The Efficacy of OWL and DL on User Understanding of Axioms and Their Entailments

Eisa Alharbi, John Howse, Gem Stapleton, Ali Hamie, and Anestis Touloumis

University of Brighton, Brighton, UK
{e.alharbi2,John.Howse,g.e.stapleton,a.a.hamie,a.touloumis}@brighton.ac.uk

Abstract. OWL is recognized as the de facto standard notation for ontology engineering. The Manchester OWL Syntax (MOS) was developed as an alternative to symbolic description logic (DL) and it is believed to be more effective for users. This paper sets out to test that belief from two perspectives by evaluating how accurately and quickly people *understand* the informational content of axioms and derive *inferences* from them. By conducting a between-group empirical study, involving 60 novice participants, we found that DL is *just as effective* as MOS for people's understanding of axioms. Moreover, for two types of inference problems, *DL supported significantly better task performance than MOS*, yet MOS never significantly outperformed DL. These surprising results suggest that the belief that MOS is more effective than DL, at least for these types of task, is unfounded. An outcome of this research is the suggestion that ontology axioms, when presented to non-experts, may be better presented in DL rather than MOS. Further empirical studies are needed to explain these unexpected results and to see whether they hold for other types of task.

Keywords: ontologies, OWL, DL, Manchester OWL Syntax, usability

1 Introduction

This paper sets out to provide evidence to support the untested belief that the Manchester syntax [6] for OWL [2] is more effective for users than Description Logic (DL) [3]. Research efforts have focused on the usability of OWL itself, demonstrating the importance placed on effectively supporting ontology engineers and other stakeholders [4, 14, 17]. In light of this, it is equally as important to determine whether OWL, or more specifically the Manchester OWL Syntax (MOS), is an effective choice of notation. After all, MOS is widely employed and was developed with the intention of being usable by people [6]. Surprisingly, however, in this paper we found no evidence that MOS is superior to DL but instead that DL was sometimes more effective than MOS.

To probe deeply into the relative usability of notations, it is necessary to consider the tasks for which they are to be used. In the context of ontology engineering, notations are used to write axioms which must then be understood. Inferences are derived from those axioms and, ideally, ontology engineers would

at least be able to accurately identify when sound inferences hold. Of course, numerous other activities are performed, such as debugging and repair [8, 12], but in this paper, we focus on relative usability from the perspective of understanding axioms and deriving inferences from them. The specific questions we address, via an empirical study, represent the first steps towards determining the relative efficacy of MOS as compared to DL and are as follows:

1. Does MOS support significantly more accurate understanding of axioms than DL? We found MOS to be no more effective than DL.
2. Does MOS support significantly more accurate identification of sound inferences than DL? We found that MOS does not and, sometimes, DL is more effective than MOS.
3. Does MOS lead to significantly fewer unsound inferences than DL? We found that MOS does not and, sometimes, DL is more effective than MOS.

Given the surprising answers to these questions, particularly with respect to MOS not significantly outperforming DL, additional research is needed. In particular, future empirical studies should evaluate MOS and DL to determine the extent to which DL can outperform MOS and to identify tasks for which MOS outperforms DL. A key take-away message is that it is not clear-cut that MOS is a more usable notation.

The paper is set out as follows. Section 2 provides an overview of related work, focusing on the usability of MOS. In section 3, we illustrate the nature of task that users were required to perform in our empirical study. The hypotheses to be tested are given in section 4 and the experiment design is described in section 5, together with the statistical methods employed. The results obtained are given in section 6 and discussed in section 7. We identify threats to validity in section 8 and conclude in section 9. The experiment materials and data collected is at https://sites.google.com/site/eisamalharbi/owlanddlefficiency.

2 Background

Ontology engineering has become a major activity with many stakeholders involved in producing ontologies. The W3C OWL working group devised several different syntaxes – e.g. RDF/XML and a functional style syntax – designed to serve different purposes. However "none of them ... are designed for ease of use by humans when building or analyzing ontologies" [6]. Given the diversity of expertise held by different stakeholders, it is important to ensure the efficacy of notations used for ontology engineering.

The Manchester syntax was created with a view that it "would be easier to write and understand, particularly for non-logicians" [6]. This is supported by the official W3C working group documentation: "The Manchester syntax is a user-friendly compact syntax for OWL 2 ontologies"[1]. Indeed, the Manchester syntax is the de facto standard notation used for ontology engineering and various tools support its use, such as Protégé [11]. It is believed that because

the Manchester syntax uses short and intuitive English words instead of logical symbols, such as those employed by DL, usability is improved [6].

Despite these beliefs, it is known that some users find interpreting OWL difficult. Warren et al. provide insight into the relative efficacy of different Manchester OWL constructs, with a focus on drawing sound inferences from given axioms [17]. Whilst this study revealed that users were "prone to certain misconceptions" it did not compare the Manchester syntax with DL or other notations. An evaluation by Sarker et al. [15] reported that ROWLtab, a Protégé plugin that allows users to enter OWL axioms by way of rules, " is much quicker than the standard interface, while at the same time, also less prone to errors for hard modeling tasks." Others have also considered the understandability of OWL, such as [14], and Horridge et al. [4] provided insight into the relative cognitive complexity of OWL justifications, but again not in comparison to DL.

