The impact of proof steps sequence on proof readability — experimental setting*

Karol Pąk[1] and Aleksy Schubert[2]

[1] Institute of Informatics, University of Białystok,
ul. K. Ciołkowskiego 1M, 15-245 Białystok, Poland
pakkarol@uwb.edu.pl
[2] Institute of Informatics, University of Warsaw
ul. S. Banacha 2, 02–097 Warsaw, Poland
alx@mimuw.edu.pl

Abstract. Different sequences of proof steps may result in different experiences of the overall formal proof clarity. There is a conjecture that sequence in which more steps refer to the content of the preceding statement (1) is more comprehensible than the one in which references span longer distances (2). We studied the claim in experimental setting where subjects indicated their reading preference with two versions of the same proof. The difference between their reports indicates that contrary to the conjecture, proofs of the kind (2) can give cognitive advantage in a statistically significant way.

1 Introduction

The readability of formal proofs, as in the case of computer programs, has significant impact on the process of proof formalisation development and maintenance. This phenomenon has already been observed in the programme of Nicolas Bourbaki — the resulting proofs written in formalist fashion were perceived as too obscure and the whole project gained the opinion of unwieldy [12]. Moreover, there are numerous situations in which proof scripts are indeed read, e.g. when a proof development is used as a library of facts [4], when users try to learn new proving techniques [8], or when users want to strengthen theorems [2].

There are many factors that have an impact on readability. In this paper we focus on the psychological readability factors associated with locality of reference. They can be summarised in a high-level statement due to Behaghel that *elements that belong close together intellectually should be placed close together* [1] and this statement concerns particularly systems such as Mizar [3] and Isabelle/Isar [13], which try to construct proofs close to natural language ones.

Discussions among Mizar Mathematical Library developers [9] led to the conclusion that proof steps that refer to the preceding step are perceived as more comprehensible. This led to construction of algorithms that rearranged

* The paper has been supported by the resources of the Polish National Science Centre granted by decision n° DEC-2012/07/N/ST6/02147.

Mizar proof scripts so that they maximise the number of such references [9] and further analysis of the complexity of the rearrangement process [10,11].

One important step in understanding the applicability of the techniques is to assess their impact on real readers. This paper is an attempt in this direction. We investigate the influence of one of the employed techniques — changing the sequence of steps so that the number of `then` steps in continuous sequences is maximal. In our study we give two versions of the same proof to a substantial number of subjects and obtain their readability assessment.

The paper is constructed as follows. In Section 2 we present the design of our experiment. Next, we present the statistical methods we employ and present the way our experiment was executed. This is done in Section 3. The results of the experiments are presented in Section 4 and discussed in Section 5. Many people helped us in organising the experiment so we acknowledge them in Section 6.

2 The Design of the Experiment

2.1 The General Idea of the Experiment

It is not straightforward to devise an experiment in which two text structures are compared for readability. First of all, the feature of readability is subjective so a comparing judgement of a single person must be the basic unit of the measurement. Moreover, one must note that it is not directly valid to ask subjects to compare texts of different proofs for readability since then there may be many other structural factors that must be recognised and well understood to make such a comparison informative. Therefore, our first design choice is that one subject at one moment compares two proofs of the same mathematical statement and the result of the judgement is accounted in the study.

This assumption brings one major difficulty in. Comparison of essentially the same proofs requires reading two texts that convey the same content. The texts are by the nature of the reading process read in sequence. Clearly, the process of reading of the second text is strongly influenced by that of the first text. We can now expect two effects:

1. either the subjects will judge the second text as more readable due to the fact that it reflects something that is already known,
2. or the subjects will judge the second text as less readable since they already familiarised themselves with the structure of the first one and perceive the text of the second one as alien and so less favourable.

It turns out that the second effect is much stronger than the first one (we provide statistical evidence for this in Section 4.2). Therefore, we decided that meaningful responses are only those that counter this strong trend and that such responses indicate visible cognitive advantage to the subjects that behave in this way.

This assumption requires us to create two contrasting situations in which subjects can counter this trend. We decided here to distribute among subjects two variants of the test: one that presents a version A of the proof first and then a version B, and one that presents the version B first and A subsequently.

One more important issue to deal with here is making sure that the tests are not neglected by the subjects. That is why we decided that the test should have a form of an assignment in which the analysis of the proofs is required to give an answer. In the end we decided that a good way to achieve this is to introduce a mistake into the proof and instruct subjects to find it. The subjects were to obtain an award for fulfilling the task.

