Experiment on Ultra Strong Learning – Documentation of Data Analysis*

Ute Schmid and Christina Zeller
Cognitive Systems Group, University of Bamberg
{ute.schmid,christina.zeller}@uni-bamberg.de

January 3, 2017

Contents

*The experiment 'UltraStrong' has been conducted at December 12 and 14 2016 at University Osnabrueck.

1 Design and Research Questions

1.1 Motivation

Michie's performance criteria for machine learning:

- Weak: high accuracy
- Strong: additionally "communicate internal updates in explicit symbolic form"
- Ultra-Strong: additionally communication of updates are "operationally effective", that is the user is required to understand the update and any consequences to be drawn from it.

Our first series of experiments (ILP'16) addressed the operational effectiveness by assessing understanding of Prolog rules with respect to correct evaluations and the ability to give them meaningful names. However, while the investigated rules could have been the learned by an ILP system, the learning setting was not explicitly presented to the participants. That is, we presented the facts and the rules but not a set of positive and negative examples for the target predicate.
In the new experiment, we will extend the previous design accordingly. We do notinclude the variation of rules with/without predicate invention (reason see below, after the design has been described). In addition, we present either the full setting for learning or – as in the previous experiment – the reduced setting where only the results are given.
We only use one problem and not four as in the previous experiment. This problem is an isomorph to the well known *grandparent/2* relation which was used as second problem in the previous experiment.

1.2 Variables

Independent variable **Setting**: learning/result

- **learning**: Subjects receive first a set of facts and a set of positive and negative observations of the searched for (target) predicate. They are asked (a) to formulate the rule (dependent variable: learning-target) and (b) to solve 7 evaluation problems (dependent variable: learning-score). Afterwards they are presented with the rules learned by a hypothetical ILP system and have to solve 7 additional evaluation problems (dependent variable: result-score).
- **result**: Subjects receive first a distractor task (to have a comparable amount of mental activities for both condition) and then they are presented with the rules learned by a hypothetical ILP system and have to solve 7 evaluation problems which are identical to the second set of the Setting/learning condition (dependent variable: result-score).

The independent variable is varied between subjects, that is we have two groups. The dependent variables for score can be complemented by dependent variables for time. The dependent variable learning-target must be rated by experts for correctness.

1.3 Material

The complete procedure including all instruction, a warming-up problem and the demographic questions are given in files `q1.pdf` and `{q3a/q3b}.pdf`[1].

The general cover story is:
Imagine you work in a chemical laboratory. Over the last days you tested several substances (named aa, ab, and so on) for two reactions q1(X,Y) and q2(X,Y). For example, q1(aa,ab) means that aa is a substrate and ab is a product of reaction q1. A list of all observations is given below.

Observations:

```
q1(ab,ac). q2(aa,ac).
q1(ab,ae). q2(aa,ae).
q1(ad,ag). q2(ac,ag).
q1(ad,ai). q2(ac,ai).
q1(ae,aj). q2(af,aj).
q1(ae,al). q2(af,al).
q1(ag,an). q2(af,am).
q1(ag,ao). q2(ah,an).
q1(aj,ap). q2(ah,ao).
q1(aj,aq). q2(ak,ap).
q2(ak,aq).
```

Setting Learning:
Today you tested whether a pair of substances are related to an exothermic reaction (a chemical reaction that releases energy by light or heat). For example, exothermic(ac,an) means that ac and an are, respectively, substrate and product of a (chain of) reaction(s) which is exothermic. You observed the following test results.

Test Results:

```
exothermic(ac,an). not exothermic(aa,ab).
exothermic(aa,al). not exothermic(ad,ai).
exothermic(ab,ag). not exothermic(ab,aq).
exothermic(ae,ap). not exothermic(aj,ap).
exothermic(aa,ag). not exothermic(an,ac).
```

Setting Result:
You have a new computer program which can support you in finding rules to characterize substances. When you presented your observations to the program, it returned the following rules:

Rules:

```
exothermic(X,Y) :- q1(X,Z), q1(Z,Y).
exothermic(X,Y) :- q1(X,Z), q2(Z,Y).
exothermic(X,Y) :- q2(X,Z), q1(Z,Y).
exothermic(X,Y) :- q2(X,Z), q2(Z,Y).
```

[1] In the place of the learning part an alternative task was introduced in two variants which is an experiment concerning concept learning. This was done to keep the amount of mental effort and of time comparable between experimental groups.

