

Testing levels of competencies in biological experimentation



Dissertation zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Thi Thanh Hoi Phan

Kiel

2007

Referent: Prof. Dr. Horst Bayrhuber

Korreferent: Prof. Dr. Marcus Hammann

Tag der mündlichen Prüfung: 04. Juni, 2007

Zum Druck genehmigt: Kiel, den.....2007

Der Dekan

Abstract

Currently, efforts are being made to improve the quality of biology education in Germany and develop tests to assess student achievement. With the introduction of national science education standards in 2005, there has been a growing need for criterion-oriented tests, i.e. tests that assess whether or not a specific level of competency has been reached. Here we present an approach to measuring students' levels of competencies in experimentation. In particular, the focus lies on the competencies of forming hypotheses, planning experiments and analysing data. These competencies were selected because they are crucial to experimentation as problem-solving, according to the SDDS model (Klahr 2000), and because they are central to the newly introduced Biology standards.

One specific reference point of this study is the international scientific literacy test for PISA 2000. Five levels of competencies were proposed for PISA 2000. Items were mapped onto the levels of competencies by dividing up the maximum total sum score into five segments and by assigning items with a low/high difficulty to a low/high level of competency. The approach chosen in the present study differs, as closed-end test items were developed with response categories that can be directly related to a specific level of competency. Each item could be answered in different ways that are indicative of a particular level of competency. Item development, thus, took into consideration the rich research literature on qualitative differences among levels of competencies in experimentation.

Three pre-tests and a main test were taken by over 2000 students (11-12 years). The test design systematically crossed different biological subject matters (e.g., seed germination, making bread, heart beat) with the three competencies in experimentation. An independent knowledge test was also administered, measuring the students' knowledge about the science content. This design allows for analysing interactions between the competencies and the influence of the students' knowledge about the subject matter on the three competencies. Item analyses revealed reliable scales for the three competencies, for example a Cronbach's alpha of 0.78 for "forming hypotheses". Thus item development was successful as it was possible to form reliable scales with this testing approach. Correlation statistics further revealed that correlations between the three competencies were quite high for all three combinations (i.e., planning experiments * forming hypotheses, analysing data * planning experiments, forming hypotheses * analysing data). Finding these high correlations was not untypical because students with a high ability in one dimension also have high abilities in the other dimensions.

Zusammenfassung

Derzeitig wird versucht, die Qualität der Biologieausbildung in Deutschland zu verbessern und Tests zur Messung der Schülerleistung zu entwickeln. Mit der Einführung der nationalen Bildungsstandards in den Naturwissenschaften im Jahre 2004 gibt es einem zunehmenden Bedarf an kriterienorientierten Tests, d.h. Tests die einschätzen können, ob ein spezifisches Niveau der Kompetenz erreicht werden konnte oder nicht. In dieser Arbeit wird ein Ansatz zum Messen der Niveaus der Kompetenz von Schülern im Bereich Experimentieren vorgestellt. Der Fokus liegt dabei auf den Kompetenzen Hypothesen formulieren, Experimente planen und Daten analysieren. Diese Kompetenzen wurden ausgewählt, weil sie beim Experimentieren als Beispiel komplexen Problemlösens nach dem SDDS-Modell von Klahr (2000) entscheidend sind und zentral für die neu eingeführten Bildungsstandards im Fach Biologie.

Ein wichtiger Bezugspunkt der vorliegenden Studie ist der internationale wissenschaftliche Leistungsfähigkeitstest von PISA 2000. Fünf verschiedene Schwierigkeitsstufen wurden für PISA 2000 vorgeschlagen. Die vorhandenen Items wurden einem Kompetenzlevel zugeordnet, indem die maximale Gesamtpunktzahl in fünf Segmente aufgeteilt wurde und daraufhin die Aufgaben mit einer geringen bzw. hohen Schwierigkeit einem niedrigen bzw. hohen Kompetenzniveau zugeordnet wurde. Das Vorgehen in der vorliegenden Untersuchung unterscheidet sich von PISA 2000 dadurch, dass ein geschlossenes Antwortformat für die Aufgaben entwickelt wurde mit Antwortkategorien die direkt einem spezifischen Kompetenzniveau zugeordnet werden können. Jedes Item kann so beantwortet werden, dass es indikativ für ein bestimmtes Niveau der Kompetenz ist. Bei der Item-Entwicklung wurde auf die zahlreiche Forschungsliteratur zu qualitativen Unterschieden zwischen Kompetenzniveaus beim Experimentieren Bezug genommen.

