In the Hand of the Beholder: Comparing Interactive Proof Visualizations

Christian Alrabbaa[1], Stefan Borgwardt[1], Nina Knieriemen[2], Alisa Kovtunova[1], Anna Milena Rothermel[2], and Frederik Wiehr[2]

[1] Institute of Theoretical Computer Science, TU Dresden, Germany
firstname.lastname@tu-dresden.de

[2] German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany
firstname[_middlename].lastname@dfki.de

Abstract. Although logical inferences are interpretable, actually explaining them to a user is still a challenging task. While sometimes it may be enough to point out the axioms from the ontology that lead to the consequence of interest, more complex inferences require proofs with intermediate steps that the user can follow. Our main hypothesis is that different users need different representations of proofs for optimal understanding. To this end, we undertook some user experiments related to logical proofs. We explored how a user's cognitive abilities influence the performance in and preference for certain proof representations. In particular, we compared tree-shaped representations with linear, text-based ones, and for each we offered an interactive and a static version. After each proof, participants had to solve some tasks measuring their level of understanding and rated each proof according to the perceived comprehensibility. At the end of the questionnaire, subjects ranked the proofs by comprehensibility. We found no differences between the general performance or the subjective ratings of the proof representations. However, in the final ranking participants preferred the conditions with tree-shaped proofs over the textual ones, with significant differences in the rankings of the higher cognitive ability group and across both groups but not in the low cognitive ability group.

1 Introduction

Research on explaining description logic inferences has mostly focused on computing *justifications* [9,21,31], i.e., minimal sets of axioms from which a consequence follows. While justifications are already very helpful for designing or debugging an ontology, depending on the complexity of the inference and the expertise of the user, more detailed proofs of the consequence are needed. Following a line of research on the understandability of description logic inferences and proofs [2–4,10,17,23,26,30,33], in this paper we investigate the usefulness of different proof representations. We compare proofs in a traditional tree shape with a linearized textual representation of proofs. For both variants, we provide an interactive as well as a static version.

Logical (abstract) reasoning is the ability to solve novel problems without task-specific knowledge and a core mechanism of human learning [28]. Logical reasoning is closely related to fundamental cognitive functions. Generally, according to [20], there are two types of intelligence. *Fluid intelligence* involves processing new information and solving novel types of problems, as opposed to those that require *crystallized intelligence*, which entails the application of consolidated knowledge typically acquired in academic settings. In this terminology, *experience* with logic can be classified as an example of crystallized intelligence. And since it is typically neither trained nor taught, logical *ability* is often assessed as part of fluid intelligence tests [12].

Our main goal in this paper is to find out whether a user's logical ability influences which of these four proof representations are preferred and lead to better performance. A study from last year [10] attempted a similar comparison, but it did not find significant differences based on the user's self-reported experience with logic. In this paper, we attempt to measure the user's logical ability level more directly and investigate the impact on performance and preferences of different proof representations. Moreover, we were able to increase the number of participants by creating a fully online survey and making it available at a study participant recruitment platform.

In a first experiment, we verified that the score on the standardized International Cognitive Ability Resource (ICAR) test[1] strongly correlates with the ability to draw logical inferences and understand logical proofs. Based on this insight, we used the ICAR in our second experiment to measure the user's ability levels and compare the different proof representations. To open our experiments to a larger population, we decided not to use the formal DL syntax, but rather hand-crafted textual representations of axioms, similar to the ones used in [1, 34].

2 Related Work

There are various fluid intelligence tests assessing logical ability, e.g. The Stanford Binet Intelligence Scale, Fifth Edition (SB:V) [25], Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV) [13], and nonverbal Raven's Progressive Matrices [32]. We refer the reader to [38] for a comprehensive survey. These tests provide reliable, and well-validated abstract reasoning items [11] and are extensively used in clinical, educational, occupational and research settings [16, 37]. However, these tasks are not free-to-use, and copyright often prevents these pen-and-paper tasks from being adapted into computerized tasks. Notable exceptions here are the freely available web-based Hagen Matrices Test [19] and the public-domain International Cognitive Ability Resource (ICAR) [39].

In parallel, several approaches for converting description logic axioms and proofs into textual representations have been developed and evaluated [1, 6, 29, 33,

[1] `https://icar-project.com/projects/icar-project/wiki`

$$\dfrac{\dfrac{\mathsf{A} \sqsubseteq \exists \mathsf{r}.\top \qquad \mathsf{A} \sqsubseteq \forall \mathsf{r}.(\mathsf{B} \sqcap \mathsf{C})}{\mathsf{A} \sqsubseteq \exists \mathsf{r}.(\mathsf{B} \sqcap \mathsf{C})} \qquad \mathsf{C} \sqcap \mathsf{B} \sqsubseteq \bot}{\mathsf{A} \sqsubseteq \bot} \qquad \begin{aligned} \mathcal{O} = \{\ & \mathsf{A} \sqsubseteq \exists \mathsf{r}.\top, \\ & \mathsf{C} \sqcap \mathsf{B} \sqsubseteq \bot, \\ & \mathsf{A} \sqsubseteq \forall \mathsf{r}.(\mathsf{B} \sqcap \mathsf{C})\ \} \end{aligned}$$

Fig. 1. A proof for the unsatisfiability of A w.r.t. $\mathcal{O}$, i.e., that $\mathcal{O} \models \mathsf{A} \sqsubseteq \bot$.

