Testing SHARP: Experimental Design

Introduction and Goals

The goal of this experiment is to verify certain predictions made with SHARP, a model that simulates human symbolic reasoning in the description logic $\mathcal{ALE}$. This will either yield trust in the model, or specific pointers on how to improve it.

Population and sample

Our population of interest consists of people aged between 20 and 60, who have experience with formal logic. The sample consists of both master and PhD students from Swedish universities that are selected on the basis of some questions about formal (description) logic.

To estimate the sample size, we looked at similar experiments in literature and we performed an estimation with certain computer software. In [1], [13] and [15] statistically similar experiments were performed with sample sizes: 338, 77 and 5571 subjects respectively. Using the G*Power 3.1 software we estimated the sample size to be at least $N = 54$ subjects for a paired t-test.[1] Based on this we aim for a sample size of 60 subjects.

We offered a 100 SEK voucher as a reward.[2]

Material

The ABoxes from the appendix are presented one by one in random order using the online LimeSurvey software. Most of these ABoxes have an associated ABox which is a variant of the other, so as to compare the subjects' performance between the two.

The subjects are asked about the consistency of each ABox. The response format contains the three options: 'Consistent', 'Inconsistent', or 'I don't know'.

An instruction video is used explaining the syntax and semantics of the description logic $\mathcal{ALE}$; explanation about the tableau deduction rules for this logic is excluded.

Procedure and Tasks

The experiment is conducted in English by means of an online questionnaire made with LimeSurvey.

Firstly, the subjects are asked about their familiarity with logic in general and description logic in particular. Secondly, the subjects watch the introduction video. The purpose of the experiment is explained to the subjects. They are encouraged to make as few mistakes as possible by promising them a score

[1] With the following settings. 'Statistical test´: 'Means: Difference between two dependent means (matched pairs)´, 'Tail(s)': 'Two', 'Effect size dz': '0,5', 'α err prob': '0,05', 'Power ($1 - \beta$ err prob)': '0,95'

[2] The SuperPresentkort from GoGift.

of their performance at the end of the experiment; this should also discourage responding 'I don't know'. They are also encouraged to respond as quickly as possible.

They then get two practice cases ($\mathcal{A}_{practice0}$ and $\mathcal{A}_{practice1}$) to make sure what is expected of them and to reduce learning effects during the rest of the experiment. Additionally, these ABoxes serve to illustrate the easiest and hardest example that the subjects will encounter.

The subjects are then shown the twenty ABoxes one by one in randomized order. There is a mandatory one-minute break after each five ABoxes to reduce fatigue effects. The time difference between presenting the ABox and when the subject presses a button is measured. For every fifth ABox, the subject is asked to score the experienced difficulty of the ABox inconsistency task on a five-point Likert response format. The subjects are not allowed to use pen, paper or other tools.

Hypotheses

Certain hypotheses follow from the simulation results in chapter 6 in [6]. The ABoxes referred to can be found in the appendix.

1. The inference times for deciding inconsistency of $\mathcal{A}_{end0}$ and $\mathcal{A}_{end1}$ are identical.
2. To decide the inconsistency of $\mathcal{A}_{rnd0}$ requires less inference time than $\mathcal{A}_{rnd1}$.
3. The inference times for deciding inconsistency of $\mathcal{A}_{oin0}$ and $\mathcal{A}_{oin1}$ are identical.
4. The inference times for deciding inconsistency of $\mathcal{A}_{nin0}$ and $\mathcal{A}_{nin1}$ are identical.
5. The inference time for deciding the inconsistency of $\mathcal{A}_{nse0}$ is smaller than for $\mathcal{A}_{nse1}$.
6. Inference time scales exponentially with the size of ABoxes $\mathcal{A}_{ts0}$ to $\mathcal{A}_{ts4}$.
7. $\mathcal{A}_{spr0}$ shows less spread in inference times than $\mathcal{A}_{spr1}$.
8. By tweaking the subsymbolic parameters it is possible to fit the simulation results to the empirical data.
9. Experienced difficulty is highly correlated with inference time.
10. The order of complexity induced by SHARP's predicted inference times is identical to the order of complexity induced by the empirically established inference times.[3]

[3]Note that this order appears to be robust under changes of ACT-R's subsymbolic parameters.