Drei Vorstudien und eine Hauptstudie wurden mit insgesamt 2000 Schülern (11-12 Jahre) durchgeführt. Das Untersuchungsdesign kreuzt systematisch verschiedene biologische Fachinhalte (z.B. Samenkeimung, Brot backen, Herzschlag) mit den drei Kompetenzen des Experimentierens. Außerdem wurde den Schülern ein zusätzlicher Wissenstest vorgelegt, der das Wissen zu den einzelnen Fachinhalten erhebt. Dieses Design ermöglicht es, Interaktionen zwischen den drei Kompetenzen und dem Wissen der Schüler zu den Fachinhalten zu analysieren. Die Analyse der Items ergab reliable Skalen für die drei Kompetenzen, z.B. ein Cronbachs Alpha von .78 für Hypothesen formulieren“. Somit war die Entwicklung der Items erfolgreich und es konnten reliable Skalen für diesen Testansatz gefunden werden. Die Korrelationen zwischen den drei Kompetenzen erwiesen sich als recht hoch für alle drei Kombinationen (d.h.

Experimente planen * Hypothesen formulieren; Daten analysieren * Experimente planen; Hypothesen formulieren * Daten analysieren). Diese hohen Korrelationen waren nicht unerwartet, da Schüler mit einer hohen Fähigkeit bei einer Kompetenz auch bei den anderen beiden Dimensionen hohe Fähigkeiten besitzen.

Acknowledgment

As a Dissertation for the acquisition of a doctor degree at the Faculty of Mathematics and Natural Sciences of Christian Albrechts University in Kiel, this thesis was prepared under the supervision of Prof. Dr. Horst Bayrhuber and Prof. Dr. Marcus Hammann. The study was begun in October 2003. Since then, I have received the help of many people.

First, I gratefully thank Prof. Dr. Horst Bayrhuber and Prof. Dr. Marcus Hammann, who always supported, encouraged and assisted me during the thesis preparation.

For statistical advice, I would like to thank Prof. Dr. Claudia Nerdel, Dr. Marcus Lücken and Dr. Martin Senkbeil.

I also thank Mrs Kirsten Borski, Mrs Maria Fries for their help in the laboratory and Mrs Erika Kolaczinsky for the graphic design of the test booklets. And many thanks I want to give Mrs Daniela Hinrichsen who helped me to translate my tests into German from English.

Especially, my thanks also apply to Dr. Burkhard Schroeter, Dr. Iris Mackensen-Friedrichs, Mrs Susanne Schroeter, and many teachers and over 2000 students from many secondary schools in Germany to participating in this study

In particular, I would like to thank all co-workers of the Leibniz Institute for Science Education (IPN) at the University of Kiel, who supported me on my way to study, especially, I thank Mrs Angelika Krolow, Mrs Renate Glawe, Mrs Ulrike Gessner, and Miss Maike Ehmer.

Finally, I would like to thank my father, my mother-in-law, my husband, my son and all my friends for supporting me so I could finish this work.

Table of contents

Abstract	iv
I. Theoretical basics	1
Chapter 1: Theory	2
1. Introduction	2
2. Theory	4
2.1. Scientific Discovery as Dual Search (SDDS) model of David Klahr.....	4
2.1.1. Two main views of the process of scientific reasoning: concept formation view and problem-solving view.....	4
2.1.2. SDDS model of Klahr.....	6
2.2. The studies about the relationship between knowledge and the three processes of experimentation.....	17
2.2.1. Domain-specific pre-knowledge and hypothesis generation.....	17
2.2.2. Domain-specific pre-knowledge and experiment design.....	18
2.2.3. Domain-specific pre-knowledge and evidence evaluation.....	26
2.2.4. Domain-specific pre-knowledge and combination of hypothesis generation and evidence evaluation.....	31
2.2.5. Domain-specific pre-knowledge and combination of experiment design & evidence evaluation.....	32
2.3. Competency model in experimentation and mistakes in experimentation ...	33
2.3.1. Models for developing competencies in experimentation developed by Hammann (2004).....	33
2.3.2. Mistakes in experimentation.....	34
2.3.2.1. Student's conceptions about experimentation.....	35
2.3.2.2. Deficits when planning experiments.....	36
2.3.2.3. Deficits in the data analysis.....	37
2.3.2.4. Deficits when setting up and testing of hypotheses	40
2.4. Assessment of levels of competency of students in education science in general and in experimentation in particularly.....	43
2.4.1. Assessment of levels of students' competency in scientific literacy...	43
2.4.2. Assessment of levels of students' competency in experimentation....	44
3. Research questions	47
4. Hypotheses	48
II. Empirical part	52
Chapter 2: Cognitive Laboratory	53
1. Method	53
2. Findings	58
2.1. Time to work.....	58
2.2. Item difficulty.....	58