2.2 The Actual Tests

The assignments given to the students were based upon a faulty proof of the wrong set-theoretic formula

$$(X \cup Y) \backslash (X \cap Y) = (X \backslash Y) \cap (Y \backslash X).$$

It has a correct counterpart

$$(X \cup Y) \backslash (X \cap Y) = (X \backslash Y) \cup (Y \backslash X)$$

the proof of which can be formalised in Mizar as presented in Fig. 1 (version A of the proof). This proof was taken as the basis for the experiment. We devised another proof (see Fig. 2) of the same fact (version B) in which certain two steps were permuted, which is reflected by the changed line numbers (7, 8) on the left margins of the proofs. The actual internal structure of the two versions is the same and is presented in Fig. 3 where labels of the nodes correspond directly to the labels on the left margins of both the proof in version A and B.[3] In this way we can see that the proofs indeed differ only in the sequence of steps and the necessary Mizar syntax reorganisations. We expected that the version B will be perceived by subjects in our experiments as less readable than the version A (and this turned out to be actually false).

Since we wanted to obtain judgements of many subjects and it is difficult to gather significantly many Mizar experts in one place, we had to rewrite the proof texts in natural language so that they are digestible by people who are not familiar with Mizar syntax. Moreover, the subjects in the experiment were Polish native speakers so the assignments had to be prepared in that language.[4] We provide a faithful translation of the tests in Appendix A.

The two versions of the proof were printed on one page of an A4 leaf. Half of the tests had version A of the proof located on the left-hand side of the page and version B on the right-hand one and half of the tests had version B located on the left-hand side and version A on the right-hand one. To make distribution of tests quick we did not strictly keep the size of two resulting groups even. However, we made sure that the groups are of substantial size (see Section 4).

[3] Graph representation of proofs was defined by Pąk [9]. We invite interested readers to consult his publication for details.

[4] The original tests are available in the package with experiment data at the address `http://www.mimuw.edu.pl/~alx/proof-readability.zip`

```
    theorem
1:    (X ∨ Y) \ (X ∧ Y) = (X \ Y) ∨ (Y \ X)
    proof
2:    for x holds x in (X ∨ Y) \ (X ∧ Y) iff x in (X \ Y) ∨ (Y \ X)
      proof
3:      let x;
4:      thus x in (X ∨ Y) \ (X ∧ Y) implies x in (X \ Y) ∨ (Y \ X)
        proof
5:        assume A1: x in (X ∨ Y) \ (X ∧ Y);
6:        then not x in (X ∧ Y) by XBOOLE_0:def 5;
7:        then A2: not x in X or not x in Y by XBOOLE_0:def 4;
8:        x in X or x in Y by A1,XBOOLE_0:def 3;
9:        then x in (X \ Y) or x in( Y \ X) by A2,XBOOLE_0:def 5;
10:       hence x in (X \ Y) ∨ (Y \ X) by XBOOLE_0:def 3;
        end;
11:     thus x in (X \ Y) ∨ (Y \ X) implies x in (X ∨ Y) \ (X ∧ Y)
        proof
12:       assume x in (X \ Y) ∨ (Y \ X);
13:       then x in (X \ Y) or x in (Y \ X) by XBOOLE_0:def 3;
14:       then A3: x in X & not x in Y or
                   x in Y & not x in X by XBOOLE_0:def 5;
15:       then A4: x in (X ∨ Y) by XBOOLE_0:def 3;
16:       not x in (X ∧ Y) by A3,XBOOLE_0:def 4;
17:       hence x in (X ∨ Y) \ (X ∧ Y) by A4,XBOOLE_0:def 5;
        end;
      end;
      hence thesis by TARSKI:2;
    end;
```

Fig. 1. The Mizar proof the experiments were based upon, version A.

```
    theorem
1:    (X ∨ Y) \ (X ∧ Y) = (X \ Y) ∨ (Y \ X)
    proof
2:    for x holds x in (X ∨ Y) \ (X ∧ Y) iff x in (X \ Y) ∨ (Y \ X)
      proof
3:      let x;
4:      thus x in (X ∨ Y) \ (X ∧ Y) implies x in (X \ Y) ∨ (Y \ X)
        proof
5:        assume A1: x in (X ∨ Y) \ (X ∧ Y);
6:        then A2: not x in (X ∧ Y) by XBOOLE_0:def 5;
8:        A3: x in X or x in Y by A1,XBOOLE_0:def 3,def 4;
7:        not x in X or not x in Y by A3,A2,XBOOLE_0:def 4;
9:        then x in (X \ Y) or x in( Y \ X) by A3,XBOOLE_0:def 5;
10:       hence x in (X \ Y) ∨ (Y \ X) by XBOOLE_0:def 3;
        end;
11:     thus x in (X \ Y) ∨ (Y \ X) implies x in (X ∨ Y) \ (X ∧ Y)
        proof
12:       assume x in (X \ Y) ∨ (Y \ X);
13:       then x in (X \ Y) or x in (Y \ X) by XBOOLE_0:def 3;
14:       then A4: x in X & not x in Y or
                   x in Y & not x in X by XBOOLE_0:def 5;
15:       then A5: x in (X ∨ Y) by XBOOLE_0:def 3;
16:       not x in (X ∧ Y) by A4,XBOOLE_0:def 4;
17:       hence x in (X ∨ Y) \ (X ∧ Y) by A5,XBOOLE_0:def 5;
        end;
      end;
18:   hence thesis by TARSKI:2;
    end;
```

Fig. 2. A Mizar proof in which steps were permuted, version B.