The data gathered to test each two hypotheses above is mostly mutually exclusive, so that the data is not used to test multiple hypotheses simultaneously; exceptions to this are hypotheses 8, 9 and 10 for which data is used that is the same as is used to test the other hypotheses. This means that the hypotheses are not completely independently tested in this experiment.

Analysis

As we are interested in the characteristics of the thought process with which humans decide $\mathcal{ALE}$ ABox inconsistency, we select only the data that supports that this task is actually being performed: if a subject makes too many mistakes, all the data generated by this subject is excluded from further analysis. To achieve this, we assign a minus point to an incorrect answer, one point to a correct answer and zero points to the answer 'I don't know', after which participants should score at least 10 points to qualify their data for analysis. All a subject's data that meets this criterion is included in the analysis, even the possible wrong answers. The reason for this is that, based on the subject's total performance, we have enough evidence to believe that the correct processes are happening inside that subject's mind; the mistakes made are therefore likely not attributed to systematic failures in this these processes, but probably have different origins (mouse slips, misreadings, memory effects, etc.).

Hypotheses 1 to 5 are analysed by paired t-tests, with significance level $\alpha = 5\%$ and using the Holm-Bonferroni method to control the family-wise error rate. Hypothesis 6 is analysed by means of a regression to the exponential function $f(x) = a * e^{b*x}$, with a and b the parameters to be estimated. The adjusted R squared value of this fit will give an indication of how accurately an exponential function describes the empirically established scaling.

To test hypothesis 7, an F-test of equality of variances is performed. Hypothesis 8 is tested by manually tweaking four of ACT-R's parameters [4] and to minimise the mean squared error between the mean inference times of both the empirical and simulated data for all ABoxes except $\mathcal{A}_{practice0}$, $\mathcal{A}_{practice1}$, $\mathcal{A}_2$, $\mathcal{A}_8$ and $\mathcal{A}_{17}$. The goodness of this fit is then evaluated by calculating the Pearson correlation coefficient for the empirical and simulated data for the ABoxes $\mathcal{A}_2$, $\mathcal{A}_8$ and $\mathcal{A}_{17}$.

Hypothesis 9 is also tested by calculating the Pearson correlation coefficient. Hypothesis 10 is tested by calculating Spearman's rank correlation coefficient for the empirical and the simulated data, using the standard $\frac{m^3-m}{12}$ correction terms for possibly tied ranks m being the repeated rank.

References

[1] D. Almaleki. "Examinee Characteristics and their Impact on the Psychometric Properties of a Multiple Choice Test According to the Item Re-

[4] Namely, the utility noise, retrieval threshold, decay and instantaneous noise

sponse Theory (IRT)". In: *Engineering, Technology and Applied Science Research* 11.2 (Apr. 2021), pp. 6889–6901.

[2] F. Bagheri and A. Ghanizadeh. "Critical Thinking and Gender Differences in Academic Self-regulation in Higher Education". In: *Journal of Applied Linguistics and Language Research* 3.3 (2016), pp. 133–145.

[3] E.P. Brandon. "Gender and Competence in Deductive Logic, Seminar paper". In: (1989).

[4] M. Brunner, S. Krauss, and M. Kunter. "Gender differences in mathematics: Does the story need to be rewritten?" In: *Intelligence* 36 (2008), pp. 403–421.

[5] G. Cohen. "Inferential reasoning in old age". In: *Cognition* 9 (1981), pp. 59–72.

[6] J. T. Fokkens. *Modelling the Logical Mind*. Gothenburg, 2023.

[7] A.S. Gilinsky and B.B. Judd. "Working Memory and Bias in Reasoning Across the Life Span". In: *Psychology and Aging* 9.3 (1994), pp. 356–371.

[8] V.P. Hamm and L. Hasher. "Age and the Availability of Inferences". In: *Psychology and Aging* 7.1 (1992), pp. 56–64.

[9] M. Jin et al. "Comparison Study on High School Students' Mathematics Belief Systems between Han and Chaoxian Nationality". In: *Journal of Mathematics Education* 1.1 (2010), pp. 138–151.