In summary, insight has been gained about the relative understandability of different Manchester OWL axioms, particularly with in the context of inference problems. However, the perceived superiority of the Manchester syntax has not been rigorously tested by empirical studies that aim to understand its relative cognitive advantages over DL. In this paper we present the first such empirical study, revealing unexpected results.

3 Tasks: Understanding Axioms and Inference

When presented with an ontology, users need to understand the informational content of axioms as well as derive insights from them. Consider the following:

1. Demon SubClassOf Elf
2. Korrigan SubClassOf Demon
3. Mermaid SubClassOf Spirit
4. Elf DisjointWith Nisse
5. Demon SubClassOf hates only Goblin
6. Elf SubClassOf chases some Spirit
7. Halfling SubClassOf watches some Fairy
8. guides Domain Mermaid

Each of these axioms needs to be *understood*. For example, axiom 2 indicates class subsumption and is taken to mean 'All Korrigans are Demons'. Axiom 4 asserts class disjointness: 'No Elf is a Nisse'. More complex axioms involve quantifiers, such as axiom 7 which tells us that 'Halflings watch at least one Fairy'.

The derivation of inferences requires people to *reason* about their informational content. Reasoning is clearly a harder task than understanding axioms, since the axioms must be understood in order to make sound inferences from them. Considering the axioms above, many inferences can be drawn, such as:

- 'No Demon is a Nisse': from axiom 1 we know 'All Demons are Elves' and from axiom 4 we see that 'No Elf is a Nisse'; so 'No Demon is a Nisse'.
- 'Korrigans hate only Goblins': follows from axioms 2 and 5.
- 'Demons chase at least one Spirit': follows from axioms 1 and 6.

In each case, two axioms have been used to derive the conclusions; more complex reasoning can also occur, but we focus on inferences drawn from two axioms.

It is also important that people do not make incorrect inferences. Examples of statements that are not semantically entailed by the axioms above include: 'No Halfling is a Spirit', 'Fairies track only Elves', and 'Things scare only Halflings.' It would be unsound to deduce any of these three statements. In summary, it is important that ontology engineers and end-users understand axioms correctly, draw sound inferences from them, and do not make unsound inferences. The study we design covers all three aspects.

4 Main Hypotheses

There is a belief that the Manchester syntax is usable, in that it is easy to read (i.e. understandable) and write, as exemplified by section 2. A possible reason for this is the use of text rather than symbols. For instance, contrast Goblin SubClassOf Imp with Goblin $\sqsubseteq$ Imp: both express 'all goblins are imps'. The MOS is likely to be easier for people to understand than the DL even though it requires people to understand what is meant by SubClassOf.

DL, by contrast to MOS, exploits purely syntactic conventions whose semantics are defined in a stipulative way; the symbols do not immediately correspond to a natural language interpretation of the axioms. Therefore, DL's syntactic objects are further removed from their semantics than those of MOS. This means that DL could provide an additional cognitive burden on users, as there is a need to learn how to read the symbols in addition to then deriving an understanding of the axioms. This suggests that users of DL need to be more conscious of semantic conventions than users of MOS. Consequently, we expect an increased cognitive load for DL users which could be a deterrent to their performance when understanding and reasoning about axioms.

Given the above discussion, we identify the following hypotheses:

- People more accurately *understand* axioms using MOS than DL.
- People identify *sound inferences* more accurately using MOS than DL.
- People make fewer *unsound inferences* using MOS than DL.
- People perform tasks more quickly *overall* using MOS than DL.

With regard to the first three hypotheses, we will present a fine-grained statistical analysis that inspects performance with respect to understanding different types of axioms and different styles of inference task. Regarding the last hypothesis, time data was collected for each question, which involved all three types of task (i.e. *understanding*, *sound inference* and *unsound inference*). This design decision was to reduce the impact of fatigue effect on the data, since fewer questions and, thus, fewer sets of axioms needed to be presented to participants; if we collected time data at the fine-grained level, participants would need to answer nine times as many questions - given our design - which is not feasible. Consequently, there was no time data specifically for measuring the understanding of different types of axioms or for different styles of inference task.