34]. For example, generation of verbalized explanations for non-trivial derivations in a real world domain was tested on computer scientists in [34]. The authors distinguish short and long textual explanations, but the participants' opinions on conciseness turned out to be mixed and not too strong. In the recent work [10], it has been shown that proofs should not contain too many trivial steps, i.e. be relatively small, not only for a textual but also for a tree-shape representation.

Concerning the representation of logical statements, we take the following studies into the account. First, in [29] it has been confirmed that statements in a controlled natural language are understood significantly better than the Manchester OWL Syntax, where DL axioms are expressed by sentences with the words like "SubTypeOf", "DisjointWith", "HasDomain", etc. Second, the experiment [1] has shown that the Manchester Syntax is not more effective than the formal DL syntax. Therefore, similarly to the approaches [29, 34] and in contrast to [22], in our experiment we do not use the formal DL notation, since that would require readers to learn the syntax of ontology languages or description logics: instead we use patterns to convert DL sentences into natural-language explanations (see Section 3). Further, we arrange these sentences in a tree-shaped representation, similarly to proofs based on consequence-based reasoning procedures [27,36], and we order them in a linear sequence using English connectives, e.g. as produced by various verbalization techniques [6, 29, 30, 34].

3 Proofs

We assume a basic familiarity with DLs, in particular $\mathcal{ALCQ}$ [8]. Let $\mathcal{O}$ be an ontology and α a consequence of $\mathcal{O}$ ($\mathcal{O} \models \alpha$). The next step is to compute *justifications*, i.e., minimal subsets $\mathcal{J} \subseteq \mathcal{O}$ such that $\mathcal{J} \models \alpha$, which already point out the axioms from $\mathcal{O}$ that are responsible for α. However, actually understanding why α follows may require a more detailed proof. Informally, a *proof* is a tree consisting of inference steps $\frac{\alpha_1 \cdots \alpha_n}{\alpha}$, where each step is sound, i.e., $\{\alpha_1, \ldots \alpha_n\} \models \alpha$ holds (see Figure 1). Our proofs conform to the framework [2,3]. Often, such a proof is built from the *inference rules* of an appropriate calculus [7, 36]. However, there also exist approaches to generate DL proofs that start with a justification, and extend it with intermediate axioms (*lemmas*) using heuristics [21, 22], concept interpolation [35], or forgetting [2].

It is important that proofs are neither too detailed nor too short. In fact, a justification can itself be seen as a one-step proof of an unintended entailment α, but if each element of the justifications seems reasonable to the user, then it can be hard to track down the precise interaction between these axioms that causes the problem. Recall that axioms may still not behave as the user expects, e.g. "every A has only rs" does not imply "every A has an r". On the other hand, by assumptions also made in [10, 34], too many small proof steps can also be detrimental for understanding, because they are distracting. For example, a reasoner may add the step $\frac{\mathsf{C}\sqcap\mathsf{B}\sqsubseteq\bot}{\mathsf{B}\sqcap\mathsf{C}\sqsubseteq\bot}$ to Figure 1 to make the two conjunctions match syntactically. However, for a user this only adds clutter to the proof, because the correspondence between "B and C" and "C and B" is intuitively clear.

A textual representation of a proof is necessarily a *linearization*, where the inference steps are explained in a sequence, for example in a top-down left-right order. A text corresponding to the formal proof in Figure 1 could be the following:

> Since every A has an r and every A has only rs that are Bs and Cs, every A has an r which is a B and a C. Since every A has an r which is a B and a C and there is no object which is a C and a B at the same time, there is no object in A.

Other aspects in which a text differs from a proof tree are that conjunctions (e.g. "since", "and") are used to illustrate proof steps and that statements may be repeated if they are reused later. For our second experiment, we used a flexible visualization of trees in which arrows are used instead of horizontal lines (see Figure 2 on page 9).

As described in the introduction, we do not use the formal DL syntax for the experiments, i.e., also in the proof tree representation, $\mathsf{A} \sqsubseteq \exists\mathsf{r}.\top$ would be shown as "Every A has an r." Moreover, we use nonsense names that vaguely look and sound English to enable more natural-sounding sentences, e.g. "Every woal is munted only with luxis that are kakes" instead of "Every A has only rs that are Bs and Cs" ($\mathsf{A} \sqsubseteq \forall\mathsf{r}.(\mathsf{B} \sqcap \mathsf{C})$). We already faced a problem concerning prior knowledge about the example domains in [10]. Thus, in this study we opted for non-existing domains.

4 Connecting Logical Abilities and Proof Understanding

We first conducted an experiment that shows a connection between participants' understanding of logical proofs and their general cognitive abilities. A printable version of the survey is available online.[2]

4.1 Description of the experiment

Introduction and Goals. In our main experiment (Section 5), we want to distinguish user preferences based on their baseline level of ability to solve logical reasoning tasks. To this end, we want to employ a standardized measure that allows us to predict the performance on such tasks. Here, we first tested whether the ICAR16 questionnaire can be used for this purpose.

