Measuring User Comprehension of Inference Rules in Euler Diagrams

Sven Linker(✉), Jim Burton, and Andrew Blake

Visual Modeling Group, University of Brighton, Brighton, UK
{s.linker,j.burton,a.l.blake}@brighton.ac.uk

Abstract. Proofs created by diagrammatic theorem provers are not designed with human readers in mind. We say that one proof, P_1, is more "readable" than another, P_2, if users make fewer errors in understanding which inference rules were applied in P_1 than in P_2, and do so in a shorter time. We analysed the readability of individual rules in an empirical study which required users to identify the rules used in inferences. We found that increased clutter (redundant syntax) in the premiss diagrams affects readability, and that rule applications which require the user to combine information from several diagrams are sometimes less readable than those which focus on a single diagram. We provide an explanation based on mental models.

1 Introduction

Interactive and automated theorem proving with diagrams has been explored in systems such as Speedith [6]. However, existing tools do not take into account the growing body of research on what the specific cognitive advantages of reasoning diagrammatically might be, and on where the source of these advantages, if they exist, might be located. This research includes neurological studies that examine brain activity of users reasoning with and without diagrams [5], and empirically-derived guidelines for producing diagrams that make good use of Gestalt principles relating to colour and form [2]. At the broadest level our research question asks *is it possible to develop a systematic understanding of readability in diagrammatic proofs?* We use the term "readable" to mean relatively easy to understand, and will use error rates and response times of users who read the proof as measures of relative readability.

Euler diagrams have been used as a formal logic since the 1990s. Figure 1 shows a theorem expressed using Euler diagrams, equivalent to the expression $B \subseteq A \wedge C - B = \emptyset \Rightarrow C \subseteq A$. In order to prove that this theorem is true, we need to apply inference rules which add and remove elements from the diagrams labelled 1 and 2 until we produce diagram 3.

In this paper we describe an empirical study in which we analyse and measure the factors that affect comprehension of individual inference steps when reasoning with Euler diagrams. The study measures the number of errors and the time taken to answer a series of questions about Euler inference rule applications.

M. Jamnik et al. (Eds.): Diagrams 2016, LNAI 9781, pp. 32–39, 2016.
DOI: 10.1007/978-3-319-42333-3_3

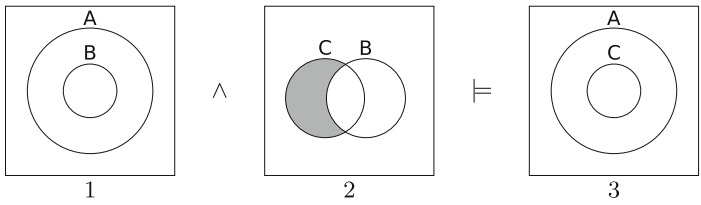


Fig. 1. An Euler diagram theorem

The present work is intended as a first step towards a notion of readability for whole proofs with Euler diagrams. To the best of our knowledge, there have been relatively few empirical studies of reasoning with Euler diagrams, e.g. the work of Sato and Mineshima such as [5]. These studies are concerned with the activity of proving itself, i.e., the creation of a proof, while we are interested in the task of understanding of a proof which already exists.

In the following section we give a brief explanation of Euler diagrams and their use in reasoning. In Sect. 3 we describe our experimental design, then in Sects. 4 and 5 we present our analysis and interpretation of the results. The training materials, questions and the (anonymised) data which was collected are available from our website, http://readableproofs.org/readability-study.

2 Euler Diagrams

Unitary Euler diagrams are drawn within a bounding rectangle representing the universe of discourse. Sets are represented by labelled *contours* (or *curves*) drawn within the rectangle. Topological relations between the curves specify the relations between sets. The curves divide the space within the diagram into *zones*. Zones may be shaded or non-shaded. The set represented by a shaded zone must be empty.

