Evaluating Colour in Concept Diagrams

Sean McGrath[1][0000−0002−5306−9218], Andrew Blake[2][0000−0001−5856−4544], Gem Stapleton[3][0000−0002−6567−6752], Anestis Touloumis[2][0000−0002−5965−1639], Peter Chapman[4][0000−0002−5524−5780], Mateja Jamnik[3][0000−0003−2772−2532], and Zohreh Shams[3][0000−0002−0143−798X]

[1] Goldsmiths, University of London, London, UK, S.McGrath@gold.ac.uk
[2] University of Brighton, Brighton, UK, {a.l.blake,a.touloumis}@brighton.ac.uk
[3] University of Cambridge, Cambridge, UK, {ges55,m.jamnik,zs315}@cam.ac.uk
[4] Edinburgh Napier University, Edinburgh, UK, p.chapman}@napier.ac.uk

Abstract. This paper is the first to establish the impact of colour on users' ability to interpret the informational content of concept diagrams, a logic designed for ontology engineering. The research is motivated by known results for Euler diagrams, which form a fragment of concept diagrams: manipulating curve colour affects user performance. In particular, using distinct curve colours is known to yield significant performance benefits in the case of Euler diagrams. Naturally, one would expect to obtain similar empirical results for concept diagrams, since colour is a graphical feature to which we are perceptually sensitive. Thus, this paper sets out to test this expectation by conducting a crowdsourced empirical study involving 261 participants. Surprisingly, our study suggests that manipulating curve colour no longer yields significant performance differences in this syntactically richer logic. This raises the exciting prospect of identifying the boundary of where manipulating graphical features is to users' cognitive advantage.

Keywords: concept diagrams · Euler diagrams · perception · colour.

1 Introduction

There is a growing body of evidence that diagrams can help people with logical reasoning, with research primarily focusing on logics with low expressiveness [16, 19, 20, 22]. As particular examples, diagrams have been found to aid some students with deductive reasoning tasks as compared to standard symbolic logic [22], and Euler diagrams have been shown to increase people's accuracy when performing syllogistic reasoning tasks [20]. In addition, fMRI studies have found, in the context of reasoning, that diagrams provide cognitive offloading and therefore aid cognition, as compared to stylized natural language [19]. Most directly related to this paper is work by Alharbi et al. which suggests that concept diagrams support more effective interpretation of information than both OWL (strictly, the Manchester OWL syntax) and description logic [2]. In summary, the prior work covered here is to compare diagrammatic representations of

information with competing notations. The takeaway message is that diagrammatic logics have been shown to effectively support users with tasks, thus giving them an accessibility advantage over their symbolic and textual counterparts. By contrast, this paper sets out to understand the impact of manipulating colour in concept diagrams, in order to increase their efficacy in logical reasoning tasks.

Euler diagrams are the underling notation of concept diagrams, as well as many other diagrammatic logics [16, 19, 20, 22]. It is known that reducing clutter in Euler diagrams improves cognition [3], as does ensuring that they possess so-called *well-formedness properties* [18]. There is no reason to suppose that low clutter and possessing well-formedness properties are not beneficial for concept diagrams. Empirical research has also focused on the *graphical features* of Euler diagrams. Prior work, such as [5], provides a series of empirically-informed guides that point towards effective graphical choices, such as how to use colour, choose curve shapes, and orient diagrams. The guide most relevant to this paper is the that for colour: *draw Euler diagrams with curves that have no fill and different colours for each represented set.* An immediate question arises, which we address in this paper: does this guide also apply to concept diagrams?

The study on which this colour-guide, for Euler diagrams, is based was limited to around eight curves. It is estimated that between eight and ten colours can be rapidly distinguished at a time by the human eye [10, 17, 24]. Whilst the reasons for this are not known [10], Miller has hypothesized that it is because humans are only able to store this number of items in their short-term memory [15]. Therefore, since concept diagrams can often include more than ten syntactic elements, varying the hues assigned to them may no longer bring performance advantages. It is important to ascertain whether manipulating colour in concept diagrams can be done in such a way that performance is significantly improved.

Appealing to [4], for our study we defined three colour treatments for concept diagrams: (1) monochrome, all diagrammatic elements were black, (2) dichrome, selected diagrammatic elements were blue and the remaining were green (as in [11, 12, 21]), and (3) polychrome, different colours were assigned to the syntactic elements; none of the curves used a colour fill. Thus, treatment (3) follows the prior guidance and could be expected to outperform (1) and (2), but with the caveat that using more than eight to ten colours may have the potential to be detrimental. The paper proceeds as follows. Section 2 gives a brief overview of concept diagrams. Our study design is described in section 3, with the study execution and results covered in section 4. We discuss our results and conclude the paper in section 6.