[2] `https://cloud.perspicuous-computing.science/s/oHp9pRaoCx5SDsF`

Design. We used LimeSurvey[3] for hosting our fully online survey. Since we did not pre-screen (i.e., impose eligibility criteria on) our participants for experience with logic or proofs, we inserted a short introduction explaining the structure of proof trees. In order to exclude the effect of tiredness, the order of the ICAR16 questions and the proof tasks was randomized.

Participants. The sample consisted of 101 participants (45 female, 56 male) with a mean age of $M = 24.52$ ($SD = 6.81$). The age range spread between 18 and 48 years. Participants were recruited using Prolific[4] and were paid 6€ for their participation. Apart from being at least 18 years old, there were no exclusion criteria. 28 participants who did not complete the survey were excluded (and not paid).

Materials.

ICAR16. To assess the participants' cognitive abilities, the abbreviated form of the International Cognitive Ability Resource (ICAR16) [15] was applied.[1] It consists of 16 questions equally distributed over four different types: matrix reasoning, letter and number series, verbal reasoning, and 3-dimensional rotation. The eight answer options were displayed in a single-choice format. In the end, a mean score was calculated by coding correct answers with 1 and incorrect answers with 0. Thus, the maximum score was 1, while the minimal score was 0. The internal consistency of the ICAR16 questionnaire is $\alpha = .81$ [15].

Logical reasoning. To test the performance with logical reasoning, participants had to solve two tasks. The first described a set of axioms (in natural language) and they should decide which of the given statements follow from the axioms. Each of the statements could be marked as "follows", "does not follow" or "I do not know". In the second task, they were given a proof in tree shape (like in Figure 2) that contained a blank node, and they were asked which of some given statements would be valid labels for the node in the context of the proof ("yes", "no", "I do not know"). The mean score of the performance in both tasks was calculated from the number of correct answers. The highest possible score was 24.

Further Information. As further information, demographic data were collected (age, gender) as well as experience with propositional logic, from 1 ("no knowledge at all/no experience") to 5 ("expert/a lot of experience"). After each proof task, the participants were asked to rate its difficulty on a scale from 1 ("very easy") to 5 ("very difficult").

Hypothesis. The ICAR16 score is an independent variable that, according to our hypothesis, predicts the performance in the logical proofs.

[3] `https://www.limesurvey.org/`

[4] `https://www.prolific.co/`

4.2 Results

Descriptive Results. The mean of the ICAR16 scores was $M = .55$ ($SD = .24$) with the participants' performance being spread in a normal distribution. The maximal achieved score was 1, the minimum was 0. The mean of the score for both logical reasoning tasks was $M = 15.99$ ($SD = 3.3$), with the maximum score being 23 and the minimum 6. The performance in these tasks was also normally distributed across the participants. Moreover, the difficulty of the first task was rated with a mean of $M = 3.64$ ($SD = .89$). The difficulty of the second task was rated as $M = 3.55$ ($SD = 1.14$).

Regression analysis. A multiple regression analysis was carried out using the performance in the logical reasoning tasks as the dependent and the ICAR16 performance as the independent variable. The ICAR16 score significantly predicted the performance in the logical tasks ($F(1, 99) = 43.15$, $p < .001$). The ICAR16 explained 30% of the variation in the score of the logical tasks ($R^2 = .3$, $p < .001$), which can be interpreted as large effect size/high explained variance [14].

5 Logical Abilities and Proof Representation Preferences

Given that ICAR16 scores are highly correlated with performance on logical reasoning tasks, we used it in our main experiment to distinguish participants by their logical ability level. A printable version of the survey is available online.[5]

5.1 Description of the experiment

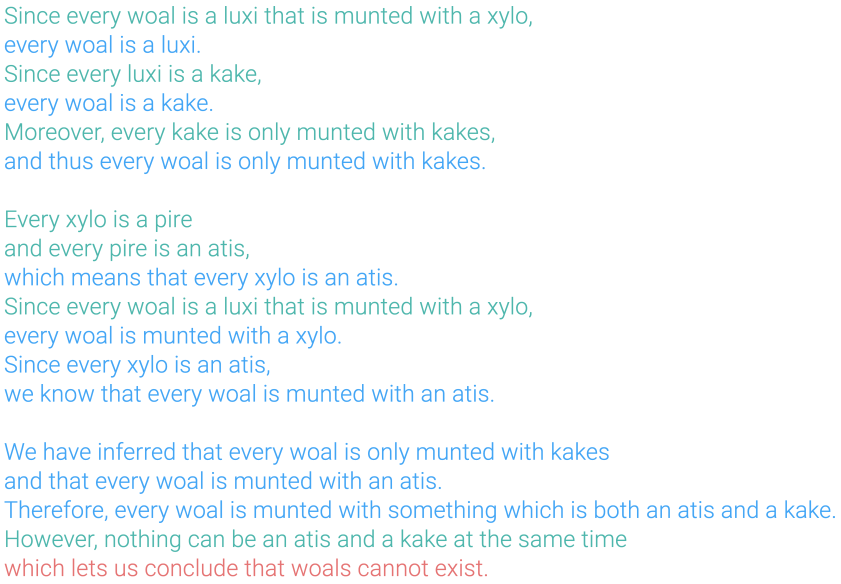
Introduction and Goals. With this experiment, we attempt to find out which proof representation is most understandable for different users. The goal is to find a difference in the (subjective) preferences and (objective) performance on each proof representation, depending on the user's level of logical reasoning ability.