Unitary Euler diagrams can be composed to create *compound diagrams*. Within this paper, we will only use conjunction to compose diagrams.

The inference rules we consider are the following: 1. Erase Contour (EC), 2. Erase Shading (ES), 3. Combine (CO), 4. Copy Contour (CC), and 5. Copy Shading (CS). The effects of the rules are as follows. *Erase Contour* removes a contour from a unitary diagram. If this contour was separating a shaded zone from a non-shaded zone, the unified zone in the result will be non-shaded. *Erase Shading* removes the shading of a single zone from a unitary diagram. Note that rules 1 and 2 make changes to a single unitary diagram. We call these the *simple rules*. When using rule 3 to *combine* two unitary diagrams, both diagrams must contain the same set of zones. In the result, these two diagrams are replaced by a single diagram with the same set of zones and in which a zone is shaded if and only if it is shaded in one of the origin diagrams. For the copy rules, we have to identify which zones in different diagrams *correspond* [3] to each other. *Copy Contour* can be used to copy a contour c_1 from a unitary diagram d_1 to a second unitary diagram, d_2, respecting the topological information within d_1 about c_1

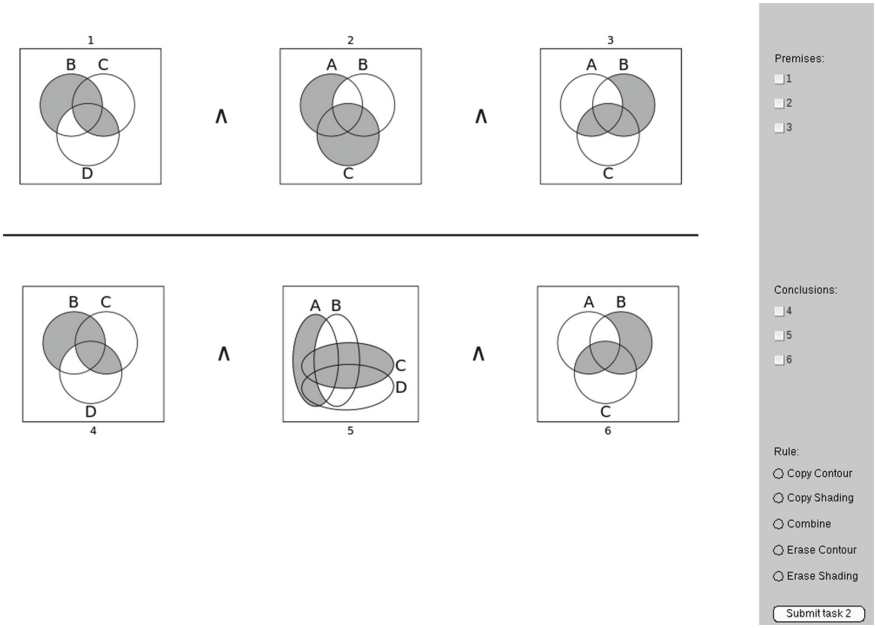


Fig. 2. Rule application: Copy Contour (cluttered version)

and the contours contained in both diagrams. Our last rule is *Copy Shading*: if a zone z_1 is not shaded in d_1 and corresponds to a shaded zone z_2 in d_2, then *Copy Shading* can be used to shade z_1. Rules 3, 4 and 5 depend on information from two unitary diagrams in the premiss. We call these rules *complex*.

An important notion for our work is the *clutter* of a diagram [1], which we measure by the *contour score* of a diagram. First, the contour score of a single zone is the number of contours it is enclosed in. The contour score of a unitary diagram is the sum of all contour scores of the zones present in the diagram. For example, within Fig. 2, all diagrams in the premiss have a contour score of 12, while diagram 5 has a score of 30.

3 Experimental Design

Our research questions are as follows: first, does the amount of clutter in the diagrams have an effect on the identification of rule applications? Secondly, are applications of complex rules less readable than applications of simple rules?