2 A Brief Introduction to Concept Diagrams

Concept diagrams include a variety of syntactic elements in order to convey information [21]. In this paper, we evaluate a fragment of the notation, since we do not need the full expressive power; concept diagrams are a second-order logic rendering them highly expressive. We introduce, by example, the syntax needed for the study in this paper. Figure 1 shows two concept diagrams. On the

left, the diagram contains one curve inside another to express that (the set of) Korrigans is a subset of Spirits: *all Korrigans are Spirits*. The righthand diagram contains two non-overlapping curves to express that the sets Demon and Elf are disjoint. The boxes are used to indicate the boundaries of each diagram. So, in figure 2 there are two juxtaposed diagrams; each individual diagram carries no meaning in this case, and the fact that there are two non-overlapping curves, Mermaid and Giant, does not convey any information since the respective curves are inside distinct rectangles. That is, spatial relationships only convey meaning inside a common bounding box.

In figure 3, there are also *two* diagrams. On the left, *the diagram comprises two boxes*, each of which encloses some syntax, with an arrow between them. This *solid* arrow, labelled scares, is sourced on Boggart and targets an unlabelled curve which is a subset of Midget. This particular diagram asserts that Boggarts scare only Midgets. On the right, the diagram is structurally similar to that on the left, but instead uses a *dashed* arrow, labelled by *annoys* and annotated with ≥ 1, which is an important (symbolic) device used to convey cardinality information in the following way: Goblins annoy *at least one* thing in the arrow's target set. Since the arrow's target set is inside Fairy, we can provide the meaning of this arrow in a much more succinct way: Goblins annoy at least one Fairy. Note that the use of a dashed arrow does not provide 'only' information as we saw in the case of the solid arrow.

As well as being sourced on curves, arrows can be sourced on the enclosing box. This box is taken to represent the universal set, so we can talk about *everything* or, more simply, *things*. Two examples are given in figure 4. On the left, the solid arrow targets a subset of Puck: things chase only Pucks. Essentially, a diagram with this syntactic construction is expressing a *range axiom*: the range of chase is Puck. In the diagram on the right, the arrow's label, *likes*, has an annotation: $^-$. The use of $^-$ is to indicate that we mean the *inverse* of the binary relation *likes*. Thus, the diagram is expressing that things 'like inverse' only Nisses. This is equivalent to *Only Nisses like things* which is a *domain axiom*. Using these basic constructions, more complex diagrams can be formed, like that in figure 5 which uses multiple colours for its syntactic elements. It expresses many facts, such as:

- All Halflings are Midgets.
- No Goblin is a Demon.
- Pucks follow only Halflings.
- Elfs chase at least on Fairy.
- Things guide only Pucks.
- Only Demons scare things.

Diagrams with this level of complexity were used to collect performance data.

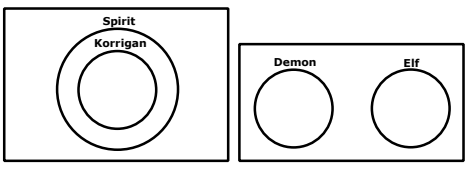

Fig. 1. Subset and disjointness.

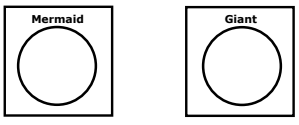

Fig. 2. Non-disjointness.

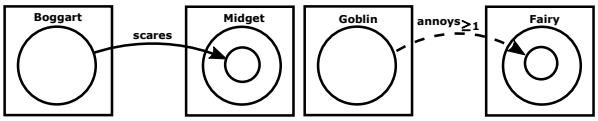

Fig. 3. Diagrams involving arrows.

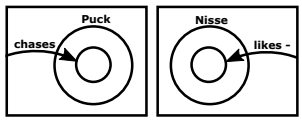

Fig. 4. Range and domain.

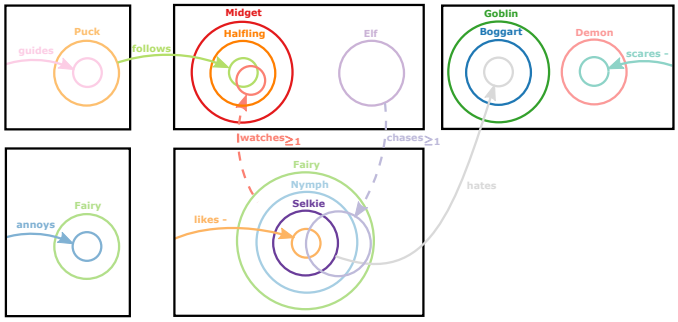

Fig. 5. A more complex diagram which expresses many different statements.

3 Study Design

We will now describe our between group study design including: the information conveyed by the diagrams, the colour treatments, participant training, strategies to manage learning effects, performance phase questions, our approach to data collection, and the statistical methods employed. Our study comprised the following phases:

1. Training phase: participants were shown a series of simple diagrams along with their interpretations.
2. Learning effect phase: participants were asked two questions, similar to those in the next phase.
3. Performance phase: participants were asked six questions, from which we recorded accuracy and time data.

Each question in the learning effect phase and the performance phase was multiple choice. Before we can describe the three phases, we need to consider the information that was to be conveyed by the performance phase diagrams.