Table of Contents

2.3. Mean level of competency in each dimension of experimentation.....	63
2.4. Interview.....	65
3. Conclusion	66
Chapter 3: Pre-test 1	67
1. Method	67
2. Findings	73
2.1. Item difficulty.....	73
2.1.1. Method.....	73
2.1.2. Findings.....	73
2.1.2.1. Item difficulty in the knowledge test.....	73
2.1.2.2. Item difficulty in the competency test.....	75
2.2. Reliability.....	78
2.2.1. Method.....	78
2.2.2. Findings.....	78
2.2.2.1. Reliability of the knowledge test.....	78
2.2.2.2. Reliability of the competency test.....	83
2.3. Latent class analysis.....	98
2.3.1. Method.....	98
2.3.2. Findings.....	99
2.4. Correlation.....	102
2.4.1. Method.....	102
2.4.2. Findings.....	102
2.4.2.1. Correlation between pre-knowledge and the three dimensions in experimentation.....	102
2.4.2.2. Correlation between the three dimensions in experimentation.	103
3. Conclusion	105
Chapter 4: Pre-test 2	106
1. Method	106
2. Findings	107
2.1. Item difficulty.....	107
2.1.1. Method.....	107
2.1.2. Findings.....	107
2.2. Reliability.....	108
2.2.1. Method.....	108
2.2.2. Findings.....	108
2.2.2.1. Reliability at unit level.....	108
2.2.2.2. Reliability at booklet level.....	111
3. Conclusion	113
Chapter 5: Pre-test 3	114
1. Method	114
2. Findings	115
2.1. Item difficulty.....	115
2.1.1. Method.....	115
2.1.2. Findings.....	116

Table of Contents

2.1.2.1. Item difficulty in the knowledge test.....	116
2.1.2.2. Item difficulty in the competency test.....	121
2.2. Reliability.....	124
2.2.1. Method.....	124
2.2.2. Findings.....	125
2.2.2.1. Reliability of the knowledge test.....	125
2.2.2.2. Reliability of the competency test.....	129
2.3. Correlation.....	133
2.3.1. Method.....	133
2.3.2. Findings.....	134
2.3.2.1. Correlation between pre-knowledge and the three dimensions...	134
2.3.2.2. Correlation between factors and the three dimensions.....	136
2.3.2.3. Correlation between the three dimensions in experimentation...	137
3. Conclusion	138
III. Main study and Conclusion	139
Chapter 6: Results of study	140
1. Method	140
2. Results	141
2.1. Item difficulty.....	141
2.1.1. Method.....	141
2.1.2. Findings.....	141
2.1.2.1. Item difficulty in the knowledge test.....	141
2.1.2.2. Assessment students' believes about variables.....	143
2.1.2.3. Item difficulty in the competency test.....	146
2.2. Factor analysis.....	149
2.2.1. Method.....	149
2.2.2. Findings.....	149
2.2.3. Discussion of the findings of the confirmatory factor analysis.....	153
2.3. Reliability.....	153
2.3.1. Method.....	153
2.3.2. Findings.....	154
2.3.2.1. Reliability in the knowledge test.....	154
2.3.2.2. Reliability in the competency test.....	157
2.4. Latent class analysis for the competency test.....	162
2.4.1. Latent class analysis for all items in dimensions of experimentation..	162
2.4.2. The relationship between groups of students and three dimensions...	163
2.4.3. Latent class analysis for each dimension in experimentation.....	166
2.4.4. Cross tables.....	171
2.5. Correlation.....	180
2.5.1. Method.....	180
2.5.2. Findings.....	181
2.5.2.1. Correlation between pre-knowledge and three dimensions of	181
2.5.2.2. Relationship between pre-knowledge and the three dimensions..	182
2.5.2.3. Relationship between pre-knowledge and the three dimensions..	184