1.4 Research Questions

Question 1: Do ILP approaches meet the ultra-strong criterion? In other words, are Prolog rules learned with an ILP approach operationally effective?
Test: Is the result-score significantly higher than the learning-score? Control: This result should be not dependent on the individual skills of the participants. We can assume that there will be a positive correlation between learning-score and result-score.

Question 2: Does it help to first think about a possible classification rule before the result of an ILP system is presented?
Test: Is result-score for Setting/learning significantly higher than for Setting/result?

1.5 Comments about the Design Decisions

Arguments for keeping the design simple, that is, do not include variations of the previous experiments concerning predicate invention and time of naming:

- More independent variables result in a complex design, in our case with an interaction of three variables. If such interactions are significant – and the main effects not – the results are typically hard to explain if we do not have a specific hypothesis which we have not.
- In the new experiment, the target predicate is named due to the cover story where substances were tested and found out to be exothermic or not. Therefore, naming naturally must be introduced *before* participants work on the evaluation problems.
- For the grandparent-isomorph target `exothermic`, there is no significant difference between the textual complexity (number of rules) between the representations with and without predicate invention. Based on the result of the previous experiment, we do not expect an effect with respect to solution scores and only very small effects with respect to time.
- For the open question of the setting learning, participants are in principle free to introduced new predicates when formulating the target rules. However, we assume that they will not do this (remember we assume that they are not even able to come up with a correct set of rules!).

2 Participants

The experiment 'UltraStrong' has been conducted at December 12 and 14 2016 at University Osnabrueck. Participants were 44 students of two cognitive science courses who have a good background in Prolog and in Logic but no knowledge of ILP. One participant did not finish the study and therefore, was excluded from the analysis. The experiment was conducted by Tarek Besold. Due to the date (close to winter holidays) of the originally expected more than 70 students, only a smaller number was available for testing.

- Setting learning: $n = 22$
- Setting result: $n = 21$

We controlled whether students already had participated or had heard of the previous experiment ('FamilyTree', see ILP'16, conducted in March 2016). This was the case for $n = 7$ participants (305, 315, 325, 327, 330, 331, 332) – Setting/learning: $n = 4$ and Setting/result: $n = 3$. Since their answer scores show no systematic pattern (e.g., these participants are all among the group with the highest score values), we did not see a reason to exclude them from analysis (see summary of scores per subject in the table below).
Remark: We should have introduced a final question whether participants did recognize the problem as 'grandparent'.
One participant (304) answered all questions with 'don't know'. However, he/she did not have consistently small values for the time spent per question. Therefore, we did not see a strong reason to exclude this participant – however, we could justify to exclude him/her.
One participant (294, setting:learning) has given the correct target rules. Since this participant already has found the correct rule, we cannot test for him/her whether giving the rules learned by an ILP system are operationally effective. Therefore, we decided that this participant should be excluded from statistical analyses. It showed that this participant is also the only one with full scores for all evaluation problems (learning score and test score are 7).

3 Formulating the Target Rules

The 22 participants of the setting/learning had to formulate the rules which characterize the target concept 'exothermic'. 12 participants tried to formualte the rules. Of these, 10 wrote Prolog code, 2 gave a natural language description.
Only one participant (294) gave a correct set of rules.

The results confirm our assumption that participants are **not** able to come up with the correct rules.

Answer (Code)	n = 22
no rule given (0)	10
wrong rule given	12
too general (2)	(2)
too specific (3)	(7)
wrong (4)	(2)
correct rule given (1)	1

Examples for wrong answers are given in the 'Detailed Results' section.