3.1 Information to be Conveyed

Concept diagrams are an expressive logic, capable of defining a broad range of axioms. It is not feasible, or even possible, to cover the rich variety of axioms that one can define using concept diagrams in an empirical study. Given the motivation for developing concept diagrams was to model ontologies, we selected six commonly occurring ontology statement (axiom) types, as was done in a

study into the relative efficacy of OWL and description logic in [1]. This restriction provided controlled variation whilst ensuring the ecological validity of the results. The six selected statement types (to which we assign the names shown in bold) are written here using English, where A and B are classes (sets) and p is a property (binary relation)[5]:

1. **Subset:** All A are B; example: All Selkies are Fairies. This type of statement is used to define class hierarchies in an ontology.
2. **Disjointness:** No A is a/an B; example: No Halfling is an Elf. This type of statement occurs when classes are required to not share individuals.
3. **Only:** A p only B; example: Selkies hate only Goblins. This type of statement is used to place a restriction on a property, p: viewing p as a binary relation, if the domain of p is restricted to A, its image must be a subset of B.
4. **Some:** A p at least one B; example: Elves chase at least one Fairy. This type of statement is used to define features of individuals that lie in A.
5. **Domain:** Only A p things; example: Only Demons scare things. This type of statement identifies the domain of a property (binary relation).
6. **Range:** Things p only A; example: Things annoy only Fairies. This type of statement identifies the range of a property.

In what follows, we always write the six types of statements following the conventions illustrated in the examples just given. To generate the eight diagrams needed for the learning effect and performance phases, we needed a systematic approach to selecting the information that they would convey, reflecting the six statement types. Now, Alharbi et al. designed a study to compare sentences expressing these six statement types, focusing on description logic and the Manchester OWL syntax [1]. With their permission, we adapted their study materials for our purposes[6]. Their study used eight sets of 14 statements; we used each set of statements to produce a single diagram, one for each question. Each set of statements had ten named sets, eight binary relations, four Subset statements and two of each of the other types of statements; more Subset statements were needed since they were necessary for what we call *indirect* statements to be derived; for an indirect statement, see section 3.3, figure 10, where we need to use the information that Ogre is a subset of Enchanter to deduce that Ogres guide only Nisses. An example of a diagram representing 14 statements can be seen in figure 5.

3.2 Colour Treatments

To test whether multiple colour use in concept diagrams brings significant performance benefits over other colour treatments, we identified three different ways of

[5] We acknowledge the blurring between syntax and semantics here; strictly speaking, A and B are monadic predicates and p is a dyadic predicate.

[6] Whilst [1] reports on OWL and DL, their study also included a third treatment: concept diagrams. None of the diagrams used in our studies were syntactically identical to Alharbi et al.'s diagrams; we adjusted the layouts and represented *Some* statements differently. Our training material was not the same as that provided by Alharbi et al., in part since we followed a crowdsourced approach.

assigning colour. We used colourbrewer [9] to define our colours ensuring suitability for visualizing qualitative or categorical information rather than sequential or diverging information:
Monochrome: all syntactic elements are coloured the same. We chose to use black, which is often employed by Euler diagram users, see figures 1 to 4.
Dichrome: two colours are used for the syntactic elements. We chose to use blue with green arrows, the de-facto standard for concept diagrams [11, 12, 21], see figures 6 to 13.
Polychrome: each set and binary relation takes a unique colour hue, see figure 5. Our tasks involved ten sets and eight binary relations, so we needed 18 colours in total. Colourbrewer can only generate sets of up to 12 colours. We generated a set of 10 colours for the named sets and a disjoint set of eight colours for the arrows. Unlabelled curves that were arrow targets took the same colour as their targeting arrow.

3.3 Training Diagrams and Explanations

It was necessary to provide participants with training in the semantics of concept diagrams. We chose to use a sequence of syntactically simple diagrams to explain how the diagrams expressed the six statement types. It should be noted that the training across participant groups differed only due to the nature of the treatment to which they were exposed. For each statement type, we included two training diagrams, which we call a *direct* version and an *indirect* version. The direct version corresponds to information that would naturally be expressed by a single axiom. For example, *All Selkies are Nymphs* and *All Nymphs are Fairies* are expressed by the diagram in figure 5 (see the bottom right box). Indirect statements correspond to information that would normally need to be *inferred* from axioms but which is readily visible in a diagram. Using the two statements *All Selkies are Nymphs* and *All Nymphs are Fairies* as a textual example, one can *infer* the *indirect* statement *All Selkies are Fairies*. Referring again to figure 5, *All Selkies are Fairies* is naturally expressed by the diagram via circle containment, by virtue of expressing the two statements from which it can be inferred. We now explain the training provided.

Subset Statements Participants were exposed to two subset training diagrams, with the direct version being shown in figure 6. The meaning of the diagram was stated using the convention illustrated in section 3.1, as can be seen in figure 6. The indirect subset training diagram can be seen in figure 7. In the remaining parts of this subsection, we omit the (simpler) training given for the *direct* statements which adopted a style similar to that illustrated here.