Table of Contents

2.5.2.4. Correlation between factors and the three dimensions.....	187
2.5.2.5. Correlation between three dimensions in experimentation	188
2.5.2.6. Correlation between the three dimensions in experimentation in three groups of students based on the knowledge test.....	190
2.5.2.7. Correlation between the three dimensions in experimentation in three groups of students based on levels of competency.....	191
3. Conclusion	198
Chapter 7: Conclusion	199
IV. Appendix	202

Table of figures

1.1	The three top-level components of the SDDS model.....	11
1.2	The two components of the process “Search Hypothesis Space”.....	12
1.3	The components of the process “Generate Frame”.....	13
1.4	The components of the process “Generate Outcome”.....	13
1.5	The components of the process “Assign Slot Values”.....	14
1.6	The four components of the process “Test Hypothesis”.....	15
1.7	The two components of the process “Evaluate Evidence”.....	16
1.8	Complete SDDS Goal Structure.....	16
2.1	Sample item: Search in the hypothesis space (Version 2).....	54
2.2	Sample item: Data analysis (Version 2).....	55
2.3	Sample item: Search in the experiment space (Version 2).....	56
2.4	Item profile in “search in the hypothesis space”-Version 1.....	59
2.5	Item profile in “data analysis”-Version 1.....	59
2.6	Item profile in “search in the experiment space”-Version 1.....	60
2.7	Item profile in “search in the hypothesis space”-Version 2.....	61
2.8	Item profile in “data analysis”-Version 2.....	62
2.9	Item profile in “search in the experiment space”-Version 2.....	62
2.10	The mean levels of competency each student gained ...(Version 1).....	64
2.11	The mean levels of competency each student gained ...(Version 2).....	65
3.1	Sample item: Search in the hypothesis space (Version 1).....	69
3.2	Sample item: Search in the hypothesis space (Version 2).....	69
3.3	Knowledge test: seed germination.....	70
3.4	Relationship between groups of students and the three dimensions in experimentation – Two-class model.....	100
3.5	Relationship between groups of students and the three dimensions in experimentation – Three-class model.....	101
6.1	Relationship between groups of students and the three dimensions in experimentation – Two-class model.....	164
6.2	Relationship between groups of students and the three dimensions in experimentation – Three-class model.....	165
6.3	Relationship between groups of students and the three dimensions in experimentation – Three-class model.....	165
6.4	Item profile for two-class model in “search in the hypothesis space”.....	167
6.5	Item profile for three-class model in “search in the hypothesis space”.....	167
6.6	Item profile for two-class model in “data analysis”.....	168
6.7	Item profile for three-class model in “data analysis”.....	168
6.8	Item profile for two-class model in “search in the experiment space”.....	169
6.9	Item profile for three-class model in “search in the experiment space”.....	170
6.10	Percentage of students who gained class 2 in one dimension also gained class 2 in other dimension – Two-class model.....	174

Table of Figures

6.11	Relationship between pre-knowledge and the three dimensions and within the three dimensions in experimentation.....	178
6.12	Correlation between pre-knowledge and the three dimensions.....	181
6.13	Relationship between pre-knowledge and the three dimensions in three groups of students basing on the knowledge test.....	183
6.14	Correlation between the three dimensions in experimentation (n = 1006).....	189
6.15	Correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions (n = 1006).....	193
6.16	Correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions (Group 1).....	194
6.17	Correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions (Group 2)	194
6.18	Correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions (Group 3).....	195
6.19	Correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions – Class 1 (n = 245).....	196
6.20	Correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions – Class 2 (n = 163).....	197
6.21	Correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions – Class 3 (n = 323).....	197