[10] J.D.Rich Jr., W. Fullard, and W. Overton. "The Relationship Between Deductive Reasoning Ability, Test Anxiety, and Standardized Test Scores in a Latino Sample". In: *Hispanic Journal of Behavioral Sciences* 33.2 (2011), pp. 261–277.

[11] M. Levels, J. Dronkers, and G. Kraaykamp. "Immigrant Children's Educational Achievement in Western Countries: Origin, Destination, and Community Effects on Mathematical Performance". In: *American Sociological Review* 73 (2008), pp. 835–853.

[12] L.L. Light and E.M. Zelinski. "Adult Age Differences in Reasoning From New Information". In: *Journal of Experimental Psychology: Learning, Memory and Cognition* 8.5 (1982), pp. 435–447.

[13] M. P. Michaelides, M. Ivanova, and C. Nicolaou. "The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items". In: *International Journal of Testing* 20.3 (2020), pp. 187–205.

[14] C. Primi et al. "Are there gender differences in cognitive reflection? Invariance and differences related to mathematics". In: *Thinking I& Reasoning* 24.2 (2018), pp. 258–279.

[15] S. D. Schneid et al. "Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting". In: *Medical Education* 48.10 (2014), pp. 1020–1027.

[16] Z. Zhu. "Gender differences in mathematical problem solving patterns: A review of literature". In: *International Education Journal* 2 (Aug. 2007), pp. 187–203.

Appendix

The following 22 ABoxes will be presented to the subjects. A * indicates that the ABox is inconsistent. There are as many consistent as inconsistent ABoxes.

- $\mathcal{A}_{practice0} = \{a : A,\ b : \neg A,\ a : \neg B\}$
- $\mathcal{A}^*_{practice1} = \{a : \forall s.(\neg A \sqcap B),\ (b, a) : r,\ c : (\neg B \sqcap C),\ b : \forall r.\exists s.(A \sqcap B),\ a : (B \sqcap \neg C),\ (b, c) : s\}$
- $\mathcal{A}^*_{end0} = \{a : A,\ a : B,\ a : \neg A\}$
- $\mathcal{A}^*_{end1} = \{a : A,\ b : B,\ a : \neg A\}$
- $\mathcal{A}_{rnd0} = \{a : \exists r.A,\ a : \forall s.\neg A,\ a : \exists r.B\}$
- $\mathcal{A}_{rnd1} = \{a : \exists r.A,\ a : \forall s.\neg A,\ a : \exists s.B\}$
- $\mathcal{A}^*_{oin0} = \{a : A \sqcap B,\ a : \neg B\}$
- $\mathcal{A}^*_{oin1} = \{a : \neg B,\ a : A \sqcap B\}$
- $\mathcal{A}_{nin0} = \{a : A,\ a : B\}$
- $\mathcal{A}_{nin1} = \{a : \neg A,\ a : \neg B\}$
- $\mathcal{A}^*_{nse0} = \{a : (A \sqcap (\neg A \sqcap (B \sqcap C)))\}$
- $\mathcal{A}^*_{nse1} = \{a : (B \sqcap (C \sqcap (\neg A \sqcap A)))\}$
- $\mathcal{A}_{ts0} = \{a : \exists r.A \sqcap \exists r.B\}$
- $\mathcal{A}_{ts1} = \{a : \exists r.(\exists r.A \sqcap \exists r.B) \sqcap \exists r.B\}$
- $\mathcal{A}_{ts2} = \{a : (\exists r.A \sqcap \exists r.B) \sqcap \forall r.(\exists r.A \sqcap \exists r.B)\}$
- $\mathcal{A}_{ts3} = \{a : (\exists r.A \sqcap \exists r.B) \sqcap \forall r.(\exists r.(\exists r.A \sqcap \exists r.B) \sqcap \exists r.B)\}$
- $\mathcal{A}_{ts4} = \{a : ((\exists r.A \sqcap \exists r.B) \sqcap \forall r.((\exists r.A \sqcap \exists r.B) \sqcap \forall r.(\exists r.A \sqcap \exists r.B)))\}$
- $\mathcal{A}^*_{spr0} = \{a : (A \sqcap (B \sqcap (C \sqcap (D \sqcap \neg A))))\}$
- $\mathcal{A}^*_{spr1} = \{a : (A \sqcap B),\ a : (B \sqcap C),\ a : (C \sqcap D),\ a : (D \sqcap \neg A)\}$
- $\mathcal{A}_{17} = \{a : (\neg A \sqcap (B \sqcap \neg C)),\ b : (\neg B \sqcap (C \sqcap A)),\ b : (A \sqcap C),\ a : B\}$
- $\mathcal{A}^*_2 = \{a : A,\ a : \neg A,\ b : \forall r.A\}$
- $\mathcal{A}^*_8 = \{a : \exists r.\exists s.\neg A,\ a : \forall r.\forall s.A,\ b : \exists s.\neg B\}$