Disjointness Statements Figure 8 shows the indirect disjointness training diagram, which conveys information by the presence of the curves within a *single* box. An important feature of concept diagrams is the use of *multiple* enclosing boxes. This allows a less cluttered representation of classes when they are not known to be disjoint [13]; high levels of clutter leads to less effective diagrams [3, 11]. It was important to train the participants that diagrams exploit distinct

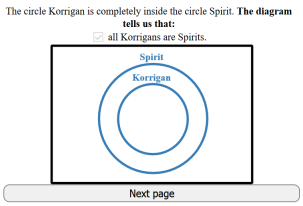

Fig. 6. Training for *direct* Subsets.

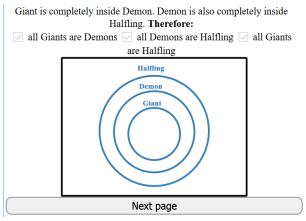

Fig. 7. Training for *indirect* Subsets.

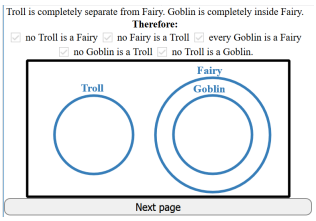

Fig. 8. Training for *indirect* Disjointness.

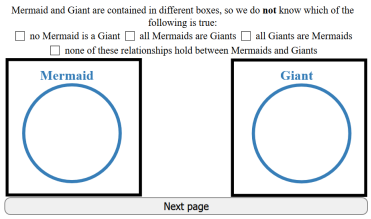

Fig. 9. Training for multiple box use.

boxes in order to avoid expressing a relationship between the represented classes. Figure 9 shows how this was done.

Only Statements Indirect statements in the case of Only axioms can arise in two ways, depending on the source and the target of the arrow. Referring to figure 10, focusing on the arrow source, we see that Ogre is a subset of Enchanter and, since Enchanters guide only Nisses, we can infer that Ogres guide only Nisses. Regarding the target, since Nisses are Demons, we can infer that Enchanters guide only Demons. These indirect statements are perhaps less obvious than those we saw for subset and disjointness statements. In the source case this is, in part, because there is no arrow emanating from Ogre.

Some Statements Indirect statements in the case of Some axioms can also arise in two ways, depending on the source and the target of the arrow. Focusing on the arrow source in figure 11, we see that Ogre is a subset of Giant and, since Giants like at least one Halfling, we can infer that Ogres like at least one Halfling. Regarding the target, since all Halflings are Mermaids, we can infer that Giants like at least one Mermaid. These indirect statements are, as in the case of Only statements, perhaps less obvious than those we saw in the case of subset and disjointness.

Domain Statements Indirect statements in the case of Domain axioms can arise in one way, from the target of the arrow; we do not get any indirect Domain statements arising from the source, since Domain axioms are always defined over the universal set which is represented by the enclosing box. In figure 12, Goblin is a subset of Fairy and, since only Goblins track things, we can infer that only Fairies track things; that is, if the domain of tracks is Goblin and all Goblins are Fairies then the domain can also be viewed as Fairy.

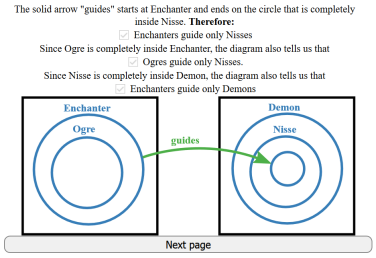

Fig. 10. Training for *indirect* Only.

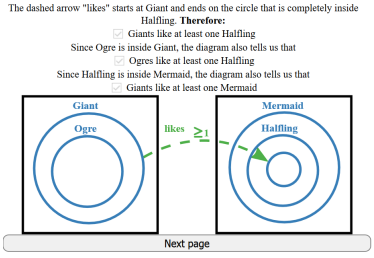

Fig. 11. Training for *indirect* Some.

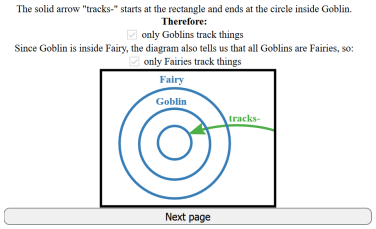

Fig. 12. Training for *indirect* Domain.

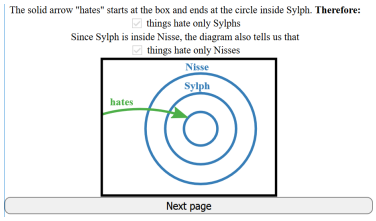

Fig. 13. Training for *indirect* Range.

Range Statements Indirect statements in the case of Range axioms can also arise in one way, depending on the target of the arrow; we do not get any indirect Range statements arising from the source, since Range axioms are always defined over the universal set via arrows sourced on the enclosing box. In figure 13, Sylph is a subset of Nisse and, since things hate only Sylphs, we can infer that things hate only Nisses; that is, if the range of tracks is Sylph and all Sylphs are Nisses then the range can also be viewed as Nisse.