5 Empirical Study Design

In order to determine whether MOS or DL most effectively helped people understand and reason about ontologies, we focused on six axiom types:

1. simple class subsumption: C_1 SubClassOf C_2 and $C_1 \sqsubseteq C_2$,
2. simple class disjointness: C_1 DisjointWith C_2 and $C_1 \sqcap C_2 \sqsubseteq \bot$,
3. complex class subsumption, involving all values from constraints: C_1 SubClassOf R only C_2 and $C_1 \sqsubseteq \forall R.C_2$,
4. complex class subsumption, involving some values from constraints: C_1 SubClassOf R some C_2 and $C_1 \sqsubseteq \exists R.C_2$,
5. domain: R Domain C_1 and $\exists R.\top \sqsubseteq C_1$,
6. range: R Range C_1 and $\top \sqsubseteq \forall R.C_1$,

where the C_i are primitive concepts. These were chosen because they are commonly occurring, especially simple class subsumption and complex class subsumption involving some values from which are prominent in biomedical ontologies. It was deemed important that participants had no prior knowledge of the information contained in the axioms, so that they could not work out the answers without reading the MOS or DL. Equally, the use of abstract-style axioms, such as in the enumerated list above, could be off-putting to participants. Therefore, the axioms presented information about mythical creatures to give some context to the questions. A between-group design was used, with participants being exposed to one of the two notations[1]. We measured relative efficacy through accuracy and time performance data. Accuracy was taken to be the primary performance indicator: one notation was more effective than another if people performed tasks significantly more accurately with it. To establish whether significant performance differences existed, we designed an empirical study that required participants to answer questions that required a set of checkboxes to be selected, corresponding to understanding the axioms and deriving sound inferences from the axioms. Further checkboxes were included that related to information that could not be deduced from the axioms. The nature of these checkboxes will be further explained below.

5.1 Designing Questions for the Study

A screenshot of a question used in the study is given in figure 1. There is a list of 14 axioms, presented in DL in this case. Participants are asked one question: 'Which of the following statements hold?' This is followed by a list of nine checkbox statements, given in natural language. Therefore, each question can be viewed as comprising nine tasks: for each checkbox, determine whether the information conveyed by the associated statement is necessarily true. Figure 1 will be used as a running example where we describe the question design process. The next three subsections consider factors that informed the question design.

[1] The study included a third group which saw a diagrammatic representation of the axioms. We do not report on that group in this paper.

Banshee ⊑ Spirit
Gnome ⊑ Dwarf
Imp ⊑ Troll
Pixie ⊑ Spirit
Gnome ⊓ Fiend ⊑ ⊥
Imp ⊓ Dwarf ⊑ ⊥
Dwarf ⊑ ∀frightens.Pixie
Elf ⊑ ∀watches.Halfling
Fiend ⊑ ∃scares.Elf
Spirit ⊑ ∃annoys.Fiend
∃guides.⊤ ⊑ Elf
∃helps.⊤ ⊑ Elf
⊤ ⊑ ∀chases.Halfling
⊤ ⊑ ∀likes.Banshee

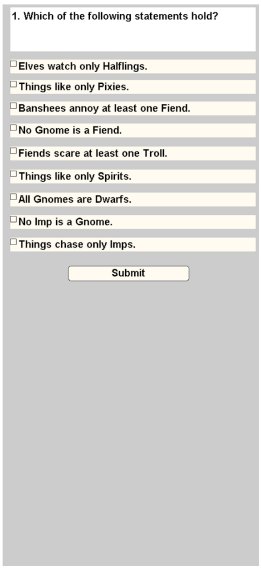

Fig. 1. Screenshot of a DL question.

Understanding Axioms As discussed earlier, an axiom can be understood through a natural language statement. For example, Gnome ⊑ Dwarf, given in figure 1 is understood as 'All Gnomes are Dwarfs'. The axiom types and their representations in MOS and DL are given in table 1 together with their associated natural language interpretation written in an abstract form; we call these interpretations *statement styles* which will be used in the context of inference as well as understanding. To obtain a sufficient number of data points to statistically analyse accuracy performance, each axiom type was tested for understandability three times, in three different questions. Since there are six axiom types, we had a total of 18 tasks (resp. 18 checkboxes) relating to understanding axioms. These 18 tasks were distributed evenly across the questions: each question included three.

Making Sound Inferences from Axioms From Banshee ⊑ Spirit and Spirit ⊑ ∃annoys.Fiend in figure 1 we can deduce Banshee ⊑ ∃annoys.Fiend, which is inter-

Table 1. Representing axioms

Axiom Type	MOS	DL	Statement Style
Simple Class Subsumption	C_1 SubClassOf C_2	$C_1 \sqsubseteq C_2$	All C_1 are C_2
Simple Class Disjointness	C_1 DisjointWith C_2	$C_1 \sqcap C_2 \sqsubseteq \bot$	No C_1 is a C_2
Complex Class Subsumption: All VF	C_1 SubClassOf p only C_2	$C_1 \sqsubseteq \forall p.C_2$	C_1 p only C_2
Complex Class Subsumption: Some VF	C_1 SubClassOf p some C_2	$C_1 \sqsubseteq \exists p.C_2$	C_1 p at least one C_2
Domain	p Domain C	$\exists p.\top \sqsubseteq C$	Only C p Things
Range	p Range C	$\top \sqsubseteq \forall p.C$	Things p only C

preted as 'Banshees annoy at least one Fiend'. We tested inferences that involved only two axioms and, in each case, one of the axioms was simple class subsumption. To give a controlled variety of inference tasks, simple class subsumption axioms were paired with each of the six axiom types we are considering. Such a pairing resulted in an inference whose interpretation is one of the six statement styles given in table 1. Each such pairing was used to give three inference tasks for each statement style and 18 sound inference tasks overall. These 18 tasks were distributed evenly across the questions: each question included three sound inference tasks.