4 Descriptive Statistics

4.1 Table with Results per Participant

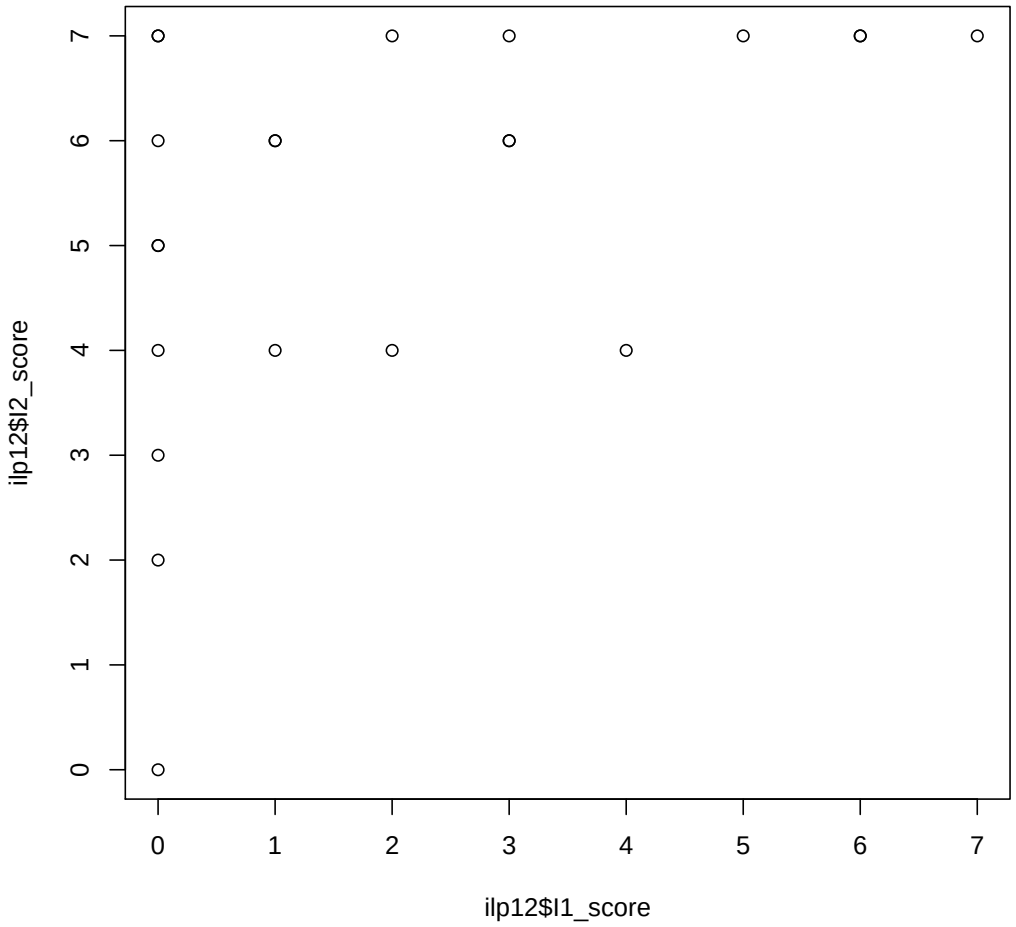
Participants Setting/learning ordered decreasing (learning-score) in the left table and result-score of Participants Setting/result in the right table. For setting/learning, *Explanation level* and *Correctness* refers to the formulation of the target rules (see Section 3).

Participants	learning-score	result-score	Explanation level	Correctness
294	7	7	1	1
325	6	7	1	2
333	6	7	2	3
327	5	7	2	3
306	4	4	1	4
328	3	7	1	3
291	3	6	1	3
305	3	6	1	4
293	2	7	3	0
309	2	4	3	0
302	1	6	1	3
323	1	6	3	0
320	1	4	1	3
301	0	7	3	0
315	0	7	3	0
314	0	6	1	3
295	0	5	1	2
324	0	5	3	0
322	0	4	3	0
298	0	3	3	0
316	0	2	3	0
297	0	0	3	0

Participants	result-score
300	7
312	7
313	7
319	7
335	7
292	6
303	6
321	6
331	6
334	6
332	5
311	4
317	4
310	3
296	2
307	2
318	2
330	2
299	1
308	1
304	0

4.2 Scatterplot for Setting/Learning

Scatterplot for the group Setting/learning with learning-score (I1_score) on the x-axis and result-score (I2_score) on the y-axis:



5 Summarized Results and Inferential Statistics

The results are given for the case where participant 294 (the only one who gave a correct set of rules) is excluded. The results for the case where this participant is included are comparable and given in the 'Detailed Results' section.

5.1 Is Result Score higher than Learning Score?

Group	Setting/learning	
	learning-score	result-score
Size n	21	
Mean (SD)	1.76 (2.07)	5.24 (1.92)

- dependent t-test with the hypothesis, that result-score is greater than learning-score: $t(21) = 7.63$, p. < 0.001
- The correlation between learning-score and result-score is r = .45, p = .038.

Question 1 can be answered positively: For setting/learning the result score is significantly higher than the learning score.

5.2 Comparison of Result Scores between Settings

Groups	Setting/learning	Setting/result
Size n	21	21
result-score Mean (SD)	5.24 (1.92)	4.33 (2.39)

- Wilcoxon rank sum test with continuity correction with the hypothesis, that result-score in the two groups differ (hopefully not significant, that is the two groups should not differ in their results):
  W = 267, p. = 0.119

Question 2 can be answered negatively: Giving the target rules is operationally effective, regardless whether participants first tried to come up with their own rules or not.