List of tables

1.1	Areas of competency in biology teaching (KMK 2004 c).....	2
1.2	Types of foci in psychological studies of scientific reasoning processes.....	17
2.1	Mean level of competency of each student in the three dimensions of experimentation (Version 1)	63
2.2	Mean level of competency of each student in the three dimensions of experimentation (Version 2).....	64
3.1	The schools and number of students participating in the test.....	68
3.2	Design of the knowledge and the competency test.....	71
3.3	Mean item difficulty for items and for units in the knowledge test.....	74
3.4	Mean item difficulty for booklets and for versions in the knowledge test.....	74
3.5	Item difficulty for items in the competency test _ Version 1.....	75
3.6	Item difficulty for items in the competency test _ Version 2.....	75
3.7	Mean item difficulty for dimensions and for units in the competency test.....	76
3.8	Mean item difficulty for booklets and for versions in the competency test....	77
3.9	Reliability of Unit 1 - In the knowledge test.....	78
3.10	Reliability of Unit 2 - In the knowledge test.....	79
3.11	Reliability of Unit 3 - In the knowledge test.....	79
3.12	Reliability of Unit 4 - In the knowledge test.....	79
3.13	Reliability of Unit 5 - In the knowledge test.....	80
3.14	Reliability of Unit 6- In the knowledge test.....	80
3.15	Reliability of Unit 7 - In the knowledge test.....	80
3.16	Reliability of Unit 8 - In the knowledge test.....	81
3.17	Reliability of Unit 9- In the knowledge test.....	81
3.18	Reliability Cronbach's alpha at unit level in the knowledge test.....	81
3.19	Reliability Cronbach's alpha at booklet level in the knowledge test.....	82
3.20	Reliability of Unit 1 - In the competency test.....	83
3.21	Reliability of Unit 2 - In the competency test.....	83
3.22	Reliability of Unit 3 - In the competency test.....	84
3.23	Reliability of Unit 4 - In the competency test.....	84
3.24	Reliability of Unit 5 - In the competency test.....	84
3.25	Reliability of Unit 6 - In the competency test.....	85
3.26	Reliability of Unit 7 - In the competency test.....	85
3.27	Reliability of Unit 8 - In the competency test.....	85
3.28	Reliability of Unit 9 - In the competency test.....	86
3.29	Reliability Cronbach's alpha at unit level in version 1_Compentency test.....	86
3.30	Reliability Cronbach's alpha at unit level in version 2_Compentency test.....	87
3.31	Reliability for three scales "forming hypothesis", "data analysis" and "planning experiment" at booklet level_ Booklet 111.....	88
3.32	Reliability for three scales_ Booklet 121.....	89
3.33	Reliability for three scales _Booklet 131.....	89
3.34	Reliability for three scales _Booklet 211.....	90

List of Tables

3.35	Reliability for three scales _Booklet 221.....	90
3.36	Reliability for three scales _Booklet 231.....	91
3.37	Reliability for three scales _Booklet 241.....	91
3.38	Cronbach’s alpha for three scales at booklet level – Item selection.....	92
3.39	Cronbach’s alpha for three scales– after using Spearman-Brown formula.....	94
3.40	Cronbach’s alpha for three scales at booklet level combined	94
3.41	Cronbach’s alpha for three scales at booklet level combined – before and after using Spearman-Brown formula.....	95
3.42	The reliability Cronbach’s alpha for three scales at booklet level combined for two scoring models.....	95
3.43	Value of the probable tested models.....	99
3.44	Mean of response probability for three latent class model.....	100
3.45	Correlation between pre-knowledge and the three dimensions in experimentation.....	102
3.46	Correlation between the three dimensions in experimentation.....	103
4.1	Mean item difficulty for items and for units.....	107
4.2	Reliability of unit 1 - In the knowledge test.....	108
4.3	Reliability of unit 2 - In the knowledge test.....	109
4.4	Reliability of unit 3 - In the knowledge test.....	109
4.5	Reliability of unit 4 - In the knowledge test.....	110
4.6	Reliability at booklet level in the knowledge test.....	111
4.7	Reliability Cronbach’s alpha at unit level in pre-test 1 and pre-test 2.....	112
5.1	Item difficulty for items and mean item difficulty for units - Knowledge test.	116
5.2	Mean scores and STD for variables in unit 1.....	118
5.3	Mean scores and STD for variables in unit 2.....	118
5.4	Mean scores and STD for variables in unit 3.....	120
5.5	Mean scores and STD for variables in unit 4.....	121
5.6	Item difficulty for items in the competency test.....	121
5.7	Mean item difficulty for units in the competency test.....	123
5.8	Mean item difficulty for dimensions and for booklet in the competency test..	123
5.9	Reliability of unit 1 - In the knowledge test.....	125
5.10	Reliability of unit 2 - In the knowledge test.....	125
5.11	Reliability of unit 3 - In the knowledge test.....	126
5.12	Reliability of unit 4 - In the knowledge test.....	126
5.13	Reliability Cronbach’s alpha at unit level for all units – Knowledge test.....	127
5.14	Reliability at booklet level - In the knowledge test.....	127
5.15	Reliability at unit level in the knowledge test in pre-test 1 and pre-test 3.....	128
5.16	Reliability Cronbach’s alpha at unit level in the competency test.....	129
5.17	Reliability at unit level in the competency test in pre-test 1 and pre-test 3.....	130
5.18	Reliability for scales “forming hypothesis”, “data analysis” and “planning experiment” at booklet level.....	130
5.19	Reliability for three scales at booklet level in pre-test 1 and pre-test 3.....	131
5.20	Reliability for three scales at booklet level combined.....	131
5.21	Cronbach’s alpha for units, for three scales and for complete booklet	132
5.22	Correlation between pre-knowledge and the three dimensions in	