Conditions. We used two different conditions (factors) with two levels each. One condition was the representational form of the proof, which was either tree-shaped or (linear) textual. The other condition was the interactivity of the proof representation, which was either static or interactive. Thus, there were the four following condition combinations: (**ir**) interactive tree, (**sr**) static tree, (**ix**) interactive text, and (**sx**) static text (see Table 1). We used a 2 × 2 within-subjects design, which means that each participant saw all four combinations. The independent variable in the main study is the ICAR16 score. Objective (the number of correct answers) and subjective (comprehensibility rating) performances on proofs as well as proof rankings are dependent variables. We conducted the experiment in English via LimeSurvey, with embedded links to the interactive proof representations that were hosted on our own server.

[5] `https://cloud.perspicuous-computing.science/s/dCSmbraoJ4RzDqG`

Table 1. Distribution of the conditions and proofs into four participant groups.

$*$ \ *Groups*	*1*	*2*	*3*	*4*
Proof 1	**ix**	**sr**	**sx**	**ir**
Proof 2	**ir**	**ix**	**sr**	**sx**
Proof 3	**sx**	**ir**	**ix**	**sr**
Proof 4	**sr**	**sx**	**ir**	**ix**



Since every woal is a luxi that is munted with a xylo,
every woal is a luxi.
Since every luxi is a kake,
every woal is a kake.
Moreover, every kake is only munted with kakes,
and thus every woal is only munted with kakes.

Every xylo is a pire
and every pire is an atis,
which means that every xylo is an atis.
Since every woal is a luxi that is munted with a xylo,
every woal is munted with a xylo.
Since every xylo is an atis,
we know that every woal is munted with an atis.

We have inferred that every woal is only munted with kakes
and that every woal is munted with an atis.
Therefore, every woal is munted with something which is both an atis and a kake.
However, nothing can be an atis and a kake at the same time
which lets us conclude that woals cannot exist.

Fig. 2. A linear textual representation of Proof 1.

Design. The survey was again implemented using LimeSurvey. As in the first experiment, the order of the ICAR16 and the proof question groups was randomized. Moreover, each participant was randomly assigned to one of the four groups in Table 1. Before the proof tasks, we added a short explanation of the proof representation and a small training example on which the participants could explore both interactive formats.

Participants. The final sample consisted of 173 participants (41% female, 59% male) with a mean age of $M = 24.8$ ($SD = 8.21$) and an age range from 18 to 65 years. The mean of the participants' experience with propositional logic was $M = 1.76$ ($SD = 1$). Furthermore, 60.7% of the participants indicated that they never worked with propositional logic. The participants were recruited using Prolific[4] and were paid 10.20€ for their participation. Apart from being at least 18 years old, there were no exclusion criteria. 83 participants who did not complete the survey were excluded also by the platform. Due to technical errors, the proofs were not displayed for 3 participants, which were also excluded. Additionally, four attention checks were implemented in the study. 13 participants with more than 2 incorrectly answered attention checks were excluded as well.

Material.

ICAR16. Again, we used the ICAR16 questionnaire to assess the participants' cognitive abilities (see page 5).

Proofs. We developed four artificial proofs of roughly the same difficulty level. The statements of each proof were given in textual form (also for the tree-shaped representations), with concept and role names replaced by nonsense words (see Section 3). For each of the four proofs, the proofs in both representation formats were created manually, not automatically. For the linear textual format, the individual statements were connected by additional words and statements were repeated if they were last mentioned more than 2 lines above. In Figures 2 and 2, we depict the textual and tree-shaped representations of Proof 1. Green color indicates the assumptions, intermediate deductions are marked in blue, and the final conclusion is colored orange.

The interactive version of the tree representation[6] started with only the final conclusion visible, and participants could interact with each node by using three buttons. A button labeled "<" revealed the immediate predecessors of the current node in the proof. The second button labeled "↑" could reveal the whole subtree above the current node, and the last button labeled "↓" allowed to collapse the subtree back to a single node. The interactive text[7] worked in a different way. At the beginning, participants saw only the first sentence, i.e., the first assumption. Using two buttons labeled ">" and "<" they could reveal the next sentence or hide the most recently revealed sentence. Additionally, clicking on a sentence highlighted the premises that were used to infer the selected statement (corresponding to the predecessors of a node in the tree representation). Moreover, both interactive representations could be freely zoomed and panned. The interactive proofs were provided by a prototypical web application for explaining DL entailments called Evonne [5, 18]. For this study, Evonne was adapted to allow for (linear) interactive textual representation of proofs. In addition, the modes of interaction in the tree representation were kept relatively basic on purpose. This was done especially to avoid overwhelming participants who had little experience with logic and proofs.

