., Ramos, G., & Czerwinski, M. (2011). Design study of LineSets, a novel set visualization technique. *IEEE Transactions on Visualization and Computer Graphics, 17*(12), 2259–2267.

Alsallakh, B., Micallef, L., Aigner, W., Hauser, H., Miksch, S., & Rodgers, P. (2014). Visualizing sets and set-typed data: State-of-the-art and future challenges. In *Eurographics Conference on Visualization (EuroVis)* (pp. 124–138).

Blake, A., Stapleton, G., Rodgers, P., Cheek, L., & Howse, J. (2014). The impact of shape on the perception of Euler diagrams. In *8th International Conference on the Theory and Application of Diagrams* (pp. 124–138). Springer.

Blake, A., Stapleton, G., Rodgers, P., & Howse, J. (2016). The impact of topological and graphical choices on the perception of Euler diagrams. *Information Sciences, 330,* 455–482.

Chapman, P., Stapleton, G., Rodgers, P., Michallef, L., & Blake, A. (2014). Visualizing sets: An empirical comparison of diagram types. In *Diagrams 2014* (pp. 146–160). Springer.

Chen, J., Menezes, N., Bradley, A., & North, T. (2011). Opportunities for crowdsourcing research on Amazon Mechanical Turk. *Human Factors, 5*(3).

Cheng, P. (2011). Probably good diagrams for learning: Representational epistemic recodification of probability theory. *Topics in Cognitive Science, 3,* 475–496.

Chow, S., & Ruskey, F. (2005). Towards a general solution to drawing area-proportional Euler diagrams. In *Euler Diagrams 2004, ENTCS* (Vol. 134, pp. 3–18). ENTCS.

Collins, C., Penn, G., & Carpendale, M. S. T. (2009). Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 1009–1016.

Gottfried, B. (2015). A comparative study of linear and region based diagrams. *Journal of Spatial Information Science, 2015*(10), 3–20.

Gurr, C. (1999). Effective diagrammatic communication: Syntactic, semantic and pragmatic issues. *Journal of Visual Languages and Computing, 10*(4), 317–342.

Harrower, M., & Brewer, C. (2003). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal, 40*(1), 27–37.

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

Meulemans, W., Henry Riche, N., Speckmann, B., Alper, B., & Dwyer, T. (2013). Kelpfusion: A hybrid set visualization technique. *IEEE Transactions on Visualization and Computer Graphics, 19*(11), 1846–1858.

Riche, N., & Dwyer, T. (2010). Untangling Euler diagrams. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 1090–1099.

Rodgers, P., Stapleton, G., & Chapman, P. (2015). Visualizing sets with linear diagrams. *ACM Transactions on Computer-Human Interaction, 22*(6), 28.

Shimojima, A. (1996). On the efficacy of representation. PhD Thesis, Indiana University.

Shimojima, A. (2015). *Semantic Properties of Diagrams and Their Cognitive Potentials*. Stanford, CA, USA: CSLI Publications.

Simonetto, P., Auber, D., & Archambault, D. (2009). Fully automatic visualisation of overlapping sets. *Computer Graphics Forum, 28*(3), 967–974.

Stapleton, G., Jamnik, M., & Shimojima, A. (2017). What makes an effective representation of information: A formal account of observational advantages. *Journal of Logic, Language and Information, 26*(2), 143–177.

Stapleton, G., Zhang, L., Howse, J., & Rodgers, P. (2011). Drawing Euler diagrams with circles: The theory of piercings. *IEEE Transactions on Visualization and Computer Graphics, 17*(7), 1020–1032.

Takemura, R. (2019). Proof-theoretical investigation of Venn diagrams: a logic translation and free rides. https://abelard.flet.keio.ac.jp/person/takemura/paper/final-rev-vennres.pdf. Accessed February 2019.

Venn, J. (1880). On the diagrammatic and mechanical representation of propositions and reasonings. Mag: Phil.

Wagemans, J., Elder, J., Kubovy, M., Palmer, S., Peterson, M., & Singh, M. (2012). A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organisation. *Psychol Bull., 138*(6), 1172–1217.

Wilkinson, L. (2012). Exact and approximate area-proportional circular Venn and Euler diagrams. *IEEE Transactions on Visualization and Computer Graphics, 18*(2), 321–331.