We designed our experimental tasks by first creating two semantical situations, being the relationships between four sets named A, B, C and D, chosen so as to ensure that all rules can be applied to such a situation. We constructed two compound diagrams representing each situation, with *high* and *low* clutter respectively. Each compound diagram consists of a conjunction of three unitary

diagrams. From each of the four premises we constructed an instance of an application of each of the five rules, resulting in 20 such tasks for a within-group study. Given an application, the participants' task is to identify which rule has been applied, which unitary diagrams comprise its input, and which unitary diagram in the conclusion is the result of applying the rule.

The cluttered versions of the premises were created by using Venn-form diagrams. Diagrams were drawn according to well-formedness principles and using consistent font, font weight, line width and so on. All parts of the study took place in a specialised usability lab using the same equipment.

The study consisted of a paper-based introduction to Euler diagrams and the inference rules, a training phase and the main study. Within the training phase, the instructor actively worked with the participants to address misunderstandings and misconceptions directly as they arose. The participants had access to a "cheatsheet" containing an exemplary application of each rule with the answers the participants had to provide.

After the training phase, participants were presented with the 20 tasks that form the main study in a randomized order (see Fig. 2). For each question, the participants had to identify the rule that had been applied, the diagram(s) that rule had been applied to, and the diagram that was changed in the conclusion. The program recorded the answers of the participants as well as the time taken to finish the task. For the main study we recruited 30 undergraduate participants, 23 male and 7 female, from the ages of 18 to 34.

4 Analysis

Error Analysis. We excluded nine data points where the participant did not come to an answer within the time limit of 120 s.

Table 1. Errors: clutter

Clutter	Errors	Correct
High	69	226
Low	56	240

Table 2. Errors: type of rule

Rule	Errors	Correct
Combine	18	102
Copy Contour	28	89
Copy Shading	53	62
Erase Contour	6	114
Erase Shading	20	99

Table 1 shows errors aggregated by clutter. A Chi-square test reveals no significant differences for clutter. Table 2 shows the number of errors according to the rules used in the tasks. A Chi-square test shows that there is a significant difference between some pairs in the set ($\chi^2(df = 4, N = 591) = 66.26, p < 0.05$). We used the Chi-square test for all pairs in this set to find the significantly different entries with confidence of $p < 0.001$.

The results of these tests are shown in Table 3. Participants performed significantly worse for *Copy Shading*. The difference between *Copy Contour* and *Erase Contour* is also significant.

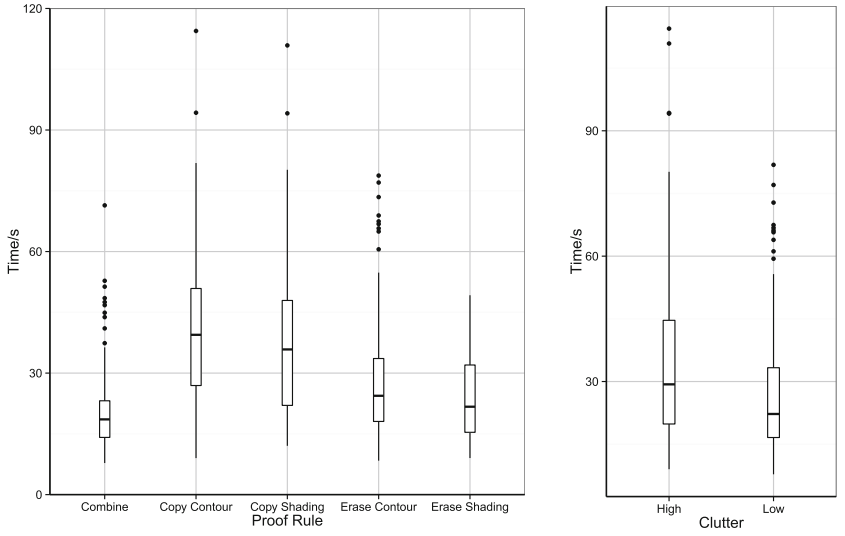
Table 3. Pairwise comparisons of rules (errors)