For each proof, there were three pages of questions. For the interactive conditions, the proofs on the first page were collapsed (to the root node or the first sentence), but on subsequent pages they started fully expanded since they had already been explored. Each question page contained a single question with 6 answer options (plus "none of these" and "I don't know"). Questions were of the

[6] https://lat.inf.tu-dresden.de/evonne/proof1

[7] https://lat.inf.tu-dresden.de/evonne/textProof1

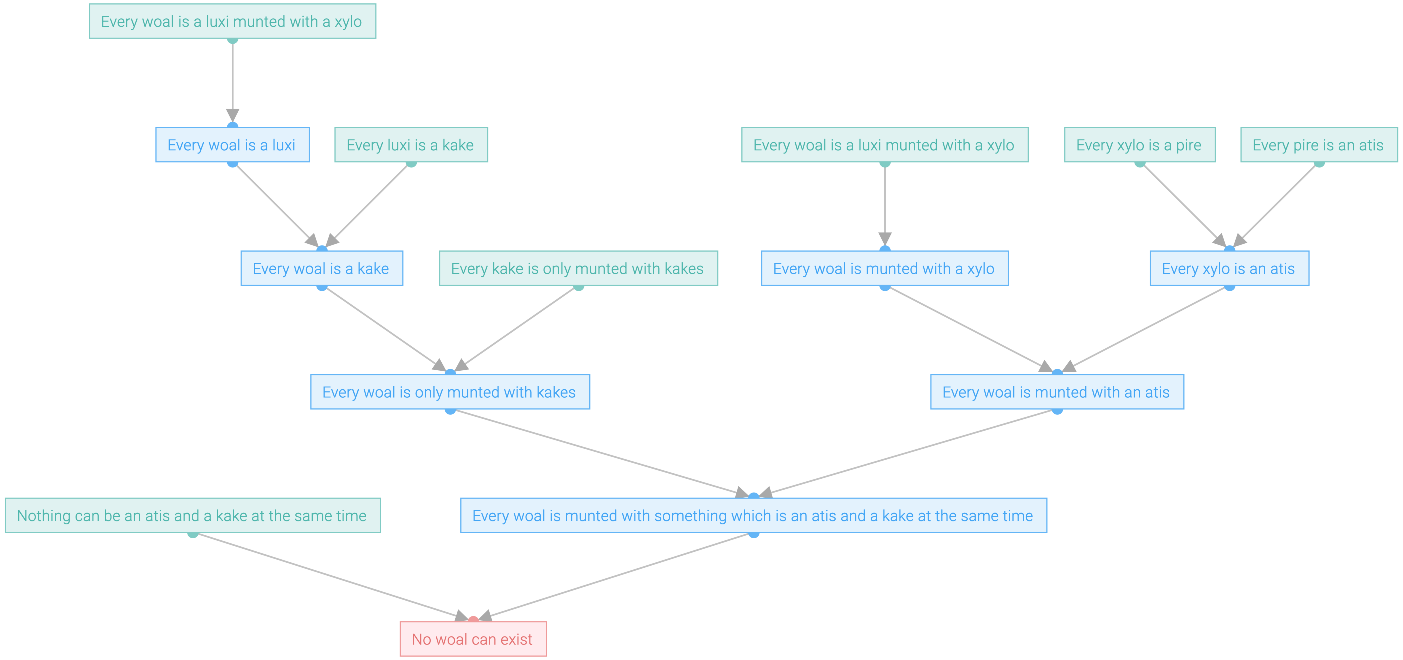

Every woal is a luxi munted with a xylo
Every woal is a luxi
Every luxi is a kake
Every woal is a luxi munted with a xylo
Every xylo is a pire
Every pire is an atis
Every woal is a kake
Every kake is only munted with kakes
Every woal is munted with a xylo
Every xylo is an atis
Every woal is only munted with kakes
Every woal is munted with an atis
Nothing can be an atis and a kake at the same time
Every woal is munted with something which is an atis and a kake at the same time
No woal can exist

Table 2. A tree-shaped representation of Proof 1.

form “Which of the following would be a correct replacement for the deduction ‘XYZ’ in the proof?” or “Which parts of the following summary/reformulation of the proof are incorrect?” The former questions require to choose a statement, which is not necessary equivalent to a given axiom but which aligns smoothly to the rest of the proof. In the end, a score was calculated based on the number of correct answers. Thus, the highest possible score for a user performance was 12.

Further Information. As further information, demographic data were collected (age, gender, nationality), as well as experience with propositional logic, from 1 (“no knowledge at all/no experience”) to 5 (“expert/a lot of experience”). Participants also indicated how frequently they use propositional logic with 1 being “Never” and 5 being “All the time”. After each proof, the participants were asked to rate its comprehensibility on a scale from 1 (“not at all”) to 5 (“very much”). At the end of the survey, the participants were asked to additionally rank all the proofs they had seen according to their relative comprehensibility.

Hypotheses. We stated two hypotheses concerning the preferences and performance differences between the proof representations.

Hypothesis 1: It is easier to understand interactive proofs than static proofs. This will be shown by an increase in performance and by a higher comprehensibility rating for the interactive conditions.

Hypothesis 2: The relative level of comprehensibility of a tree-shaped vs. textual proof depends on the cognitive abilities. This will be shown by a difference in performance and difficulty rating between the condition combinations and in the final comprehensibility ranking, in dependence of the ICAR16 scores.