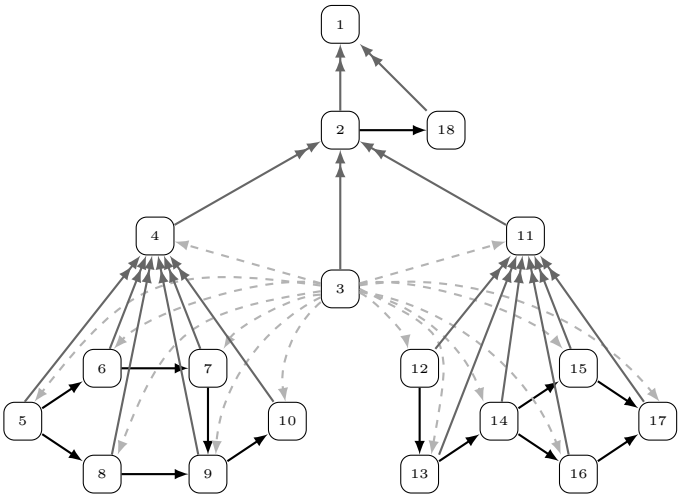

Fig. 3. The structures of the two versions of the proof.

3 The Execution of the Experiment

The experiments were conducted on students of informatics at Faculty of Mathematics, Informatics and Mechanics, University of Warsaw and Faculty of Mathematics and Computer Science, University of Białystok. They were executed in three rounds of two slightly different settings.

Round I took place in January 2015 and was conducted as a separate assignment in an exam with a slot of 10 minutes dedicated exclusively to the task of comparing the proofs. The award for students were additional points in the exam score. The topic of the exam was *Logic for Computer Scientists*. The subjects of the experiment were students of first year postgraduate studies. These students had studied the full curriculum of undergraduate informatics, including *Foundations of Mathematics* as taught on the first year of undergraduate studies. We can assume that their experience with formal systems, including programming languages, was good enough for this test.

The experiment was conducted in parallel in two lecture halls which contained groups of students of approximately 35 people each. The assignments were given in parallel and in the conditions of the exam, which guarantee the lack of communication between subjects. Therefore, we can assume that the results were independent.

The instruction of the assignment was as follows

> *You are presented two proofs of the same fact. Both have the same flaw. Choose the version of the reasoning for which you can easier perform the following task: please show the place where the flaw is located and describe in one sentence what is the reason of the mistake.*

Therefore, the students were given information that the two texts are essentially equivalent and they were instructed to indicate which of the texts gave them cognitive advantage.

Both versions of the proof were formulated using the same natural language phrases so that one formulation corresponds to the same kind of inference step.

Round II took place in February 2015 and was conducted in Białystok. The procedure and tests were the same as in the case of Round I. The topic of the exam was *Logic and Set Theory*. The subjects of the experiment were students of first year undergraduate studies of mathematics and informatics. We can assume that their experience with formal systems was fair and enough for this test, since they were taught specifically to prove facts using natural deduction format that was close to the one used in Mizar. Still, we obtained significantly more answers than in Round I which indicated that subjects did not understand the subject matter of the proof. The exam was taken in two groups of approximately 20 and 40 people respectively

Round III took place in January 2016 in Warsaw and was conducted as a separate assignment during blackboard classes with a slot of 10 minutes dedicated exclusively to the task of comparing the proofs. The award for students were additional points in their score of the classes. The topic of the classes was *Foundations of Mathematics*. The subjects of the experiment were students of first year undergraduate studies. The experience of the students with formal systems was limited. The experiment was conducted at the end of the course on foundations of mathematics so their understanding of the notion of proof should be satisfactory for the purpose of our procedure.

The experiment was conducted in 4 different groups of classes at different times. The size of the groups was approximately 10 people. However, we can assume that the measurements were independent since the students of the first year of undergraduate studies rarely communicate between different groups (all other classes are given in groups of the same composition) and the details of the assignment necessary to complete it successfully are not very specific and so very difficult to explain without having the actual test at hand. Therefore, we can safely assume that despite the assignments were not given in parallel their results are independent. Of course, the assignments within each of the group were executed so that subjects did not communicate one with another.