6 Final Comments

- The results clearly support our assumpation that ILP fulfills not only the strong but also the ultra-strong criterion of machine learning!
- Maybe we can do additional statistical analyses by splitting the participants into groups: It is possible to divide the Setting/result group in participants with a high and a low learning-score (e.g. learning-score> 4; Participant 294, 325, 333, 327).
- It is probably reasonable to do a follow-up experiment with a more difficult problem (e.g. greatgrandparent or Michalski train tasks).

7 Detailed Results, R Scripts

7.1 Participants' Knowledge of Previous Experiment

Did you already participate in a previous study (Spring 2016) where comprehension of Prolog programs was investigated?

n = 7
Participants: 305, 315, 325, 327, 330, 331, 332
Setting/learning: n = 4
Setting/result: n = 3

CODE:

```
> participated <- fin[fin$DE07 == "yes",]
> nrow(participated)
[1] 7
> participated$Row.names
[1] "305" "315" "325" "327" "330" "331" "332"
> p_ilp12 <- participated[participated$cilp12 == "ilp12",]
> nrow(p_ilp12)
[1] 4
> p_cilp2 <- participated[participated$cilp12 == "cilp2",]
> nrow(p_cilp2)
[1] 3
```

Did you read or hear about a previous study (Spring 2016) where comprehension of Prolog programs was investigated?

n = 5
Participants: 305, 315, 325, 327, 331
Setting/learning: n = 4
Setting/result: n = 1

```
CODE:
> hr <- fin[fin$DE08 == "yes",]
> nrow(hr)
[1] 5
> hr$Row.names
[1] "305" "315" "325" "327" "331"
> hr_ilp12 <- hr[hr$cilp12 == "ilp12",]
> nrow(hr_ilp12)
[1] 4
> hr_cilp2 <- hr[hr$cilp12 == "cilp2",]
> nrow(hr_cilp2)
[1] 1
```

7.2 All Answers 'Don't Know'

Participant 304 is the only participant who answered all 7 questions for the result-score with "don't know".
Time spent on the Example Problem: Page 06–Page12 (Question-Feedback-Question-Feedback-...)
Time spent on the Test Problem (leading to the result-score): Page 13–Page 21 (First Question on Page 15):

Participant/Page	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
304	8	46	5	22	2	37	3	16	15	24	10	4	4	4	9	5

```
CODE:
> cilp2$I2_dn
 [1] 0 0 1 0 0 7 1 0 0 0 0 0 0 0 0 0 2 0 1 0 0
> ilp12$I1_dn
 [1] 0 1 0 0 0 5 5 5 3 1 1 0 0 7 3 5 4 7 0 0 0 0
> ilp12$I2_dn
 [1] 0 0 0 0 5 2 0 0 0 0 0 0 0 2 0 2 0 0 0 0 0 0

> I2_dn7 <- cilp2[cilp2$I2_dn==7,]
```

7.3 Examples for Wrong Target Rules

Question: *Can you find general rules (systematic characterization) for exothermic reactions based on q1 and q2? These rules should cover all your observations and generalize over them. That is, they should predict for non yet tested pairs of substances whether they are related to an exothermic reaction or not. Formulate the rules as Prolog programs, in the form*

```
exothermic(X,Y) :- ...
```

If you are not able to do so, formulate them in natural language as precise as possible.
Coding:

- Explanation level—1: code; 2: textual; 3: none/no idea
- Correctness—0: none; 1: correct; 2: too general; 3: too specific; 4: wrong

Examples
Participant 327: Natural language description (Explanation level—2); too specigic – does not cover all positive examples (Correctness level—3)

```
exothermic if the substrate appears as a substrate and
the product appears as a product in the same type of q.
if they are both substrates or both products,
or if they appear like that but in different q's,
then it's not exothermic
```

Participant 295: Prolog rules (Explanation level—1); too general – does not exclude some negative examples (Correctness level—2)

```
not_exothermic(X,Y) :- q2(X,Z), q1(Y,Z).
not_exothermic(X,Y) :- q1(X,Y).
exothermic(X,Y) :- not(not_exothermic(X,Y)).
```

Participant 314: Prolog rules (Explanation level—1); too specific – does nor cover all positive examples (Correctness level—3)

```
exothermic(X,Y) :- q2(X,Z), q1(Z,Y).
```