Making Unsound Inferences from Axioms To test the ability of each notation to reduce the likelihood of unsound reasoning, tasks were included that corresponded to statements that were not semantically entailed by the axioms. For example, in figure 1, the statement Things chase only Imps cannot be inferred from the axioms. To ensure the unsound inference tasks were non-trivial, we used statements that contained classes and properties that were present in the axiom list. For consistency with the other two test types and to facilitate the statistical analysis we produced three statements that were unsound inferences from the axioms for each of the six statement styles. This gave a total of 18 statements that are unsound inferences from the axioms, again distributed evenly across the questions: each question included three unsound inference tasks.

Generating Axioms for Questions Overall, we required participants to perform 54 tasks: 18 each for understanding axioms, sound inferences, unsound inferences. Each question had nine checkboxes, so we required six questions. Each question needed a set of axioms from which three statements could be understood, three sound inferences made, and three unsound inferences identified.

It was important to have tasks of sufficient complexity to reveal statistically significant difference – should they exist – and, therefore, a reasonable number of axioms was required for each question. However, if too many axioms were involved, participants may have found the tasks too difficult to perform with minimal training. Informal experimentation indicated that providing two axioms of each type was appropriate. As discussed, the sound inference tasks involved just two axioms, one of which was always simple class subsumption. Hence to include, for each question, three sound inference tasks and, for half the questions, a simple class subsumption understanding task - each question contained four simple class subsumption axioms. So each of the six questions involved a list of 14 axioms: four simple class subsumption axioms and two of each other type.

Each set of 14 axioms was randomly generated in order to avoid selection bias, using ten named classes and eight named properties; in total 27 different class names and 15 different property names were used across the six questions. Each class name started with a different letter to avoid potential misreading, the same was true for property names. The axioms were ordered according to axiom type: simple class subsumption, simple class disjointness, followed by complex class subsumption involving all values from, some values from constraints,

domain, and, lastly, range. Within these axiom types the axioms were ordered alphabetically; see figure 1. The checkbox statements were generated randomly by statement style and task type, whilst ensuring the required distribution of checkboxes. The statements for each question were presented in fixed random order, see figure 1. The presentation of each question (i.e. order of axioms, order of checkbox statements, position of items on the screen), was identical for each participant, except for the use of MOS and DL.

Summary The six questions were designed to test participants' ability to understand axioms, to make sound inferences, and to recognize unsound inferences. Each question involved a list of 14 axioms and nine checkbox statements. The 14 axioms consisted of four simple class subsumption axioms and two each of the other axiom types. The nine checkbox statements involved three statements testing axiom understanding, three testing sound inference and three testing unsound inference. In total participants were required to consider 54 checkbox statements. To be answered correctly, the axiom understanding statements and sound inference statements boxes should be checked, but the unsound inference statements boxes should be unchecked.

5.2 Experiment Phases

The experiment had three phases: paper-based training, software-based training, and the main study. The paper-based training taught participants how to understand axioms. This consisted of one A4 sheet containing one training axiom for each of the six axiom types. They were written using the mythical creatures scenario and presented alongside English language explanations, like those in section 3. For example, the MOS group were shown the statement Boggart SubClassOf scares only Midget (among others) and were told this meant 'Boggarts scare only Midgets'; the DL group saw Boggart $\sqsubseteq \forall$scares.Midgets alongside the same meaning. Participants retained their training sheet throughout the study.

In the second phase of the study, participants were taught how to answer the questions using the software that collected performance data, familiarizing them with the user interface. This involved participants answering questions similar to those designed for the main study. The training material was identical for each group, except that the notation used was different. They were told that some of the possible answers required inferences to be made from the axioms presented. The participants attempted the training questions, then the experiment facilitator told them the correct answers and explained why they were correct.

The third phase collected performance data based on the six questions described in section 5.1. Participants were told that the information presented in each question was independent of the information in the other questions, so inferences should only be made from the axioms on the screen. They could not re-attempt questions but were able to refer to the single side of A4 paper training material from phase 1. To reduce the impact of learning effect, the questions were presented in a random order generated separately for each participant. After the

answers to a question were submitted, the software showed a pause screen, allowing participants to decide when to start the next question. This feature was designed to reduce fatigue effect and to ensure that the time recorded to answer each question was appropriate; the recorded time was the duration from when the question was displayed until an answer was submitted, not the time taken to select individual checkboxes as this would not be meaningful. No time limit was imposed on the participants, allowing them to spend as long as they needed to answer each question.

5.3 Experiment Execution

Participants were recruited by word-of-mouth and were all students, studying a variety of subjects, at the University of Brighton, none associated with the authors' research group. Some participants were not native English speakers but all had proficiency in English. Participants were randomly divided into groups, one for MOS the other for DL. A pilot study was conducted to test the experiment design, the software used to display the questions, and the data collection process. Ten participants (6F, 4M, ages 18-38) took part in the pilot, five per group. No changes were required after the pilot study. A further 50 participants (18F, 32M, ages 18-45) took part in the main study, 25 in each group.