The instruction of the assignment was as follows

> *You are presented two proofs of the same fact. Find in them as many flaws as you can. In case a mistake occurs in both proofs, mark with a star the version in which you found it first. Is the structure of some of the proofs more readable for you?*

Therefore, the students were given information that the two texts are essentially equivalent and they were instructed to indicate which of the texts gave them cognitive advantage. Additionally, they were instructed to search for as many flaws as possible and mark places where the mistake was found first. We intro-

duced this change to get deeper confidence that the proofs were read thoroughly and check if this change has impact on the effects (E1) and (E2)

In this case the two versions of the proof were formulated in a different fashion. Moreover, as the proofs actually demonstrate equivalence, they divide into two parts: one for ($\Rightarrow$) and one for ($\Leftarrow$). We decided to have a different sequence of these parts in each version. The rationale behind these changes was to make line to line comparison more difficult and force subjects to perceive the structure of the proofs first and then judge their clarity.

We would like to point out that in Round I and II the instruction was less direct and referred to *easiness of the task* while in Round III the instruction appealed directly to the intuitive understanding of the term *readability*. However, we would like to stress that each of the formulations actually means that the subjects were instructed to indicate their cognitive preference for the two versions of the proof, i.e. which of the texts is easier to digest.

3.1 The Mann-Whitney Test

For our verification, we use the non-parametric Mann-Whitney test (also called Willcoxon test), as the use of standard parametric tests assumes the distributions are normal. This test is used to statistically refute a null hypothesis H_0 in favour of its negation, i.e. the alternative hypothesis H_A.

This kind of test can be applied when the following prerequisites are met:

Hypotheses The null hypothesis H_0 for the study should be formulated in the fashion such that the probability of an observation from the population G_1 exceeding an observation from the second population G_2 is less then or equal the probability of an observation from G_2 exceeding an observation from G_1, i.e. $P(score(G_1) > score(G_2)) \leq P(score(G_2) > score(G_1))$.
Consequently, the alternative hypothesis H_A is that the probability of an observation from the group G_1 exceeding an observation from the group G_2 is greater than the probability of an observation from G_2 exceeding an observation from G_1, i.e. $P(score(G_1) > score(G_2)) > P(score(G_2) > score(G_1))$.

Uniformity of populations We should also be able to assume that in case the distribution of results from G_1 and G_2 is the same, the probability of an observation from G_1 exceeding one from G_2 is equal to the probability of an observation from G_1 exceeding one from G_2. In other words the background of both groups is the same and their members were chosen randomly from the point of view of the experiment.

Independence We should be able to assume that the observations in G_1 and G_2 are independent.

Comparable responses The scores should be numeric so they can be easily compared one with another so a single ranking of the subjects can be formed.

The basic idea of the test is that we build a ranking list of the scores from the lowest to the highest (with possible ties). Then we try to get a numerical representation on which of the groups has more results close to the top one.

Consider the following pattern

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=1}^{n_2} R_i$$

where U is the value of the Mann-Whitney U test, n_1 is the sample size of G_1, n_2 is the sample size of G_2 and R_i is the ranking position of the i-th score in the group G_2. We can see that the more scores of G_2 are closer to the top the number U is greater.

Contemporary statistics packages return a normalised value Z of the statistical test that is computed according to the pattern

$$Z = \frac{U - m_U}{\sigma_U}$$

where m_U is the mean of U and σ_U is the standard deviation of U, which are in turn computed using the formulas

$$m_U = \frac{n_1 n_2}{2}, \qquad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

However, the formula for σ_U must be corrected in case ties (caused by equal scores of certain subjects) are present and it is then

$$\sigma_U = \sqrt{\frac{n_1 n_2}{12}((n+1) - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{n(n-1)})}$$

where k is the number of tied positions, t_i is the number of subjects that attained the i-th tied rank, and $n = n_1 + n_2$.

3.2 Stouffer-Lipták Method

Since the tests given to the subjects were different and the students were of different maturity we cannot directly sum the groups that took part in the two rounds. Therefore we need a method to combine results of three experiments. For this we use a meta-analysis technique called Stouffer-Lipták method [7]. In this method we compute a combined Z value based upon Z-values obtained from the Mann-Whitney test using the pattern (tailored to the case where three experiments are combined):

$$Z = \frac{w_1 Z_1 + w_2 Z_2 + w_3 Z_3}{\sqrt{w_1^2 + w_2^2 + w_3^2}}$$

where Z_1, Z_2, Z_3 are the cumulative normal distribution values that match the probabilities $1-p_1, 1-p_2$ and $1-p_3$ with p_1, p_2, p_3 being p-values obtained in two sub-experiments while w_1, w_2, w_3 are weights of the sub-experiments that can be used to accommodate for different sample sizes (note that sub-experiment with

a bigger number of subjects should weigh more than the sub-experiment with the smaller one, according to Lipták square roots of sample sizes are optimal here).