ABoxes $\mathcal{A}_{practice0}$ and $\mathcal{A}_{practice1}$ are presented first to the subjects to reduce learning effects; they are chosen to be one of the simplest and one of the most complex ABoxes the subjects will encounter in the experiment, thereby 'calibrating' the Likert response format for the subjective difficulty of the ABoxes.

ABoxes $\mathcal{A}_{17}$ and $\mathcal{A}_2$ are selected because they have a very high and a very low value according to the complexity measure $\mathrm{CM}^{\mathrm{SRP}}_{avg}$ respectively; $\mathcal{A}_8$ has a value close to the average of those two values. These last three ABoxes are somewhat less artificially constructed, because they are not designed to test specific hypotheses. Instead, they are designed by manual construction after randomly selecting the number of negations, conjunctions, restrictions and their consistency, thereby aiming to form an unbiased sample drawn from the population of all ABoxes.

On including a question about gender

The literature seems inconclusive about whether (and if, how) gender differences affect performance on logico-deductive reasoning tasks; although slight difference in favour of males do exist, these differences can be explained by (a wide variety of) factors other than biological sex. Firstly, [2] find no evidence of gender affecting performance on both the Watson-Glaser's Critical Thinking Appraisal and the Self-Regulation Trait Questionnaire. Also [10] find no evidence for gender affecting performance on the Selection task, TerraNova examinations and PSSA assessments, most of which contain mathematical reasoning tasks. Moreover, [3] concludes that "the males we have tested do not perform any differently from the females on tests of deductive logical competence".

On the other hand, some other papers do report gender differences on deductive thinking, for example [14] reports slight advantages for males on performance on the Cognitive Reflection Test, although these differences are probably attributed to factors like math anxiety. In [4] the authors conclude that gender differences in mathematical reasoning exist, but that they can be explained by the factors 'general cognitive ability' and 'specific mathematical ability', suggesting that sex is not a relevant factor affecting performance in mathamatical reasoning. Similar conclusions are reached in [16], where even more factors are taken into consideration, suggesting that sex alone does not influence performance on mathematical reasoning tasks. In light of the above, we decide to include a question about gender to be able to tell whether we reach a representative sample of the population.

On including a question about age

The literature seems to suggest that reasoning abilities change with age, although again no complete consensus is reached. [12] reports only moderate change of reasoning abilities with age. [8] reports, however, no age-dependence of performance on two different deductive reasoning tasks.

On the other hand, in [7] it is reported that age affects reasoning performance negatively, even after correcting for working memory and vocabulary.

Similarly, [5] reports experiments showing both a significant decline of reasoning performance with age and a decline in extracting information from text. The literature, therefore, seems to suggest that reasoning performance depending on age cannot be ruled out. For this reason, a question about the subject's age will be added to the survey.

On including a question about nationality

As opposed to the above two topics, not much research has been done on how nationality affects deductive reasoning. One study compared certain beliefs about mathematics between people from Han and Chaoxian nationalities [9] and found substantial differences, although the methods were mostly qualitative. [11] reports significant differences of immigrant students' mathematical performance based on their different background nationalities. So, although the evidence is scant and circumstantial, it is probably wise to make sure our sample is balanced in terms of subjects' nationalities to ensure generalisability over this variable. Private correspondence confirmed this.