The experiment was performed in a usability laboratory, providing a quiet environment without interruption. Participants were treated equally with the same environment, equipment, materials and procedures. They performed the experiment individually, and were provided with full details about the purpose of their role by an experiment facilitator. Upon completion, each participant was provided with a debrief summary, telling them how to access the study's results. Participants were offered a £6 canteen voucher for their time spent in the study (approximately 30 minutes).

5.4 Statistical Methods

Statistical analysis was based on the Generalized Estimation Equations (GEE) method [10] implemented in the R package geepack [18]. In addition, the function `ComparisonStats` was used to evaluate the statistical significance of the desired comparisons for the accuracy data. The notation type (participant group), the axiom type, and checkbox type were used as explanatory variables that are linearly connected with the probability of providing a correct answer. The significance of the explanatory variables and their interaction will be assessed to determine whether they affect the probability of correctly performing a task.

The following model was fitted to the accuracy data:

$$\begin{aligned}\log\left[\frac{\Pr(Y_{ij}=1)}{1-\Pr(Y_{ij}=1)}\right] =& \beta_0+\beta_1x_{ij1}+\beta_2x_{ij2}+\beta_3x_{ij3}+\beta_4x_{ij4}+\beta_5x_{ij5}+\beta_6x_{ij6}\\ &+\beta_7x_{ij7}+\beta_8x_{i8}+\beta_9x_{ij1}x_{ij6}+\beta_{10}x_{ij2}x_{ij6}\\ &+\beta_{11}x_{ij3}x_{ij6}+\beta_{12}x_{ij4}x_{ij6}+\beta_{13}x_{ij5}x_{ij6}+\beta_{14}x_{ij1}x_{ij7}\\ &+\beta_{15}x_{ij2}x_{ij7}+\beta_{16}x_{ij3}x_{ij7}+\beta_{17}x_{ij4}x_{ij7}+\beta_{18}x_{ij5}x_{ij7}\\ &+\beta_{19}x_{ij1}x_{i8}+\beta_{20}x_{ij2}x_{i8}+\beta_{21}x_{ij3}x_{i8}+\beta_{22}x_{ij4}x_{i8}\\ &+\beta_{23}x_{ij5}x_{i8}+\beta_{24}x_{ij6}x_{i8}+\beta_{25}x_{ij7}x_{i8}+\beta_{26}x_{ij1}x_{ij6}x_{i8}\\ &+\beta_{27}x_{ij2}x_{ij6}x_{i8}+\beta_{28}x_{ij3}x_{ij6}x_{i8}+\beta_{29}x_{ij4}x_{ij6}x_{i8}\\ &+\beta_{30}x_{ij5}x_{ij6}x_{i8}+\beta_{31}x_{ij1}x_{ij7}x_{i8}+\beta_{32}x_{ij2}x_{ij7}x_{i8}\\ &+\beta_{33}x_{ij3}x_{ij7}x_{i8}+\beta_{34}x_{ij4}x_{ij7}x_{i8}+\beta_{35}x_{ij5}x_{ij7}x_{i8}\end{aligned}$$

where $\Pr(Y_{ij}=1)$ is the probability for participant i to answer checkbox j correctly (i.e. ticked for understanding and sound inference tasks, not ticked for unsound inference tasks) and

- x_{ij1} is the indicator for the *simple class disjointness* statement style,
- x_{ij2} is the indicator for the *domain* statement style,
- x_{ij3} is the indicator for the *range* statement style,
- x_{ij4} is the indicator for the *simple class subsumption* statement style,
- x_{ij5} is the indicator for the complex class subsumption statement style involving *some values from*,
- x_{ij6} is the indicator for the unsound inference task type,
- x_{ij7} is the indicator for the sound inference task type,
- x_{i8} is the indicator for the *MOS* group,

for $i=1,\ldots,60$, corresponding to the individual participants, and $j=1,\ldots,54$, corresponding to the individual checkboxes. The βs are coefficients of the model computed using the data. `ComparisonStats` uses the βs to produce the p-value and the confidence interval for the contrast under study. Using this GEE-based model, we could determine whether the odds of providing a correct answer for any one combination statement style and task type is significantly different between groups (i.e. notation); this model also takes into account the expected correlation among the responses provided by each individual participant.

The regression model $\log(Z_{ik})=\gamma_0+\gamma_1x_{i8}$ was fitted to the time data where Z_{ik} is the time needed for participant i to answer question k, x_{i8} is the indicator for the *MOS* group, for $i=1,\ldots,60$ and $k=1,\ldots 6$. This GEE-based model allowed us to determine whether the time taken to answer questions for one notation was significantly different from the other.

6 Results

The following results are based on the data collected from 60 participants (30 per group); as no changes were made after the pilot study, we carried forward

the data when performing the statistical analysis. Each participant answered six questions providing a total of 3240 accuracy observations: 1620 for each group, 1080 for task type and 540 for each statement style. For each statistical comparison, arising from the 18 combinations of task type and statement style, there were 180 accuracy observations (90 each group). For the time data there were $60 \times 6 = 360$ observations, 180 for each group. Throughout, results were taken to be statistically significant at the 5% level.