The p-value for Z-value obtained in this way is $1-v$ where v is the probability that a normally distributed random number will be less than that this Z-value.

3.3 The Binomial Test

The binomial test is used to check which of two outcomes has greater probability. We assume here that our process has n outcomes represented by random variables $X_1, \ldots, X_n$. These outcomes are either 0 (failure) or 1 (success) and now $Y = \sum_{i=1}^{n} X_i$ represents the number of successes in the process. We observe now that

$$P(Y = k) = \binom{n}{k} q^k (1-q)^{n-k} \text{ for } k \in \{0, \ldots, n\}$$

where q is the probability of success (i.e. that $X_i = 1$). We can now use the value $Q_{n,q}(\alpha)$ of the quantile of order α for the distribution to reject H_0 telling that $q \geq p_0$ versus the alternate hypothesis H_A that $q < p_0$ if and only if $Y \leq Q_{n,p_0}(\alpha)$ where α is the desired level of confidence. Typically $p_0 = 0.5$ and we actually check that the probability of success is greater than the one of failure.

4 Results of the Experiment

Since we use the Mann-Whitney test, we have to make sure that the assumptions of the test are met. First of all we have to ensure that the scores are numeric. In our setting we give the score 1 to those assignments that clearly indicate the right-hand side of the test. The other score, 0, is given when the left-hand side is indicated as preferred, when there is no preference and also when the assignment was solved incorrectly by the subject.

We can now compare two populations depending on what was the content of the left-hand side (i.e. first side) of obtained assignment: (G_1) these who obtained version A there with (G_2) those who obtained version B. We understand that subjects in G_1, by choosing the right-hand side in their test, judged that version B gives them cognitive advantage while subjects in G_2, by choosing the same right-hand side, judged that version A gives them this. Our main statistical hypotheses can be formulated in the following way

- H_0^a: the probability that a subject from G_2 has the score 1 is greater or equal to the probability that a subject from G_1 does it;
- H_A^a: the probability that a subject from G_2 has the score 1 is smaller than the probability that a subject from G_1 does it.

We have two more populations, G_1' with subjects that obtained 0 in the assessment of assignments and G_2' with subjects that obtained 1 there. This time we assign score 1 to all subjects of the groups. We can now formulate two other hypotheses

– H_0^f: the probability that a subject from G'_2 has the new score 1 is greater than the probability that a subject from G'_1 does it;
– H_A^f: the probability that a subject from G'_2 has the new score 1 is smaller or equal to the probability that a subject from G'_1 does it.[5]

Let us take a look at the assumptions of the Mann-Whitney test. This formulation of the two pairs of hypotheses directly falls under the format prescribed in Section 3.1. In each case the leaves with assignments were distributed in sequence without any taking into account the subjects' history (in particular scores in earlier exams) so we can assume that the populations of all the groups had the same background. The assignments in each case were worked out by the subjects in the exam-like conditions, i.e. without any communication with their colleagues solving the same assignment. Therefore, we may assume that the measures for each of the assignment were taken independently. At last, the final scores were numeric so they fulfil the assumption of comparability of responses.

Round I

Version	size	mean	sd
V. A first (G_1)	33	0	0
V. B first (G_2)	43	0.1627907	0.3735437
Total	76	0.09210526	0.2910959

Round II

Version	size	mean	sd
V. A first (G_1)	34	0.08823529	0.2879022
V. B first (G_2)	31	0.09677419	0.3005372
Total	65	0.09230769	0.2917125

Round III

Version	size	mean	sd
V. A first (G_1)	17	0.2941176	0.4696682
V. B first (G_2)	17	0.1764706	0.3929526
Total	34	0.2352941	0.4305615

Fig. 4. The basic statistics of the experiment.

4.1 General Overview of Results

The overall number of subjects who took part in Round I of the experiment was 76. Out of the number, 33 people got the assignment leaf with version A of the proof on the left-hand side and 43 people got the assignment with version B on the left-hand side. In the case of Round II the total number of subjects was 65 with 34 assignments where version A of the proof was on the left-hand side and 31 where version B was there. In Round III the total number of subjects was 34 and each of the groups had 17 people. We have to note here that the uneven number of subjects in the two groups does not prevent applicability

Round I

Version	score 0	score 1
V. A first (G_1)	36	7
V. B first (G_2)	33	0

Round II

Version	score 0	score 1
V. A first (G_1)	28	3
V. B first (G_2)	31	3

Round III

Version	score 0	score 1
V. A first (G_1)	14	3
V. B first (G_2)	12	5

Fig. 5. The distribution of the assigned scores.