3.4 Learning Effect Questions

Recall, from section 3.1, each of the six main phase tasks was derived from a set of 14 statements. This meant that the diagrams used in the performance phase were syntactically more complex and more expressive than the relatively simple training diagrams (for example, contrast figures 1 and 5). Therefore, two questions were included to reduce the impact of any learning effect that may be present. The diagrams for these questions were derived from Alharbi et al.'s two sets of 14 statements used to train participants in their study [1], which are different from those used in the main phase tasks. These two questions were associated with ten checkboxes of which *seven* should be selected. In total, this gave 14 correct answers across the two questions, one for each of the six direct statements, one for each of the Subset, Disjoint, Range and Domain indirect statements, and two each for the Only and Some indirect statements; for Only and Some indirect, there were two variants of true statement depending on whether the arrow source or target was used to make the derivation. This left

six false statements, across the two questions, one for each of the six statement types. This ensured that participants had been exposed to each type of checkbox (direct, indirect and false) for each statement type before performance data was gathered and the two ways in which Only and Some indirect statements could arise. The participants were unaware that the data collected for these two questions would not be used in the analysis. See table 1 for an illustration.

3.5 Performance Phase Questions

Given the six sets of 14 statements, from which diagrams were derived for the performance phase, we needed to identify suitable sets of checkbox responses for each of them. Again following [1], each type of textual statement occurred as a correct answer six times. This meant we needed 36 statements which appeared as correct answers. For each type of textual statement, we included three *direct* versions of the task and three *indirect* versions. As we are also interested in ensuring that people do not read incorrect information from diagrams, we also included each type of statement as an incorrect answer three times. This gave a total of 54 checkboxes, which were distributed across the six tasks. Thus, each task was associated with nine checkboxes; six of the statements were correct and three were incorrect.

3.6 Data Collection Method

We adopted a between group design. Participants were randomly assigned to a group and were paid £3.25 for their participation. Prolific Academic was used to crowdsource participants from the general population. It is recognised that in crowdsourced studies, participants do not always give questions their full attention, or have difficulties with the language used, and this is hard to control [7]. Varying techniques can be employed for avoiding the recruitment of participants who may have issues with the language or do not give questions their full attention. We chose to limit the participant pool to those who are fluent in English, as well as including other pre-screening criteria covered in section 4. We also

Option	**Checkbox Type**	**Checkmark**
Elves chase at least one Fairy	Some – direct	✓
All Selkies are Fairies	Subset – indirect	✓
Things guide only Elves	Range - false	
Selkies hate only Goblins	Only – indirect	✓
All Boggarts are Midgets	Subset – false	
No Halfling is an Elf	Disjoint – indirect	✓
Only Demons scare things	Domain – direct	✓
Pucks follow only Goblins	Only – false	
Things annoy only Fairies	Range – direct	✓

Table 1. Nine options, six of which are correct.

included two questions, designed to catch *inattentive* participants, that were trivial to answer if the associated text was read. Answering either of these two questions incorrectly meant that the participant was classified as inattentive. The first of these questions was included in the training phase and the second one was in the performance phase. The inattentive participants were unable to proceed with the study as soon as they answered one of these two questions incorrectly and any data collected up to that point was not retained.

In each phase of the study, each diagram and its associated question was displayed on a unique page. Participants could not return to pages and subsequent pages were not revealed until the 'Next page' button was clicked. The training pages were presented in a fixed order for each participant and, as just indicated, included one of the inattentive questions. The order was: Subset, Subset – indirect, Disjoint, Disjoint – indirect, Disjoint – multiple boxes, Only, Only – indirect, *inattentive question*, Some, Some – indirect, Domain, Domain – indirect, Range, Range – indirect. It was felt that this order began training people using simpler concepts since there were no arrows in the first five diagrams. Once the participant clicked the next button, they were asked to answer the two questions included to reduce learning effect. After that, the next two questions were randomly selected from the six performance phase questions, followed by the second inattentive question, and then the remaining four performance phase questions in a random order.

3.7 Statistical Analysis Method

We view accuracy 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. If no significant accuracy difference exists and performance is significantly quicker then the quicker notation is judged to be more effective. For the analysis, we employed two local odds ratios generalized equation models [23] to analyse the *accuracy* data. For the *time* data, we used a generalized estimation model [14] that allowed us to estimate whether the time taken to provide answers was significantly different. Alternative models such as ANOVA were not deemed appropriate as our data violated their assumptions.

4 Study Execution and Statistical Analysis

Here we describe our pilot study before presenting the statistical analysis yielding an overall comparison between treatment types and a comparison by task type.

4.1 Pilot Study

When running a pilot study, we pre-screened participants. Pre-screening criteria included having a Prolific approval rate of 95% or higher, the requirement to have completed at least 5 studies on Prolific previously, being fluent in English, and being aged between 18 and 100 (this was imposed by Prolific). This left a

pool of 27561 potential participants, out of 63577, so over half were disqualified. Further, we indicated that the study would be supported on desktop and tablet devices, but not mobile devices; Prolific does not guarantee that this means participants will refrain from using a mobile device. A total of 39 participants began the pilot study. Of these, two were classified as inattentive, three timed-out after 50 minutes, a further three withdrew before completion, and for one did not have their data saved due to a read/write error. This left us with data from 30 participants.