	CC	CS	EC	ES
CO	$\chi^2(1, 237) = 2.48/$ $p = 0.116$	$\chi^2(1, 235) = 25.46/$ $p < 0.001$	$\chi^2(1, 240) = 5.60/$ $p = 0.018$	$\chi^2(1, 239) = 0.04/$ $p = 0.838$
CC	-	$\chi^2(1, 232) = 11.57/$ $p < 0.001$	$\chi^2(1, 237) = 15.77/$ $p < 0.001$	$\chi^2(1, 236) = 1.43/$ $p = 0.231$
CS		-	$\chi^2(1, 235) = 50.56/$ $p < 0.001$	$\chi^2(1, 234) = 22.02/$ $p < 0.001$
EC			-	$\chi^2(1, 239) = 7.42/$ $p = 0.006$

Time Analysis. In this analysis we removed all errors, since we are interested in the time the participants needed to perform the tasks correctly, reducing our dataset to 466 data points. We distinguish the means according to the amount of clutter and according to the type of rule. Figures 3a and b show the interquartile ranges of the performance times. Participants took longest to identify *Copy Contour* and *Copy Shading*, while *Combine* has the lowest median.

Due to the removal of the errors and timeouts, our data is unbalanced. Hence, we analysed the set by using a RM-ANOVA test, comparing the performance time for clutter and the type of rules. Clutter levels had a significant impact on performance time ($F(1, 19) = 37.83, p < 0.05$). The effect size is $d = 0.45$, which corresponds to a percentile of 66 %–69 %. That is, approximately 2/3 of participants were, on average, faster at completing the tasks with low clutter than the average person completing the high clutter tasks.

Furthermore, the RM-ANOVA showed that a significant difference exists between at least one pair of rules ($F(4, 19) = 36.97, p < 0.05$). The results of a



(a) Rules (b) Clutter

Fig. 3. Performance times

Table 4. RM-ANOVA for pairwise comparisons of rules (time)

	CC	CS	EC	ES
CO	$F(1,19) = 107.46/$ $p < 0.001/$ $d = 1.32/88–92\%$	$F(1,19) = 61.43/$ $p < 0.001/$ $d = 1.07/84–88\%$	$F(1,19) = 16.89/$ $p < 0.001/$ $d = 0.53/69–73\%$	$F(1,19) = 3.66/$ $p = 0.057$
CC	-	$F(1,19) = 1.54/$ $p = 0.217$	$F(1,19) = 34.97/$ $p < 0.001/$ $d = 0.75/76–79\%$	$F(1,19) = 69.37/$ $p < 0.001/$ $d = 1.17/84–88\%$
CS		-	$F(1,19) = 15.51/$ $p < 0.001/$ $d = 0.55/69–73\%$	$F(1,19) = 36.79/$ $p < 0.001/$ $d = 0.92/82–84\%$
EC			-	$F(1,19) = 4.88/$ $p = 0.028$

pairwise analysis, testing for an increased confidence level < 0.001, are shown in Table 4. This table allows us to group the rules into different (not necessarily disjoint) subsets. *Copy Contour* and *Copy Shading* are significantly different from all other rules. While *Erase Contour* and *Combine* are significantly different from each other, the difference to *Erase Shading* is not significant. Hence *Erase Contour* and *Erase Shading* constitute one subset, and *Combine* and *Erase Shading* constitute the last one.

The RM-ANOVA shows that there is significant interaction between the type of rule and the amount of clutter ($F(4,19) = 2.97, p < 0.05$). However, applying the test to pairs of rules and distinguishing the clutter level does not show a significant difference with a confidence < 0.01.