[5] The small difference in the formulation of the hypotheses H_0^a, H_A^a and H_0^f, H_A^f is probabilistically negligible, but it is easier to handle in computations by R.

of the Mann-Whitney statistical test, it only impairs its accuracy.

The basic statistics of the groups are presented in Fig. 4. We can see the sizes of the groups (column size) as well as the mean score (column mean) and standard deviation of the scores (column sd). We immediately see that the mean of the score in the group G_1 during Round I is 0. This is due to the fact that none of the subjects returned the assignment with indicated version A on the right-hand side. We can confirm this by looking in the table displayed in Fig. 5 where the numerical scores of all the six subgroups are presented.

4.2 First Text Is Favoured

Round	p-value
Round I	$3.208984 \cdot 10^{-14}$
Round II	$2.482325 \cdot 10^{-12}$
Round III	0.001467528
Stouffer-Lipták	0

Fig. 6. The p-values for the test on which side is favoured.

We evaluate first the hypotheses H_0^f versus H_A^f as they give the background for the whole experiment. For this we use a more direct *binomial test* based upon the Bernoulli distribution. The p-values for this test are $3.208984 \cdot 10^{-14}$, $2.482325 \cdot 10^{-12}$, and 0.001467528 for the samples taken in Round I, Round II, and Round III respectively. The combined p-value obtained with the Stouffer-Lipták method is so small that it is below the precision range of our tools. Therefore the overall result is 0.[6] All the figures are gathered in the table presented in Fig. 6. Summing up, we obtained a prevalent evidence that the left hand side is favoured in such comparison tests.

4.3 Readability Assessment

Round	Z-value	p-value
Round I	2.41645	0.007836335
Round II	0.117872	0.4530845
Round III	−0.7966275	0.7871663
Stouffer-Lipták	1.31315	0.0945662

Fig. 7. The statistics of the readability experiment.

The picture in the readability assessment is more complicated. Already a look at the table in Fig. 5 reveals that in Round I the preference for version B was strong, in Round II the scores were even and in Round III turned in the opposite direction. This is reflected in obtained Z-values, which are 2.41645, 0.117872 and −0.7966275 respectively. The negative value of the second statistics reflects the fact that we are closer to rejecting H_A^a. The p-values obtained in the tests are 0.007836335, 0.4530845 and 0.7871663 respectively.[7] This shows that the first test gives a strong statistically significant result, which supports our claim that the proofs in version A format can give cognitive advantage at the statistically significant level, actually the significance here is higher than 99%. This evidence is so strong that even when we combine the three rounds with Stouffer-Lipták

[6] The results were obtained with help of the standard R function `binom.test`.

[7] The results were obtained with help of the R package **coin** [5,6] with its heuristics `mid-ranks` to handle ties in input data.

test we obtain result of reasonable significance at the level over 90%. These results are summarised in the table presented in Fig. 7.

5 Discussion

The fact that the left-hand side of tests devised in our experiment is strongly favoured has very strong support in our data. The case of readability is more curious. We can see that Round I resulted in strong preference for version B of the proof, which is more convoluted. However slight changes to the conditions in Rounds II and III gave a picture that tends to support the opposite claim. We can observe that Round I (during an exam) was likely more stressful than Round III, and this could make the short-term memory less effective. As a result, the bigger number of labels present in version B could turn out to be advantageous for the understanding process. Some of the subjects shared with us this opinion on their assessments. However, this line of argument is not supported by the result of Round II, which was also taken under exam conditions.

One more phenomenon that can also play role here is that the proof in version A indeed more often referred to the previous step. However, each step depends not only on the directly preceding one, but also on some other ones. In the case of version A we have a step (line 8. in Fig. 1) which required subjects to refer 3 steps back, while the proof in version B required subjects to refer at most 2 steps back. Further investigations are necessary to check which of the reasons actually holds. However, this investigation shows that the tools to improve readability based upon the principle used here should be used cautiously. Readability is a feature that depends highly on individual abilities of proof readers so the tools to improve it should not enforce any fixed method of proof improvement, but rather offer a number possibilities that can be chosen by users.

The subjects in this experiment actually were not trained in extensive reading of mathematical proofs. We can say that they were accidental proof readers. This experiments showed that they prefer proofs in version B. Still, experienced users, as implied by our initial expectations, seem to prefer proofs structured as in version A. It may be that the experienced users learn this preference as they gain experience. This conjecture is also worth further investigation.