The overall accuracy rate and the average time to answer each performance phase question during the pilot study are given in table 2, accompanied by a breakdown for each group. The pilot data do not indicate a ceiling or floor effect, which would have suggested that the tasks are either too easy or too hard to reveal significant differences. That is, the pilot data suggest that the tasks require some cognitive effort to perform but are not so difficult that the participants are essentially guessing the answers.

We noted a very high error rate for *Disjoint - false* checkboxes: overall, the *Disjoint - false* accuracy rate was 27.78%, with the Monochrome and Polychrome groups both scoring 23.33% and the Dichrome group getting 36.67% correct. Ticking a *Disjoint - false* checkbox would suggest that the participant believed two sets are disjoint when in fact they are not. This could be due to misunderstanding the information provided by multiple rectangle use and added further text to the associated training page: "In particular, because Mermaid and Giant are in different rectangles, the diagram **does not tell** us that no Mermaid is a Giant and **does not tell us** that no Giant is a Mermaid." We also observed that the accuracy rates were low overall for *Domain* (50.00%) and *Domain - indirect* questions (37.78%) but we suspected that this was due to the difficulty of understanding the use of inverse. As such no change was made to the study based on this observation.

4.2 Main Study

For the main study, we included an additional pre-screening criterion: no participant who took part in the pilot could take part in the main study. Three participants self-reported as colourblind, one in each group. The statistical analysis is performed on the entire data set; we did not perform a subsequent analysis with the colourblind participants removed. The accuracy rates and mean times are summarised in table 3. Whilst the accuracy rates and mean times can be seen as an *indicator* of relative performance across groups, it is important to note

Group	No. of Participants	Accuracy Rate	Mean Time
Overall	30	81.17%	2m 52.98s
M	10	85.37%	3m 11.10s
D	10	80.37%	2m 46.66s
P	10	77.78%	2m 38.18s

Table 2. Summary of the pilot data.

Group	No. of Participants	Accuracy Rate	Mean Time
Overall	261	80.11%	2m 48.17s
M	81	77.98%	2m 52.81s
D	89	80.96%	2m 54.74s
P	91	81.16%	2m 37.60s

Table 3. Main study performance data.

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.

Learning Effect Questions We evaluated whether the learning effect questions yielded significantly lower accuracy performance than the six performance phase questions. Based on a Wald test, the learning effect questions had statistically significant lower overall accuracy rate than the performance phase questions (0.8011 vs. 0.7571 with p-value < 0.001. This suggests that participants did improve their accuracy performance during these first two questions. This does not, however, mean that the learning effect was eliminated but suggests that it was reduced by the inclusion of these two questions. It is not appropriate to compare the times taken for the learning effect questions with the performance phase questions due to their differing number of checkboxes (ten versus nine).

Statistical Analysis: Overall Comparison Here we report on the overall comparison between the three colour treatment groups. Using a Generalised Estimating Equations (GEE) based [23] 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. Recall that a correct answer means either correctly ticking a checkbox (where the associated statement is true) or correctly not ticking a checkbox (where the associated statement is false). We computed p-values to determine whether the treatments gave rise to significantly different accuracy performance. The estimated odds of correctly answering questions with Dichrome was 1.20 (to 2d.p.) times that of Monochrome with a 95% CI of (0.94,1.53) and p-value of 0.1363 (to 4d.p.). Therefore, there was no significant difference in accuracy performance between Dichrome and Monochrome. Results for the other pairwise comparisons are given in table 4: there wer no significant differences overall in accuracy across treatments.

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 complete a task with one treatment compared to another. The derived CI and its corresponding p-value allowed us to determine whether two treatments were significantly different. The model estimated that the time needed to complete a task with Dichrome was 1.11 times (2d.p.) that with Monochrome with a 95% CI of $(0.89, 1.37)$ and p-value of 0.3553. Therefore, there is no significant difference in time performance between Dichrome and Monochrome. Results for the other pairwise comparisons are given in table 5. The analysis revealed no significant differences overall in time taken across the three treatments. Therefore, our accuracy and time analysis consistently indicate that there is no *overall difference* in the three colour treatments: *the overall ranking is Monochrome = Dichrome = Polychrome.*

Statistical Analysis: Comparison by Checkbox Type When seeking to establish whether significant performance differences exist for each of the three

Treatments	Odds	CI	p-value
D versus M	1.20	(0.94,1.53)	0.1363
D versus P	0.99	(0.78,1.24)	0.9145
M versus P	0.82	(0.64,1.05)	0.1174

Table 4. Overall comparison: accuracy.

Treatments	Ratio	CI	p-value
D versus M	1.11	(0.89,1.37)	0.3553
D versus P	1.12	(0.92,1.36)	0.2606
M versus P	1.01	(0.82,1.24)	0.9179

Table 5. Overall comparison: time.