6.1 Understanding Tasks

We present a full explanation for understanding tasks where the statement style was All C_1 are C_2; the remaining cases are similar and are in table 2. Both treatments yielded a mean accuracy rate of 90.00%. Using the GEE-based model, the odds of providing a correct answer in the OWL group are 1.00 times that in the DL group, with a 95% confidence interval of $(0.40, 2.52)$ and p-value of 1.00. Therefore, there is no significant difference between MOS and DL when understanding simple class subsumption axioms. No significant differences were found between MOS and DL for any of the understanding tasks.

Table 2. Results for understanding tasks.

Statement Style	MOS	DL	Odds	CI	p-value	Significant
All C_1 are C_2	90.00%	90.00%	1.00	$(0.40, 2.52)$	1.00	×
No C_1 are C_2	82.22%	83.33%	0.93	$(0.34, 2.53)$	0.88	×
C_1 p only C_2	87.78%	92.22%	0.61	$(0.21, 1.76)$	0.36	×
C_1 p at least one C_2	85.56%	93.33%	0.42	$(0.31, 1.33)$	0.14	×
Only C p things	84.44%	80.00%	1.36	$(0.58, 3.18)$	0.48	×
Things p only C	81.11%	83.33%	0.32	$(0.32, 2.33)$	0.76	×

6.2 Sound Inference Tasks

We present a full explanation for sound inference tasks where the statement style was Things p only C since this case yielded a significant result; the remaining cases are given in table 3. MOS and DL yielded mean accuracy rates of 58.89% and 83.33%. Using the GEE-based model, the odds of providing a correct answer in the MOS group are 0.29 times that in the DL group, with a 95% confidence interval of $(0.13, 0.61)$ and a p-value < 0.005. Therefore, there is a significant difference between MOS and DL when performing sound inference tasks for this statement style: DL better supports sound reasoning using a simple class subsumption axiom with a complex class subsumption axiom involving range. In terms of effect size for this task type, on average we would expect 24 more correct answers per 100 tasks when people use DL instead of MOS.

Table 3. Results for sound inference tasks.

Statement Style	MOS	DL	Odds	CI	p-value	Significant
All C_1 are C_2	78.89%	63.33%	2.16	$(0.84, 5.60)$	0.11	×
No C_1 are C_2	70.00%	56.67%	1.78	$(0.90, 3.53)$	0.10	×
C_1 p only C_2	67.78%	82.22%	0.45	$(0.20, 1.04)$	0.06	×
C_1 p at least one C_2	71.11%	80.00%	0.62	$(0.31, 1.20)$	0.16	×
Only C p things	64.44%	67.78%	0.86	$(0.40, 1.87)$	0.71	×
Things p only C	58.89%	83.33%	0.29	$(0.13, 0.61)$	0.00	✓

6.3 Unsound Inference Tasks

We present the case for unsound inference tasks where the statement style was All C_1 are C_2; the remaining cases are given in table 4. MOS and DL yielded mean accuracy rates of 93.33% and 100.00%. Using the GEE-based model to compare MOS and DL for this task, we obtained a p-value < 0.005. Therefore, there is a significant difference between MOS and DL when identifying unsound inference tasks for this statement style: DL better prevents unsound reasoning. In terms of effect size for this task type, on average we would expect 7 more correct answers per 100 tasks when people use DL instead of MOS.

Table 4. Results for unsound inference tasks.

Statement Style	MOS	DL	Odds	CI	p-value	Significant
All C_1 are C_2	93.33%	100.00%	0.00	$(0.00, 0.00)$	0.00	✓
No C_1 are C_2	77.78%	77.78%	1.00	$(0.41, 2.47)$	1.00	×
C_1 p only C_2	90.00%	90.00%	1.00	$(0.30, 3.32)$	1.00	×
C_1 p at least one C_2	91.11%	91.11%	1.00	$(0.30, 3.30)$	1.00	×
Only C p things	90.00%	92.22%	0.76	$(0.26, 2, 23)$	0.62	×
Things p only C	88.89%	81.11%	0.91	$(0.91, 3.81)$	0.09	×

6.4 Time Performance

The fastest mean time was for DL, where participants answered questions in 2 minutes 22.46 seconds, on average, which increased to 2 minutes 37.88 seconds for MOS. Using the regression model for the time data, no significant differences were found, with $p = 0.075$. Therefore, we have not found evidence that using OWL supports significantly improved task performance, with respect to time.

7 Discussion

The participants were not familiar with MOS or DL, so by that measure they were novices. They were trained to understand the axioms types in the appropriate notation (MOS or DL) by considering a natural language form. They

were also trained how to perform the inference tasks used in the study. We hypothesized that participants using MOS would perform significantly better than those using DL. The results of this empirical study are surprising: there were few significant differences between MOS and DL and, where there were significant differences, it was DL that performed better. This result does, however, chime with Keet [9] who reported that non-English language modellers preferred Protégé v3 with a symbolic DL interface over Protégé v3 using MOS.