6 Acknowledgements

A number of people helped us to execute the experiment. We would like to thank professors Jerzy Tyszkiewicz from University of Warsaw and Krzysztof Prażmowski from University of Białystok for allowing us to take the test during exams of their classes. The tests were also part of classes conducted by Daria Walukiewicz-Chrząszcz, Jacek Chrząszcz, and Piotr Wasilewski so we are also grateful for their help. These people were also very kind to discuss with us the design of the experiment which helped us in its better organisation.

References

1. Otto Behaghel. Beziehungen zwischen umfang und reihenfolge von satzgliedern. *Indogermanische Forschungen*, (29):110–142, 1909.
2. Georges Gonthier. Formal proof the four-color theorem. *Notices of the AMS*, 55(11):1382–1393, 2008.
3. Adam Grabowski, Artur Korniłowicz, and Adam Naumowicz. Four decades of Mizar. *Journal of Automated Reasoning*, 55(3):191–198, October 2015.
4. Adam Grabowski and Christoph Schwarzweller. Revisions as an essential tool to maintain mathematical repositories. In *Proceedings of the 14th Symposium on Towards Mechanized Mathematical Assistants: 6th International Conference*, Calculemus '07 / MKM '07, pages 235–249, Berlin, Heidelberg, 2007. Springer-Verlag.
5. Torsten Hothorn, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28(8), 2008.
6. Torsten Hothorn, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. *Package 'coin': Conditional Inference Procedures in a Permutation Test Framework*. CRAN, 2013.
7. T. Lipták. On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Közl.*, 3:171–196, 1958.
8. Adam Naumowicz and Czesław Byliński. Improving Mizar texts with properties and requirements. In Andrea Asperti, Grzegorz Bancerek, and Andrzej Trybulec, editors, *Mathematical Knowledge Management, Third International Conference, MKM 2004 Proceedings*, volume 3119 of *MKM'04, Lecture Notes in Computer Science*, pages 290–301, 2004.
9. Karol Pąk. The Algorithms for Improving and Reorganizing Natural Deduction Proofs. *Studies in Logic, Grammar and Rhetoric*, 22(35):95–112, 2010.
10. Karol Pąk. Improving Legibility of Natural Deduction Proofs is Not Trivial. *Logical Methods in Computer Science*, 10(3), 2014.
11. Karol Pąk. Improving Legibility of Formal Proofs Based on the Close Reference Principle is NP-Hard. *Journal of Automated Reasoning*, 55(3):295–306, 2015.
12. Ernst Snapper. The three crises in mathematics: Logicism, intuitionism and formalism. *Mathematics Magazine*, 52(4):207–216, 1979.
13. Makarius Wenzel. *The Isabelle/Isar Reference Manual*. University of Cambridge, 2013.

A English Translations of the Tests Used in the Experiment

In each of the two versions of the experiment the assignment text and the formulation of the statement repeat on each side of the page. Only the proof part is different. We give the repeating parts only once in this translation.

A.1 Version Used in January 2015

Assignment You are presented two proofs of the same fact. Both have the same flaw. Choose the version of the reasoning for which you can easier perform the following task: please show the place where the flaw is located and describe in one sentence what is the reason of the mistake.

Statement $(X \cup Y)\backslash(X \cap Y) = (X\backslash Y) \cap (Y\backslash X)$

Proof (version A) We show first that for each x the equivalence holds
$x \in (X \cup Y)\backslash(X \cap Y)$ if and only if $x \in (X\backslash Y) \cap (Y\backslash X)$.
For the proof from left to right:
- *A1:* Let us take any $x \in (X \cup Y)\backslash(X \cap Y)$.
- We immediately obtain that $x \notin (X \cap Y)$ (def. of set difference),
- *A2:* and further that $x \notin X$ or $x \notin Y$ (def. of set intersection).
- Additionally $x \in X$ or $x \in Y$ (see A1, def. of set sum)
- therefore $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (see A2, def. of set difference),
- which gives the expected $x \in (X\backslash Y) \cap (Y\backslash X)$ (def. of set intersection).

For the proof from right to left:
- Let us take any $x \in (X\backslash Y) \cap (Y\backslash X)$.
- We immediately obtain that $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (def. of set intersection),
- *A4:* and further $x \in X$ and $x \notin Y$ or $x \in Y$ and $x \notin X$ (def. of set difference),
- and further $x \in (X \cup Y)$ (def. of set sum).
- Additionally $x \notin (X \cap Y)$ (see A4, def. of set intersection)
- which gives the expected $x \in (X \cup Y)\backslash(X \cap Y)$ (def. of set difference).

From the proved equivalence and definition of set equality we immediately obtain the goal statement.