Participant 320: Prolog rules (Explanation level—1); too specific (Correctness level—3)

```
exothermic(X, Y) : -
    not(X = an, aj, ad),
    not(Y = ac, aq, ab, ai),
    X = ac, Y = an.

exothermic(X, Y) : -
    not(X = an, aj, ad),
    not(Y = ac, aq, ab, ai),
    X = aa, Y = al, ag.

exothermic(X, Y) : -
    not(X = an, aj, ad),
    not(Y = ac, aq, ab, ai),
    X = ab, Y = ag.

exothermic(X, Y) : -
    not(X = an, aj, ad),
    not(Y = ac, aq, ab, ai),
    X = ae, Y = ap.
```

7.4 Table with Results per Participant

CODE:

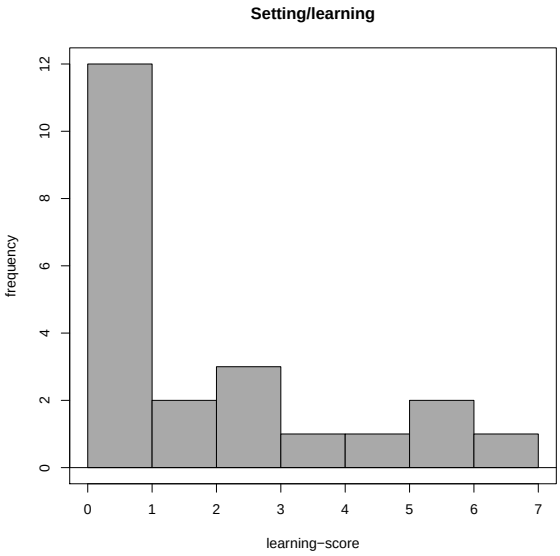
```
> df <- data.frame(ilp12$Row.names,ilp12$I1_score,ilp12$I2_score)
> df_or <- df[order(df$ilp12.I1_score,df$ilp12.I2_score,decreasing=TRUE),]
> show(df_or)

> df <- data.frame(cilp2$Row.names, cilp2$I2_score)
> df_or <- df[order(df$cilp2.I2_score,decreasing=TRUE),]
> show(df_or)
```

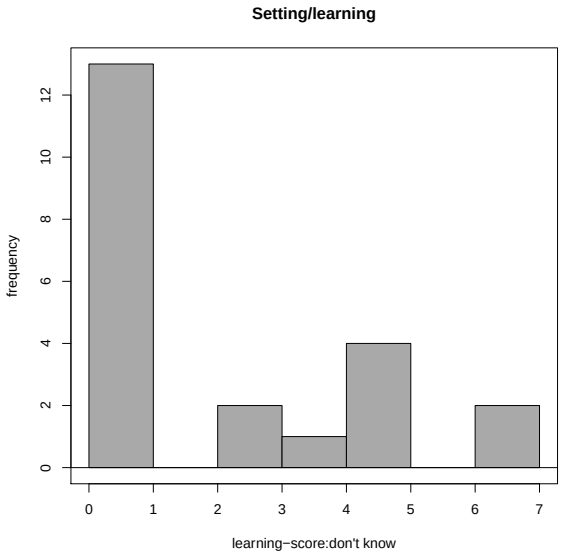
7.5 Scatterplot for Setting/Learning

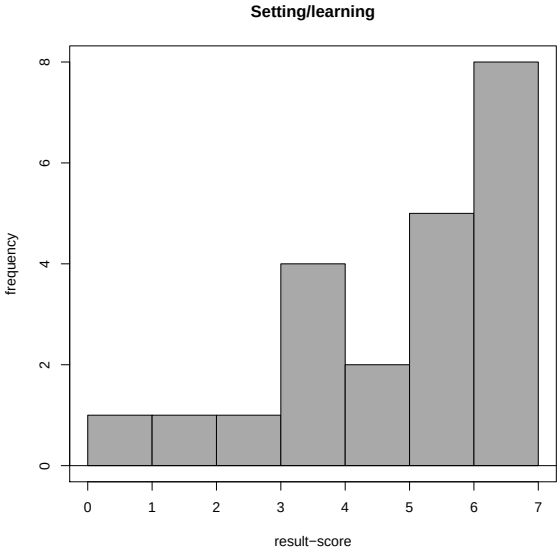
CODE:

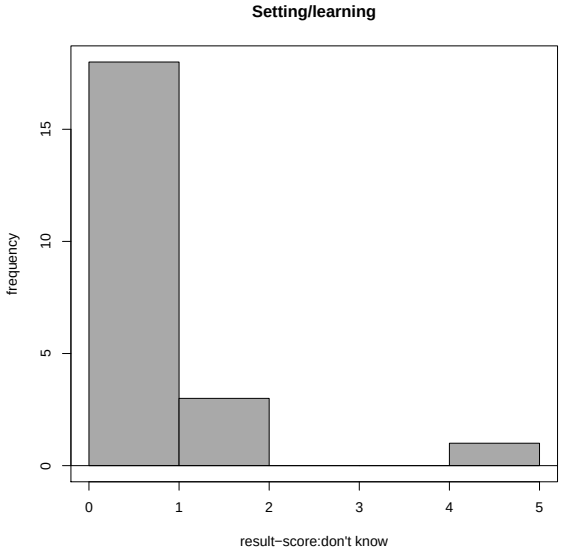
```
> pdf("I1-I2-score.pdf")
> plot(ilp12$I1_score,ilp12$I2_score)
> dev.off()
```

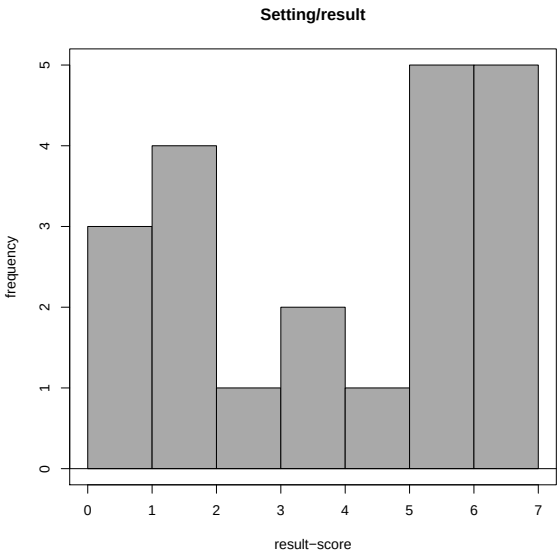
7.6 Histograms of Answer Behavior

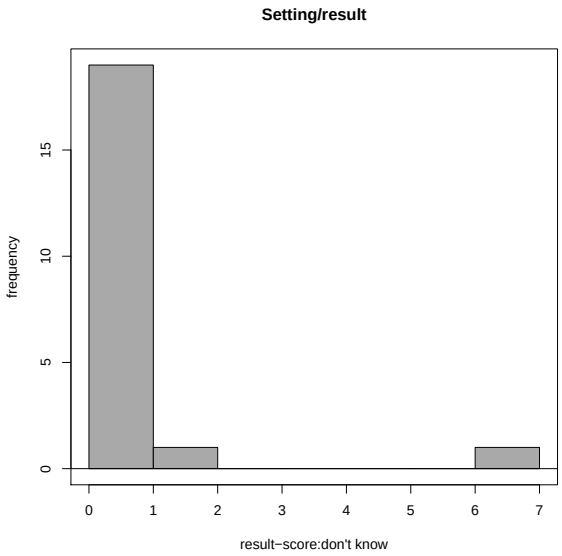












```
CODE:
pdf("cilp2-I2_score.pdf")
with(cilp2, Hist(I2_score, scale="frequency", breaks="Sturges", col="darkgray",main="Setting/result", xlab="result-score"))
dev.off()

pdf("cilp2-I2_dn.pdf")
with(cilp2, Hist(I2_dn, scale="frequency", breaks="Sturges", col="darkgray",main="Setting/result", xlab="result-score:don't know"))
dev.off()

pdf("ilp12-I2_score.pdf")
with(ilp12, Hist(I2_score, scale="frequency", breaks="Sturges", col="darkgray",main="Setting/learning", xlab="result-score"))
dev.off()

pdf("ilp12-I2_dn.pdf")
with(ilp12, Hist(I2_dn, scale="frequency", breaks="Sturges", col="darkgray",main="Setting/learning", xlab="result-score:don't know"))
dev.off()

pdf("ilp12-I1_score.pdf")
with(ilp12, Hist(I1_score, scale="frequency", breaks="Sturges", col="darkgray",main="Setting/learning", xlab="learning-score"))
dev.off()

pdf("ilp12-I1_dn.pdf")
with(ilp12, Hist(I1_dn, scale="frequency", breaks="Sturges", col="darkgray",main="Setting/learning", xlab="learning-score:don't know"))
dev.off()
```

7.7 Descriptive Statistics Summary

Descriptive statistics over all participants who finished the questionnaire:

Groups	Size	learning-score Mean (SD)	result-score Mean (SD)
Setting/learning	22	2 (2.31)	5.32 (1.91)
Setting/result	21	—	4.33 (2.39)

Descriptive statistics excluding Participant 294 (changes in **bold**):