7.1 Understanding Axioms

The success rates for understanding the axioms were high for both notations, indicating that participants had a strong understanding of their meaning. Participants using MOS achieved between 81.11% (range) and 90.00% (simple class subsumption) accuracy, with the DL group achieving between 80.00% (domain) and 93.33% (complex class subsumption involving some values from). We hypothesized that MOS would, however, outperform DL due to its textual nature: MOS appears more closely aligned with its natural language interpretation, potentially placing a lower cognitive burden on users. The lack of significant differences show, at least for tasks of this type, no difference in cognitive burden. The axioms considered in this study were chosen due to their simple form and their frequent use in ontologies but future work should consider more complex axioms to determine whether MOS brings performance benefits.

7.2 Sound Inferences

We expected the sound inference tasks to be cognitively more demanding than understanding tasks for both notations. This is confirmed by the accuracy rates which are higher for understanding axioms than for sound inference. Participants using MOS achieved between 58.89% ('range' statement styles) and 78.89% ('simple class subsumption' statement styles), with the DL group achieving between 56.67% ('disjointness' statement styles) and 82.22% ('complex class subsumption involving all values' from statement styles). These lower accuracy rates are consistent with Warren et al. [17] who found "users are prone to certain misconceptions. These include confusion ... about the inheritance of property characteristics," although the sound inference tasks in the study involved class inheritance only. Despite increased difficulty, we still expected the OWL group to perform significantly better, in part due to the expected improved understanding that did not materialize. Since the accuracy rates reduced, as compared to the understanding tasks, we can be sure that the sound inference tasks required reasonable cognitive effort to perform. As cognitive effort was demonstrably required, we cannot readily attribute lack of significant differences - found in five of the six cases - to triviality of the inference tasks. Thus, we suggest that our hypothesis is not supported: MOS does not support more accurate inferences to be made. DL can, in fact, sometimes outperform MOS. Further work needs to consider more complex inference tasks to reinforce, or otherwise, these results.

The evidence for cognitive burden arising from the task difficulty further supports the significant difference found in the 'range' statement style case, i.e. 'Things p only C'. The 'range' statement style is expressed as p Range C in MOS and $\top \sqsubseteq \forall p.C$ in DL, an example is in figure 1. The checkbox statement 'Things like only Spirits' can be inferred from the axioms $\top \sqsubseteq \forall$like.Banshee (like Range Banshee) and Banshee $\sqsubseteq$ Spirit (Banshee SubClassOf Spirit). Of the participants using DL, 25 out of 30 correctly made this inference against, surprisingly, only 13 out 30 for MOS users. It is not immediately clear why there is a significant difference only in this case. One possible explanation is that participants may be misunderstanding Range to mean the image of the relation (as is the case in some languages such as Z [16]) implying that two different classes cannot be the range. So participants interpreting Range in this way would only partially understand range axioms, leading to lack of ability to make sound inferences. If this conjecture is correct, it indicates a problem with using natural language in notations: some people may interpret natural language in a reasonable but incorrect way; a case of a little knowledge being a dangerous thing.

7.3 Unsound Inferences

The success rates for the unsound inference tasks were high for both notations, indicating that they were effective. Again, we expected MOS to outperform DL, but this was not the case. Interestingly, the DL group performed significantly better than the MOS group for unsound inferences involving 'simple class subsumption' style statements that were unsound inferences. Further work is needed to understand why these results were obtained.

8 Threats to Validity

Threats to validity are categorized as internal, construct and external [13]. With respect to internal validity, a major consideration related to carry-over effect which can arises when the measure of one treatment is affected by the measurement of another treatment. Using a between-group design ensured that each participant was only exposed to one notation and this threat was eliminated.

Construct validity focuses on dependent variables (accuracy rate, false negatives, and time) and independent variables (questions and treatments). Errors could arise if the axioms were ordered in such a way that cognition was hindered (this could also increase time taken). To manage this effect, all axioms were carefully ordered, ensuring that simple class subsumption axioms appeared first and so forth, minimizing unwanted variation between questions. The classes and properties in each question did not share a common first letter in an attempt to reduce false negatives due to misreading. Careful consideration was paid to the time taken to submit an answer: the inclusion of a pause screen between each question ensured that the question was only displayed when the participant was ready and they used the same PC with no applications running in the background. These steps were taken to ensure that the time to answer the questions was measured accurately, so far as is reasonably possible.

Lastly, we focus on external validity, by examining the limitations of the results and the extent to which they can be generalized. We observe the following. The questions involved three types of task: understanding axioms, drawing sound inferences from them, and identifying unsound inferences. Thus, our results are for these types of task only and exclude, for example, writing axioms or identifying incoherence and subsequently repairing the ontology (see [7]). Moreover, the sound inference tasks only required two axioms to be used to make the desired inference. More complex reasoning tasks were not considered.