5.2 Results

For the analyses, IBM SPSS Statistics (Version 26) predictive analytics software for Windows [24] was used. A p-value threshold of 0.05 was used for the entire analyses. After the assumptions were considered as tenable, a regression analysis was carried out, to confirm the results of the pre-study. Again, the predictive effect of the ICAR16 on the performance in the proofs was significant, $F(1, 171) = 24.8$, $p < .001$. With an $R^2 = .13$ (corrected $R^2 = .12$), the model shows a moderate explained variance (Cohen, 1988).

A median split ($mdn = .44$) was carried out to divide the participants into those who achieved high scores in the ICAR16 and thus presumably also have higher cognitive abilities and those who scored lower.

For ICAR16 the mean was $M = .46$, while it was $M = 2.36$ for the proof performance. The group containing those participants who scored low in the ICAR16 achieved $M = 1.9$ across all proofs. In contrast, the group of participants with high ICAR16 scores showed an overall proof performance of $M = 2.87$.

Performance and Comprehensibility Ratings. To compare the performance in the proof tasks and the subjective comprehensibility ratings after

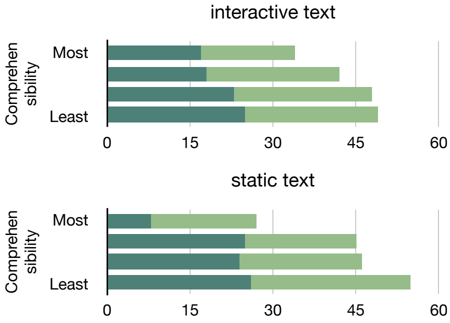


Fig. 3. Rankings of all 173 participants (light bars) and of the 83 participants with high ICAR scores (dark bars) for each condition combination.

each proof, we ran a multivariate analysis of variance (MANOVA). All the assumptions were considered as tenable. We found no significant overall difference between the conditions across the two ICAR groups, Pillai's Trace = .01, $F(6, 1376) = 1.41$, $p = .206$. Also when looking at the groups separately, we could not find any significant differences between the representations, neither in the low-ICAR group (Pillai's Trace = .03, $F(6, 712) = 1.90$, $p = .078$) nor in the group with high scores (Pillai's Trace = .01, $F(6, 656) = .53$, $p = .788$). Thus, we could not detect differences in the comprehensibility ratings as well as the performance between the various representations in each cognitive ability group and across the two groups.

Ranking. To evaluate the ranking of the four representations (1 = most comprehensible, 4 = least comprehensible), we ran a Friedman's test revealing a significant difference across both ICAR groups, $\chi^2(3) = 17.16$, $p = .001$, $n = 173$ (see Figure 3, light bars). Post-hoc pairwise comparisons were Bonferroni-corrected and showed three significant comparisons. The interactive tree was significantly more often ranked higher than the interactive text ($z = .40$, $p = .024$, Cohen's effect size $r = .03$) and also higher than static text ($z = -.50$, $p = .002$, Cohen's effect size $r = .04$). The static tree representation was also ranked significantly higher than static text, $z = .39$, $p = .032$, Cohen's effect size $r = .03$ (see Figure 3).

A Friedman's test in the group with high ICAR performance showed a significant difference in the ranking of representations, $\chi^2(3) = 12.73$, $p = .005$, $n = 83$ (see Figure 3, dark bars). Bonferroni-corrected post-hoc pairwise comparisons revealed two significant comparisons. There is a significant difference between static tree and static text ($z = .59$, $p = .019$, Cohen's effect size $r = .06$) with static tree being ranked higher than static text. Interactive tree was also preferred before static text, ($z = -.54$, $p = .041$, Cohen's effect size $r = .06$).

The low-ICAR-performers showed no significant difference in the ranking of representations, $\chi^2(3) = 6.70$, $p = .082$, $n = 90$.

6 Discussion

Neither of our two hypotheses could be conclusively confirmed. We did not find a significant advantage of using some proof representations over others, even when distinguishing groups by their levels of cognitive abilities. The analysis of the final ranking of the proof representations indicates a subjective preference of the conditions with tree-shaped proofs over their textual counterparts, but this did not seem to impact the objective performance measure nor the subjective ratings the participants gave after each proof. These preferences are largely driven by the group with higher ICAR performance (cf. Figure 3).

Limitations. According to the aims of our study, we did not pre-select participants according to their experience with logic or field of studies. 55.5% of the participants had no experience with propositional logic and 60.7% had never worked with it. For many participants, even the ones with higher ICAR scores, the proof tasks were very challenging, resulting in a mean score of M=2.36 out of a total of 12. 15 people commented about the high difficulty level in the end, and only 3 said the proofs were easy to understand. This resulted in many data points being clustered on the lower end of the scale and differences being more difficult to detect. For similar future studies, it is important to calibrate the difficulty of the tasks well; they should be neither too hard nor too easy.

Conclusion. In addition to previous observations that shorter proofs are better [10, 34], we observed a subjective preference for tree-shaped proofs, although this was not reflected by increased performance in our study. As a side result, we demonstrated that cognitive abilities tested by the ICAR16 predict the reasoning performance in formal logics. In future work, we want to further investigate the trade-off between giving no details (i.e., justifications) and giving too many details (i.e., full proofs) in various representation formats. For laypersons, it may be better to quickly communicate the gist of a proof in natural language, whereas experts may require access to the formal details.