Treatments	Odds	CI	p-value	Most Accurate
Only – direct				
Dichrome versus Polychrome	0.58	(0.33,0.99)	0.0474	Polychrome
Only – indirect				
Dichrome versus Polychrome	0.57	(0.33,0.98)	0.0436	Polychrome
Some – false				
Monochrome versus Dichrome	2.15	(1.05,4.41)	0.0364	Monochrome
Monochrome versus Polychrome	2.09	(1.04,4.22)	0.0391	Monochrome

Table 6. Comparison of Treatments by Accuracy by Statement Type.

variants (direct, indirect, false) of each of the six statement types, we can only consider the accuracy data as it was not meaningful to collect time date for individual checkboxes. As with the overall analysis, we produced a GEE based statistical model. Results for the pairwise comparisons where significant differences were observed are given in table 6. We can see from the significant results that in two cases Dichrome is significantly less accurate than Polychrome. In the two other cases, Monochrome is significantly more accurate than both Dichrome and Polychrome.

5 Discussion

The results from this study suggest that the use of multiple colours when drawing concept diagrams does not, in general, significantly enhance task performance. We did observe that for *Only* and *Only – indirect* statements, Polychrome outperformed Dichrome. In these cases, the correct response would have been to select the associated checkbox. These results indicate that Dichrome did not facilitate the extraction of the respective information as well as Polychrome. Surprisingly, however, we also observed that using black curves and arrows was significantly more effective in the case of *Some – false* tasks; these are tasks where the response is incorrect if the associated checkbox is ticked. These results indicate that the Dichrome and Polychrome treatments, which were statistically indistinguishable from each other, did not facilitate the extraction of the respective information as well as Monochrome. Of course, it would be remiss not to remark on the fact that we only had four significant results out of the 60 statistical tests conducted. One would expect to obtain three type-I errors when conducting this number of tests, at the 5% level. Thus, we cannot say with any confidence that any of the treatments significantly differ.

It is particularly surprisingly that we obtained notable evidence that the choice of colour treatment made no difference for those tasks with checkboxes whose associated statements only involved sets (i.e Subset or Disjoint statements). Here, we expected the polychrome treatment to yield superior task performance because only the information conveyed by the underlying Euler diagrams was necessary for the task. However our observations suggest that, for diagrams with a high level of complexity, the effectiveness of multiple colours for the curves in the underlying Euler diagram is indeed diminished; the polychrome treatment required 18 colours. Perceptual theory suggests that up to eight to ten colours can be rapidly distinguished by the human eye, after which the perceptual distinction between these colours diminishes [10]. Therefore, we posit that a reduction in the ability to easily distinguish between 10 or more colours has compromised the efficacy that we otherwise anticipated using the polychrome treatment. Moreover, efficacy is compromised to such an extent that there was no benefit of using multiple colours, compared to Monochrome, and sometimes performance was actually inferior (noting the caveat concerning type-I errors).

In the case of Monochrome, graphical shape is the only differentiator between syntactic elements that represents sets (circles) and those which represents binary relations (arrows). Similarity theory [8] indicates that using different syntactic devices for semantically different entities is sensible: using syntactically similar entities leads to increased search times when seeking to find a particular piece of 'target' syntax. Thus, in the Monochrome case, shape is the only graphical property that may aid a visual search through the diagram when seeking to establish the truth of a given statement. Now, colour can also be used to group syntactic devices that have some semantic commonality, as seen in Dichrome treatment: colours are assigned to syntactic items that represent semantically different types of things: blue curves represent sets and green arrows represent binary relations. In this sense, colour is being used to reinforce the semantic differentiation of syntactic devices via shapes when performing visual search. Thus, we see that using two colours or two shapes has the potential to aid information extraction, with the Dichrome treatment exploiting both and the monochrome treatment exploiting only shape. It is known that colour is a more salient graphical property than form [6], indicating that the use of two colours may be more beneficial than just the use of different shapes. However, our study suggests that using two shapes and two colours (Dichrome) is *not more effective* than using two shapes and just black (Monochrome). We posit that circles and arrows have *sufficiently different visual characteristics* meaning that the additional graphical element of colour does not bring about performance benefits.

6 Conclusion

Based on prior research into Euler diagrams [4], there was evidence to suggest that manipulating colour in concept diagrams had the potential to impact user task performance. However, the case was not clear-cut: concept diagrams express more complex statements than Euler diagrams, exploiting a more diverse set of

graphical symbols with which to make statements. Indeed, being designed for use in ontology engineering, the kinds of statements that concept diagrams express can require *each diagram* to include many syntactic elements, as in figure 5. These facts suggested that further empirical insight was required in order to understand the role that colour plays in the effective interpretation of concept diagrams. Our study suggests that colour is no longer a useful visual variable to manipulate when seeking to improve user task performance, at least for the kinds of tasks we have evaluated. The take-away message from our study is that, whilst colour is a useful graphical property to manipulate for Euler diagrams, the benefits may be lost in the case of concept diagrams.