Our tasks were limited in that each question involved 14 axioms of six commonly occurring types. Other styles of axioms may yield different results. In terms of inference, we realise that, in practice, ontologies can contain thousands of axioms. This makes the task of identifying axioms from which inferences can be made more difficult. Horridge et al. [5] identify minimal sets of axioms from which entailments holds, making inference tasks closer in cognitive complexity to the tasks in our study. Despite being able to focus on only the axioms involved in an entailment, it is important to extend our findings to inference tasks involving more than two axioms; the authors of [4] stated that "fewer than 10" axioms can still give rise to "difficult justifications" from the perspective of cognition.

The participants were all novices and were (minimally) trained in the notations. With ontologies being developed in a range of areas, where stakeholders need not have expertise in MOS or DL, our results are particularly relevant. We might obtain different results for expert participants who are familiar with one of DL and MOS. Ultimately, our results should be taken to be valid within the constraints imposed by the study design and execution.

9 Conclusion

The belief that the Manchester syntax for OWL is more usable than competing notations is widespread. Our findings suggest that for a range of task types - understanding axioms, deriving sound inferences from them, and preventing unsound reasoning - the Manchester syntax for OWL is *not* more effective than DL. This result itself is surprising, but our study also suggests that DL can sometimes better support users than the Manchester syntax. These results begin to challenge the belief that the Manchester syntax is easier for people to use.

Further work is needed to determine the extent to which DL better supports task performance than the Manchester syntax and our research raises more questions than it answers. For instance, for more complex versions of the three task types considered in our study, does the Manchester syntax support more accurate understanding than DL or other notations? Other types of task were not considered, such as writing axioms and ontology debugging and repair: does the Manchester syntax support more accurate task performance than DL, or other notations, for these other tasks? Would we see similar results if our study was re-run with expert users? Answering these questions could yield exciting new insights into the relative cognitive complexity of competing notation choices and the different types of task that ontology engineers must perform. Indeed, not

only are such answers important for ontology engineers and end-users, but also more widely in that they could impact the design of future notations. Beyond this, our major takeaway message is that it is not clear-cut that the Manchester syntax for OWL is a more usable notation than competing alternatives.

Acknowledgement Gem Stapleton was funded by a Leverhulme Trust Research Project Grant (RPG-2016-082).

References

1. OWL 2 Web Ontology Language Manchester Syntax (Second Edition): W3C WG Note 11. https://www.w3.org/TR/owl2-manchester-syntax/, acc. May 2017
2. The OWL 2 Web Ontology Language. http://www.w3.org/TR/owl2-overview/ (2016), accessed April 2016
3. Baader, F., Calvanese, D., McGuinness, D., Nadi, D., (eds), P.P.: The Description Logic Handbook. CUP (2003)
4. Horridge, M., Bail, S., Parsia, B., Sattler, U.: The cognitive complexity of OWL justifications. In: ISWC. pp. 241–256. Springer (2011)
5. Horridge, M., Parsia, B., Sattler, U.: Laconic and precise justifications in OWL. In: ISWC. pp. 323–338. Springer (2008)
6. Horridge, M., Patel-Schneider, F.: Manchester syntax for OWL 1.1. In: 4th international workshop OWL: Experiences and Directions (2008)
7. Hou, T., Chapman, P., Blake, A.: Antipattern comprehension: an empirical evaluation. In: FOIS. pp. 211–224. IOS Press (2016)
8. Kalyanpur, A., Parsia, B., Sirin, E., Hendler, J.: Debugging unsatisfiable classes in OWL ontologies. Web Semantics 3(4), 268–293 (2005)
9. Keet, C.: The use of foundational ontologies in ontology development: An empirical assessment. In: ESWC. pp. 321–335. LNCS 6643. Springer (2011)
10. Liang, K., Zeger, S.: Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22 (1986)
11. Musen, M.: The Protégé Project: A look back and a look forward. AI Matters 4(1) (2015)
12. Neuhaus, F., Vizedom, A., Baclawski, K., Bennett, M., Dean, M., Denny, M., Grueninger, M., Hashemi, A., Longstreth, T., Obrst, L.: Towards ontology evaluation across the life cycle. Applied Ontology 8(3), 179–194 (2013)
13. Purchase, H.: Experimental Human Computer Interaction: A Practical Guide with Visual Examples. Cambridge University Press (2012)
14. Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., Wroe, C.: OWL Pizzas: Practical Experience of Teaching OWL-DL. In: Engineering Knowledge in the Age of the SemanticWeb, LNCS, vol. 3257, pp. 63–81. Springer (2004)
15. Sarker, M., Krisnadhi, A., Carral, D., Hitzler, P.: Rule-based OWL modeling with ROWLTab Protégé plugin. In: ESWC. pp. 419–433. LNCS 10249. Springer (2017)
16. Spivey, J.: The Z Notation: A Reference Manual. Prentice Hall (1989)
17. Warren, P., Mulholland, P., Collins., T., Motta: The usability of description logics: Understanding the cognitive difficulties presented by desciption logics. In: Extended Semantic Web Conference. pp. 550–564. Springer (2014)
18. Yan, J., Fine, J.: Estimating equations for association structures. Statistics in Medicine 23, 859–880 (2004)