Acknowledgements This work was supported by the DFG grant 389792660 as part of TRR 248 (`https://perspicuous-computing.science`), and QuantLA, GRK 1763 (`https://lat.inf.tu-dresden.de/quantla`).

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: d'Amato, C., Fernández, M., Tamma, V.A.M., Lécué, F., Cudré-Mauroux, P., Sequeda, J.F., Lange, C., Heflin, J. (eds.) The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10587, pp. 20–36. Springer (2017). `https://doi.org/10.1007/978-3-319-68288-4_2`

2. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding small proofs for description logic entailments: Theory and practice. In: Albert, E., Kovacs, L. (eds.) LPAR-23: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning. EPiC Series in Computing, vol. 73, pp. 32–67. EasyChair (2020). https://doi.org/10.29007/nhpp
3. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: On the complexity of finding good proofs for description logic entailments. In: Borgwardt, S., Meyer, T. (eds.) Proceedings of the 33rd International Workshop on Description Logics (DL 2020) co-located with the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020), Online Event [Rhodes, Greece], September 12th to 14th, 2020. CEUR Workshop Proceedings, vol. 2663. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2663/paper-1.pdf
4. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding good proofs for description logic entailments using recursive quality measures. In: Platzer, A., Sutcliffe, G. (eds.) Proceedings of the 28th International Conference on Automated Deduction (CADE'21). Lecture Notes in Computer Science, vol. 12699, pp. 291–308. Springer-Verlag (2021). https://doi.org/10.1007/978-3-030-79876-5_17
5. Alrabbaa, C., Baader, F., Dachselt, R., Flemisch, T., Koopmann, P.: Visualising proofs and the modular structure of ontologies to support ontology repair. In: Borgwardt, S., Meyer, T. (eds.) Proceedings of the 33rd International Workshop on Description Logics (DL 2020) co-located with the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020), Online Event [Rhodes, Greece], September 12th to 14th, 2020. CEUR Workshop Proceedings, vol. 2663. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2663/paper-2.pdf
6. Androutsopoulos, I., Lampouras, G., Galanis, D.: Generating natural language descriptions from OWL ontologies: The NaturalOWL system. Journal of Artificial Intelligence Research **48**, 671–715 (2013). https://doi.org/10.1613/jair.4017
7. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Kaelbling, L.P., Saffiotti, A. (eds.) Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI'05). pp. 364–369. Professional Book Center (2005), http://ijcai.org/Proceedings/09/Papers/053.pdf
8. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017). https://doi.org/10.1017/9781139025355
9. Baader, F., Peñaloza, R., Suntisrivaraporn, B.: Pinpointing in the description logic $\mathcal{EL}^+$. In: Proc. of the 30th German Annual Conf. on Artificial Intelligence (KI'07). Lecture Notes in Computer Science, vol. 4667, pp. 52–67. Springer, Osnabrück, Germany (2007). https://doi.org/10.1007/978-3-540-74565-5_7
10. Borgwardt, S., Hirsch, A., Kovtunova, A., Wiehr, F.: In the Eye of the Beholder: Which Proofs are Best? In: Borgwardt, S., Meyer, T. (eds.) Proc. of the 33rd Int. Workshop on Description Logics (DL 2020). CEUR Workshop Proceedings, vol. 2663 (2020), http://ceur-ws.org/Vol-2663/paper-6.pdf
11. Burke, H.R.: Raven's progressive matrices: Validity, reliability, and norms. The Journal of Psychology **82**(2), 253–257 (1972). https://doi.org/10.1080/00223980.1972.9923815
12. Chierchia, G., Fuhrmann, D., Knoll, L.J., Pi-Sunyer, B.P., Sakhardande, A.L., Blakemore, S.J.: The matrix reasoning item bank (mars-ib): novel, open-access abstract reasoning items for adolescents and adults. Royal Society Open Science **6**(10), 190232 (2019). https://doi.org/10.1098/rsos.190232