5 Interpretation

While higher amounts of clutter had no significant impact on the error rate, it did in fact increase the time the participants needed to solve the tasks. Our interpretation of this is as follows. Identifying the rules was a difficult task for our non-expert participants, requiring a meticulous analysis of the diagrams. A higher amount of clutter increases the number of single parts within the diagrams they had to look at. This explains the increased time the participants needed to solve the tasks. However, the high levels of concentration they had to maintain prevented them from being distracted by these additional elements, i.e., from making additional errors. Furthermore, in comparison to the diagrams used in the preceeding studies on clutter (e.g. [1]), our diagrams can still be considered to have a low amount of clutter.

The type of the rule had a much stronger impact. Our interpretation adopts the perspective of mental models, assuming that readers create and manipulate internal representations of diagrams [4]. We assume that the experimental tasks required participants to manipulate mental models in ways that correspond with syntactical manipulations of diagrams. By asking the participants to identify the premises and conclusion of a rule, we require them to consider each unitary diagram separately. That is, we expect that the participants create a mental model

of each unitary diagram to analyse. As described by Johnson-Laird, numerous studies have corroborated the prediction that the "the more models we need to take into account to make an inference, the harder the inference should be" [4].

With regard to the performance time, we can group our rules into subsets. The first is the *complex* rules, where the participants needed to identify that information contained within one diagram in the premises was added to another diagram to yield the conclusion. Doing this would require them to inspect each diagram, p, from the premiss to decide whether their mental model of p can be manipulated in ways consistent with their mental model of the conclusion.

Our second subset comprises applications of *simple* rules. To identify these, the participants needed to find the right conclusion, create and manipulate a mental model of the diagram above it (by forgetting a part of the information within) and compare them. This difference in cognitive effort is reflected in the time the participants needed to perform the tasks.

Combine stands out from the other four rules, however, since it alone results in a conclusion containing only two unitary diagrams. Even though the participants would need to compare two diagrams from the premiss to see whether they yield the conclusion, the strong visual difference between the premiss and conclusion makes it obvious that this rule was applied.

6 Conclusion

We presented the results of a study which examined the impact of clutter and differences in inference rules on the ability of participants to identify applications of these rules. The amount of clutter did not significantly influence the number of errors made but did impact significantly on performance time. We found significant differences in performance time between the rules based on the number of unitary diagrams that need to be considered as input to the rule, modulo *Combine*. We attributed these differences to the cognitive effort the participants needed to make while identifying and validating rules applications.

Acknowledgements. This work is supported by EPSRC grant EP/M011763/1.

References

1. Alqadah, M., Stapleton, G., Howse, J., Chapman, P.: Evaluating the impact of clutter in Euler diagrams. In: Dwyer, T., Purchase, H., Delaney, A. (eds.) Diagrams 2014. LNCS, vol. 8578, pp. 108–122. Springer, Heidelberg (2014)
2. Blake, A., Stapleton, G., Rodgers, P., Cheek, L., Howse, J.: Improving user comprehension of Euler diagrams. In: VL/HCC, pp. 189–190. IEEE (2013)
3. Howse, J., Stapleton, G., Flower, J.: Corresponding regions in Euler diagrams. In: Hegarty, M., Meyer, B., Narayanan, N.H. (eds.) Diagrams 2002. LNCS (LNAI), vol. 2317, pp. 76–90. Springer, Heidelberg (2002)
4. Johnson-Laird, P.N.: Mental models and human reasoning. Proc. Nat. Acad. Sci. U.S.A **107**(43), 18243–18250 (2010)
5. Sato, Y., Masuda, S., Someya, Y., Tsujii, T., Watanabe, S.: An fMRI analysis of the efficacy of Euler diagrams in logical reasoning. In: VL/HCC, pp. 143–151. IEEE (2015)
6. Urbas, M., Jamnik, M., Stapleton, G.: Speedith: a reasoner for spider diagrams. J. Log. Lang. Inf. **24**, 1–54 (2015)