The discussion in section 5 alludes to the fact that one reason using two colours may not yield performance benefits – when used consistently with graphical shape to segregate syntactic elements that have differing semantic properties – is that circles and arrows have particularly different shapes. Thus, our research raises an important question: *when using colour to visually group syntactic elements that have a common semantic property, how different do the elements' shapes need to be in order for there to be performance benefits arising from using colours?* This question is not relevant for just concept diagrams, but all diagrammatic notations that employ a range of graphical shapes to convey information.

Acknowledgements This research was partially funded by a Leverhulme Trust Research Project Grant (RPG- 2016-082) for the project entitled Accessible Reasoning with Diagrams. Thanks to Eisa Alharbi for supplying experimental materials, associated with [1], on which some of our training materials, learning effect and performance phase questions were based.

References

1. Alharbi, E., Howse, J., Stapleton, G., Hamie, A., Touloumis, A.: The efficacy of OWL and DL on user understanding of axioms and their entailments. In: 16th International Semantic Web Conference, pp. 20–36. Springer (2017)
2. Alharbi, E., Howse, J., Stapleton, G., Hamie, A., Touloumis, A.: Visual logics help people: An evaluation of diagrammatic, textual and symbolic notations. In: IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 255–259. IEEE (2017)
3. Alqadah, M., Stapleton, G., Howse, J., Chapman, P.: Evaluating the impact of clutter in Euler diagrams. In: 8th International Conference on the Theory and Application of Diagrams, pp. 108–122. Springer (2014)
4. Blake, A., Stapleton, G., Rodgers, P., Howse, J.: How should we use colour in Euler diagrams. In: 7th International Symposium on Visual Information Communication and Interaction. ACM (2014)
5. Blake, A., Stapleton, G., Rodgers, P., Howse, J.: The impact of topological and graphical choices on the perception of Euler diagrams. Information Sciences **330**, 455–482 (2016)
6. Callaghan, C.: Interference and dominance in texture segregation: Hue, geometric form, and line orientation p. 299?311 (1989)

7. Chen, J., Menezes, N., Bradley, A., North, T.: Opportunities for crowdsourcing research on Amazon Mechanical Turk. Human Factors **5**(3) (2011)
8. Duncan, J., Humphreys, G.: Visual search and stimulus similarity. Psychological review **96**(3), 433 (1989)
9. Harrower, M., C.Brewer: ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. The Cartographic Journal **40**(1), 27–37 (2003), accessed May 2019
10. Healey, C.: Choosing effective colours for data visualization. In: 7th Conference on Visualization. pp. 263–ff. IEEE (1996)
11. Hou, T., Chapman, P., Blake, A.: Antipattern comprehension: An empirical evaluation. In: 9th International Conference on Formal Ontology in Information Systems. pp. 211–224 (2016)
12. Howse, J., Stapleton, G., Taylor, K., Chapman, P.: Visualizing ontologies: A case study. In: International Semantic Web Conference. pp. 257–272. Springer (2011)
13. John, C., Fish, A., Howse, J., Taylor, J.: Exploring the notion of clutter in Euler diagrams. In: 4th International Conference on the Theory and Application of Diagrams. pp. 267–282. Springer, Stanford, USA (2006)
14. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)
15. Matsin, L.: Short-term (working) memory. http://www.human-memory.net/types_short.html (2010)
16. Mineshima, K., Sato, Y., Takemura, R., Okada, M.: How diagrams can support syllogistic reasoning: An empirical study. Journal of Visual Languages and Computing **25**, 159–169 (2014)
17. Moody, D.: The "physics" of notations: Toward a scientific basis for constructing visual notations in software engineering. IEEE Transactions on Software Engineering **35**(6), 756–779 (2009)
18. Rodgers, P., , Zhang, L., Purchase, H.: Wellformedness properties in Euler diagrams: Which should be used? IEEE Transactions on Visualization and Computer Graphics **18**(7), 1089–1100 (2012)
19. Sato, Y., Masuda, S., Someya, Y., Tsujii, T., Watanabe, S.: An fMRI analysis of the efficacy of Euler diagrams in logical reasoning. In: IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 143–151. IEEE (2015)
20. Sato, Y., Mineshima, K.: How diagrams can support syllogistic reasoning: An empirical study. Journal of Logic, Language and Information **24**, 409–456 (2015)
21. Stapleton, G., Howse, J., Chapman, P., Delaney, A., Burton, J., Oliver, I.: Formalizing concept diagrams. In: 19th International Conference on Distributed Multimedia Systems, pp. 182–187. KSI (2013)
22. Stenning, K., Cox, R., Oberlander, J.: Contrasting the cognitive effects of graphical and sentential logic teaching: Reasoning, representation and individual differences. Language and Cognitive Processes **10**, 333–354 (1995)
23. Touloumis, A., Agresti, A., Kateri, M.: Generalized estimating equations for multinomial responses using a local odds ratios parameterization. Biometrics **69**(3), 633–640 (2013)
24. Ware, C.: Information Visualization: Perception for Design, chap. 6.1: Gestalt Laws, pp. 189 – 205. Morgan Kaufmann Publishers Inc., 2nd edn. (2004)