Groups	Size	learning-score Mean (SD)	result-score Mean (SD)
Setting/learning	**21**	**1.76 (2.07)**	**5.24 (1.92)**
Setting/result	21	—	4.33 (2.39)

```
> finExc <- fin[fin$Row.names != 294,]
>
> # descriptive information
> nrow(finExc)
[1] 42
> describeBy(finExc$I1_score,finExc$cilp12)
group: cilp2
  vars  n mean sd median trimmed mad min max range skew kurtosis se
1    1 21    0  0      0       0   0   0   0     0  NaN      NaN  0
------------------------------------------------------------
group: ilp12
  vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
1    1 21 1.76 2.07      1    1.47 1.48   0   6     6 0.82    -0.74 0.45
> describeBy(finExc$I2_score,finExc$cilp12)
group: cilp2
  vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
1    1 21 4.33 2.39      5    4.47 2.97   0   7     7 -0.31    -1.53 0.52
------------------------------------------------------------
group: ilp12
  vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
1    1 21 5.24 1.92      6    5.53 1.48   0   7     7 -1.05     0.38 0.42
```

7.8 Inferential Statistics – All Participants

Question 1 Is the result-score significantly higher than the learning-score?

- appropriate test: dependent t-test

- dependent t-test requirements:
  - normal distribution of the dependent variable (or more than 30 participants in each group)
  - no negative correlation of learning-score and result-score
- test-group: Setting/result

```
> # normal distribution of the dependent variable
> # hopefully: n.s.
> # dependent variable: result-score - learning-score for the group Setting/result
> # that is: I2_score - I1_score for the group ilp12
> s1s2ilp12 <- data.frame(ilp12$I1_score, ilp12$I2_score)
> s12dif <- s1s2ilp12$ilp12.I2_score - s1s2ilp12$ilp12.I1_score

> lillie.test(s12dif)

Lilliefors (Kolmogorov-Smirnov) normality test

data:  s12dif
D = 0.1446, p-value = 0.2701

> # normal distribution of the dependent variable is given

> # no negative correlation of learning-score (I1_score) and result-score (I2_score)
> rcorr(s1s2ilp12$ilp12.I1_score, s1s2ilp12$ilp12.I2_score)
     x    y
x 1.00 0.49
y 0.49 1.00

n= 22


P
  x      y
x        0.0221
y 0.0221

> # correlation is not negative

> # dependent t-test
> t.test(s1s2ilp12$ilp12.I2_score, s1s2ilp12$ilp12.I1_score, paired = TRUE, alternative = "greater")

Paired t-test

data:  s1s2ilp12$ilp12.I2_score and s1s2ilp12$ilp12.I1_score
t = 7.1763, df = 21, p-value = 2.245e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.522547      Inf
sample estimates:
mean of the differences
               3.318182
```

Question 2 Is the result-score for Setting/learning significantly higher than for Setting/result?

- appropriate test: independent t-test
- independent t-test requirements:
  - normal distribution of the dependent variable (or more than 30 participants in each group)

  - variance homogeneity of the dependent variable

- test-group: Setting/result

```
> # normal distribution of the dependent variable
> # hopefully: n.s.
> # result-score (I2_score) of the group Setting/learning (ilp12)
> lillie.test(ilp12$I2_score)

Lilliefors (Kolmogorov-Smirnov) normality test

data:  ilp12$I2_score
D = 0.2302, p-value = 0.003646

> # result-score (I2_score) of the group Setting/result (cilp2)
> lillie.test(cilp2$I2_score)

Lilliefors (Kolmogorov-Smirnov) normality test

data:  cilp2$I2_score
D = 0.233, p-value = 0.004176

> # normal distribution of the dependent variable not given

> # variance homogenity of the dependent variable
> # hopefully: n.s.
> # result-score (I2_score) with respect to Setting/learning and Setting/result (cilp12)
> leveneTest(fin$I2_score,fin$cilp12)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  2.7605 0.1042
      41

> # variance homogenity of the dependent variable is given

> normal distribution not given therefore alternative test: wilcoxon rank sum test
> hopefully: n.s. (no differences between result-score of Setting/learning and Setting/result)
> wilcox.test(x = ilp12$I2_score, y = cilp2$I2_score, paired = FALSE, alternative="greater")

Wilcoxon rank sum test with continuity correction

data:  ilp12$I2_score and cilp2$I2_score
W = 285.5, p-value = 0.08975
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(x = ilp12$I2_score, y = cilp2$I2_score, paired = FALSE,  :
  cannot compute exact p-value with ties
```

7.9 Inferential Statistics – Without Participant 294

Question 1 Is the result-score significantly higher than the learning-score?