Proof (version B) We show first that for each x the equivalence holds
$x \in (X \cup Y)\backslash(X \cap Y)$ if and only if $x \in (X\backslash Y) \cap (Y\backslash X)$.
For the proof from left to right:
- *A1:* Lat us take any $x \in (X \cup Y)\backslash(X \cap Y)$.
- *A2:* We immediately obtain that $x \notin (X \cap Y)$ (def. of set difference),
- *A3:* Additionally $x \in X$ or $x \in Y$ (see A1, def. of set sum)
- Observe that $x \notin X$ or $x \notin Y$ (see A2, def. of set intersection)
- Consequently $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (see A3, def. of set difference),
- which gives the expected $x \in (X\backslash Y) \cap (Y\backslash X)$ (def. of set intersection).

For the proof from right to left:
- Let us take any $x \in (X\backslash Y) \cap (Y\backslash X)$.
- We immediately obtain that $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (def. of set intersection),
- *A3:* and further $x \in X$ and $x \notin Y$ or $x \in Y$ and $x \notin X$ (def. of set difference),
- and further $x \in (X \cup Y)$ (def. of set sum).
- Additionally $x \notin (X \cap Y)$ (see A3, def. of set intersection)
- which gives the expected $x \in (X \cup Y)\backslash(X \cap Y)$ (def. of set difference).

From the proved equivalence and definition of set equality we immediately obtain the goal statement.

A.2 Version Used in January 2016

Assignment You are presented two proofs of the same fact. Find in them as many flaws as you can. In case a mistake occurs in both proofs, mark with a star

the version in which you found it first. Is the structure of some of the proofs more readable for you?

Statement $(X \cup Y)\backslash(X \cap Y) = (X\backslash Y) \cap (Y\backslash X)$

Proof (version A) We show first that for each x the equivalence holds

$$x \in (X \cup Y)\backslash(X \cap Y) \text{ if and only if } x \in (X\backslash Y) \cap (Y\backslash X).$$

For the proof from left to right:

- [1] Let us take any $x \in (X \cup Y)\backslash(X \cap Y)$.
- We immediately obtain that $x \notin (X \cap Y)$ (def. of set difference),
- [2] and further that $x \notin X$ or $x \notin Y$ (def. of set intersection).
- Additionally $x \in X$ or $x \in Y$ (see [1], def. of set sum)
- therefore $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (see [2], def. of set difference),
- summing up, we obtain in the end that $x \in (X\backslash Y) \cap (Y\backslash X)$ (def. of set intersection).

For the proof from right to left:

- Let us take any $x \in (X\backslash Y) \cap (Y\backslash X)$.
- We immediately obtain that $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (def. of set intersection),
- [3] and then $x \in X \wedge x \notin Y$ or $x \in Y \wedge x \notin X$ (def. of set difference),
- [4] and then $x \in (X \cup Y)$ (def. of set sum).
- Additionally $x \notin (X \cap Y)$ (see [3], def. of set intersection)
- summing up, we obtain in the end that $x \in (X \cup Y)\backslash(X \cap Y)$ (see [4] and def. of set difference).

From the proved equivalence and definition of set equality we immediately obtain the goal statement.

Proof (version B) We show first that for each x the equivalence holds

$$x \in (X \cup Y)\backslash(X \cap Y) \text{ if and only if } x \in (X\backslash Y) \cap (Y\backslash X).$$

For the proof from right to left:

- Let us fix arbitrary $x \in (X\backslash Y) \cap (Y\backslash X)$.
- From this, we obtain that $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (def. of set intersection),
- [1] and then $x \in X \wedge x \notin Y$ or $x \in Y \wedge x \notin X$ (def. of set difference),
- [2] and then $x \in (X \cup Y)$ (def. of set sum).
- Except from that $x \notin (X \cap Y)$ (see [1], def. of set intersection)
- which gives the expected $x \in (X \cup Y)\backslash(X \cap Y)$ (see [2] and def. of set difference).

For the proof from left to right:

- [3] Let us fix arbitrary $x \in (X \cup Y)\backslash(X \cap Y)$.
- [4] From this we obtain that $x \notin (X \cap Y)$ (def. of set difference),
- [5] Additionally $x \in X$ or $x \in Y$ (see [3], def. of set sum)
- Observe that $x \notin X$ or $x \notin Y$ (see [4], def. of set intersection).
- therefore $x \in (X\backslash Y)$ or $x \in (Y\backslash X)$ (see [5], def. of set difference),
- which gives the expected $x \in (X\backslash Y) \cap (Y\backslash X)$ (def. of set intersection).

From the proved equivalence and definition of set equality we immediately obtain the goal statement.