13. Climie, E., Rostad, K.: Test review: Wechsler adult intelligence scale. Journal of Psychoeducational Assessment **29**, 581–586 (12 2011). https://doi.org/10.1177/0734282911408707
14. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, 2nd edn. (1988). https://doi.org/10.4324/9780203771587
15. Condon, D.M., Revelle, W.: The international cognitive ability resource: Development and initial validation of a public-domain measure. Intelligence **43**, 52–64 (2014). https://doi.org/10.1016/j.intell.2014.01.004
16. Daniel, M.H.: Intelligence testing: Status and trends. American psychologist **52**(10), 1038 (1997). https://doi.org/10.1037/0003-066X.52.10.1038
17. Engström, F., Nizamani, A.R., Strannegård, C.: Generating comprehensible explanations in description logic. In: Informal Proceedings of the 27th International Workshop on Description Logics, Vienna, Austria, July 17-20, 2014. pp. 530–542 (2014), http://ceur-ws.org/Vol-1193/paper_17.pdf
18. Flemisch, T., Langner, R., Alrabbaa, C., Dachselt, R.: Towards designing a tool for understanding proofs in ontologies through combined node-link diagrams. In: Ivanova, V., Lambrix, P., Pesquita, C., Wiens, V. (eds.) Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual Conference (originally planned in Athens, Greece), November 02, 2020. CEUR Workshop Proceedings, vol. 2778, pp. 28–40. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2778/paper3.pdf
19. Heydasch, T.: The Hagen matrices test (HMT). Ph.D. thesis, University of Hagen, Germany (2014). https://doi.org/10.13140/RG.2.2.31433.75361
20. Horn, J., Cattell, R.: Refinement and test of the theory of fluid and crystallized general intelligences. Journal of educational psychology **57**(5), 253—270 (October 1966). https://doi.org/10.1037/h0023816
21. Horridge, M.: Justification Based Explanation in Ontologies. Ph.D. thesis, University of Manchester, UK (2011), https://www.research.manchester.ac.uk/portal/files/54511395/FULL_TEXT.PDF
22. Horridge, M., Bail, S., Parsia, B., Sattler, U.: Toward cognitive support for OWL justifications. Knowledge-Based Systems **53**, 66–79 (2013). https://doi.org/10.1016/j.knosys.2013.08.021
23. Horridge, M., Parsia, B., Sattler, U.: Justification oriented proofs in OWL. In: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. pp. 354–369 (2010). https://doi.org/10.1007/978-3-642-17746-0_23
24. IBM Corp.: IBM SPSS Statistics for Windows [computer software], https://www.ibm.com/products/spss-statistics
25. Janzen, H.L., Obrzut, J.E., Marusiak, C.W.: Test review: Roid, GH (2003). Stanford-Binet intelligence scales, (SB: V). Itasca, IL: Riverside Publishing. Canadian Journal of School Psychology **19**(1-2), 235–244 (2004). https://doi.org/10.1177/082957350401900113
26. Kazakov, Y., Klinov, P., Stupnikov, A.: Towards reusable explanation services in Protege. In: Artale, A., Glimm, B., Kontchakov, R. (eds.) Proc. of the 30th Int. Workshop on Description Logics (DL'17). CEUR Workshop Proceedings, vol. 1879 (2017), http://www.ceur-ws.org/Vol-1879/paper31.pdf
27. Kazakov, Y., Krötzsch, M., Simancik, F.: The incredible ELK – from polynomial procedures to efficient reasoning with $\mathcal{EL}$ ontologies. J. Autom. Reasoning **53**(1), 1–61 (2014). https://doi.org/10.1007/s10817-013-9296-3

28. Krawczyk, D.C.: The cognition and neuroscience of relational reasoning. Brain Research **1428**, 13–23 (2012). https://doi.org/10.1016/j.brainres.2010.11.080
29. Kuhn, T.: The understandability of OWL statements in controlled english. Semantic Web **4**(1), 101–115 (2013). https://doi.org/10.3233/SW-2012-0063
30. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Measuring the understandability of deduction rules for OWL. In: Proceedings of the First International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2012, Galway, Ireland, October 8, 2012. pp. 1–12 (2012), http://www.ida.liu.se/~patla/conferences/WoDOOM12/papers/paper4.pdf
31. Peñaloza, R., Sertkaya, B.: Understanding the complexity of axiom pinpointing in lightweight description logics. Artificial Intelligence **250**, 80–104 (2017). https://doi.org/10.1016/j.artint.2017.06.002
32. Raven, J., Raven, J.: Raven progressive matrices. In: Handbook of Nonverbal Assessment, pp. 223–237. Springer US, Boston, MA (2003). https://doi.org/10.1007/978-1-4615-0153-4_11
33. Schiller, M.R.G., Glimm, B.: Towards explicative inference for OWL. In: Informal Proceedings of the 26th International Workshop on Description Logics, Ulm, Germany, July 23 - 26, 2013. pp. 930–941 (2013), http://ceur-ws.org/Vol-1014/paper_36.pdf
34. Schiller, M.R.G., Schiller, F., Glimm, B.: Testing the adequacy of automated explanations of EL subsumptions. In: Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. (2017), http://ceur-ws.org/Vol-1879/paper43.pdf
35. Schlobach, S.: Explaining subsumption by optimal interpolation. In: Alferes, J.J., Leite, J.A. (eds.) Proc. of the 9th Eur. Conf. on Logics in Artificial Intelligence (JELIA'04). Lecture Notes in Computer Science, vol. 3229, pp. 413–425. Springer-Verlag (2004). https://doi.org/10.1007/978-3-540-30227-8_35
36. Simancik, F., Kazakov, Y., Horrocks, I.: Consequence-based reasoning beyond horn ontologies. In: IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. pp. 1093–1098 (2011). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-187
37. Sternberg, R.J., Grigorenko, E., Bundy, D.A.: The predictive value of IQ. Merrill-Palmer Quarterly **47**, 1–41 (2001). https://doi.org/10.1353/mpq.2001.0005
38. Urbina, S.: Tests of intelligence. In: Sternberg, R.J., Kaufman, S.B. (eds.) The Cambridge Handbook of Intelligence, p. 20–38. Cambridge Handbooks in Psychology, Cambridge University Press (2011). https://doi.org/10.1017/CBO9780511977244.003
39. Young, S.R., Keith, T.Z.: An examination of the convergent validity of the ICAR16 and WAIS-IV. Journal of Psychoeducational Assessment **38**(8), 1052–1059 (2020). https://doi.org/10.1177/0734282920943455