- appropriate test: dependent t-test
- dependent t-test requirements:
  - normal distribution of the dependent variable (or more than 30 participants in each group)
  - no negative correlation of learning-score and result-score

- test-group: Setting/result

```
> # normal distribution of the dependent variable
> # hopefully: n.s.
> # dependent variable: result-score - learning-score for the group Setting/result
> # that is: I2_score - I1_score for the group ilp12
> s1s2ilp12 <- data.frame(ilp12$I1_score, ilp12$I2_score)
> s12dif <- s1s2ilp12$ilp12.I2_score - s1s2ilp12$ilp12.I1_score

> lillie.test(s12dif)

Lilliefors (Kolmogorov-Smirnov) normality test

data:  s12dif
D = 0.1481, p-value = 0.2666

> # normal distribution of the dependent variable is given

> # no negative correlation of learning-score (I1_score) and result-score (I2_score)
> rcorr(s1s2ilp12$ilp12.I1_score, s1s2ilp12$ilp12.I2_score)
     x    y
x 1.00 0.45
y 0.45 1.00

n= 21


P
  x      y
x        0.0383
y 0.0383

> # correlation is not negative

> # dependent t-test
> t.test(s1s2ilp12$ilp12.I2_score, s1s2ilp12$ilp12.I1_score, paired = TRUE, alteater")e = "gre

Paired t-test

data:  s1s2ilp12$ilp12.I2_score and s1s2ilp12$ilp12.I1_score
t = 7.6274, df = 20, p-value = 1.208e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.690147      Inf
sample estimates:
mean of the differences
                3.47619
```

Question 2 Is the result-score for Setting/learning significantly higher than for Setting/result?

- appropriate test: independent t-test
- independent t-test requirements:
  - normal distribution of the dependent variable (or more than 30 participants in each group)
  - variance homogeneity of the dependent variable
- test-group: Setting/result

```
> # normal distribution of the dependent variable
> # hopefully: n.s.
> # result-score (I2_score) of the group Setting/learning (ilp12)
> lillie.test(ilp12$I2_score)

Lilliefors (Kolmogorov-Smirnov) normality test

data:  ilp12$I2_score
D = 0.2256, p-value = 0.006599

> # result-score (I2_score) of the group Setting/result (cilp2)
> lillie.test(cilp2$I2_score)

Lilliefors (Kolmogorov-Smirnov) normality test

data:  cilp2$I2_score
D = 0.233, p-value = 0.004176

> # normal distribution of the dependent variable not given

> # variance homogenity of the dependent variable
> # hopefully: n.s.
> # result-score (I2_score) with respect to Setting/learning and Setting/result (cilp12)
> leveneTest(finExc$I2_score,finExc$cilp12)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  2.4905 0.1224
      40

> # variance homogenity of the dependent variable is given

> normal distribution not given therefore alternative test: wilcoxon rank sum test
> hopefully: n.s. (no differences between result-score of Setting/learning and Setting/result)
> wilcox.test(x = ilp12$I2_score, y = cilp2$I2_score, paired = FALSE, alternative="greater")

Wilcoxon rank sum test with continuity correction

data:  ilp12$I2_score and cilp2$I2_score
W = 267, p-value = 0.1186
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(x = ilp12$I2_score, y = cilp2$I2_score, paired = FALSE,  :
  cannot compute exact p-value with ties
```

7.10 Summary of Inferential Statistics without Exclusion of Participant 294

Group	Setting/learning	
	learning-score	result-score
Size n	22	
Mean (SD)	2 (2.31)	5.32 (1.91)

- dependent t-test with the hypothesis, that result-score is greater than learning score: $t(21) = 7.18$, p. < 0.001

Groups	Setting/learning	Setting/result
Size n	22	21
result-score Mean (SD)	5.32 (1.91)	4.33 (2.39)

- Wilcoxon rank sum test with continuity correction[2] with the hypothesis, that result-score in the two groups differ (hopefully not significant, that is the two groups should not differ in their results):
  W = 285.5, p. = 0.08975

[2]If both 'x' and 'y' are given and 'paired' is 'FALSE', a Wilcoxon rank sum test (equivalent to the Mann-Whitney test) is carried out. In this case, the null hypothesis is that the distributions of 'x' and 'y' differ by a location shift of 'mu' and the alternative is that they differ by some other location shift (and the one-sided alternative '"greater"' is that 'x' is shifted to the right of 'y').