List of Tables

experimentation.....	134
5.23 Correlation between variables and the three dimensions in experimentation..	136
5.24 Correlation between the three dimensions in experimentation.....	137
6.1 Item difficulty for items in the knowledge test.....	141
6.2 Mean item difficulty for units in the knowledge test.....	142
6.3 Frequency of students and mean score for each variable in unit 1.....	143
6.4 Frequency of students and mean score for each variable in unit 2.....	144
6.5 Frequency of students and mean score for each variable in unit 3.....	145
6.6 Frequency of students and mean score for each variable in unit 4.....	145
6.7 Item difficulty for items in the competency test.....	146
6.8 Mean item difficulty for units in the competency test.....	147
6.9 Mean item difficulty for dimensions and for booklet in the competency test..	147
6.10 Results of the factor analysis of the competency test.....	150
6.11 Factor analysis – two factors.....	151
6.12 Means for four randomly assigned items.....	152
6.13 Factor analysis – two factors.....	152
6.14 Reliability for Unit 1.....	154
6.15 Reliability for Unit 2.....	154
6.16 Reliability for Unit 3.....	155
6.17 Reliability for Unit 4.....	155
6.18 Cronbach’s alpha for all units.....	155
6.19 Reliability Cronbach’s alpha at booklet level in the knowledge test.....	156
6.20 Reliability Cronbach’s alpha at unit level in the competency test.....	157
6.21 Reliability at unit level in the competency test – in 3 groups of students.....	157
6.22 Reliability for scales “forming hypothesis”, “data analysis” and “planning experiment” at booklet level	158
6.23 Reliability for three scales at booklet level in three groups of students.....	159
6.24 Reliability for three scales at booklet level combined.....	160
6.25 Reliability for three scales at booklet level combined in 3 groups of students	161
6.26 Value of the probable tested models.....	162
6.27 Mean of response probability for three latent class model.....	163
6.28 Mean score for each dimension in experimentation in 2 classes of students...	163
6.29 Mean score for each dimension in experimentation in 3 classes of students...	164
6.30 Cross table between “search in the hypothesis space” and “data analysis”- Two-class model.....	171
6.31 Cross table between “data analysis” and “search in the experiment space” – Two-class model.....	172
6.32 Cross table between “search in the hypothesis space” and “search in the experiment space”- Two-class model.....	173
6.33 Cross table between “search in the hypothesis space” and “data analysis” Three-class model.....	175
6.34 Cross table between “data analysis” and “search in the experiment space” – Three-class model.....	176
6.35 Cross table between “search in the hypothesis space” and “search in the experiment space”- Three class model.....	177
6.36 Correlation between pre-knowledge and the three dimensions in	

List of Tables

experimentation.....	181
6.37 Difference between two independent correlation coefficients.....	181
6.38 Mean score and STD in each dimension of experimentation	183
6.39 Correlation between pre-knowledge and the three dimensions in experimentation for three groups of students based on the knowledge test....	183
6.40 Correlation between pre-knowledge and the three dimensions in experimentation for three classes of students based on levels of competency.	184
6.41 Difference between two independent correlation coefficients.....	185
6.42 Correlation between variables and three dimensions in experimentation.....	187
6.43 Correlation between independent and independent variables and three dimensions in experimentation.....	187
6.44 Correlation between the three dimensions in experimentation.....	188
6.45 Difference between two independent correlation coefficients.....	189
6.46 Correlation between the three dimensions in experimentation for three groups of students based on total scores in the knowledge test.....	190
6.47 Difference between two independent correlation coefficients.....	190
6.48 Correlation between the three dimensions in experimentation for three classes of students based on levels of competency in experimentation.....	191
6.49 Difference between two independent correlation coefficients.....	192

Part I

Theory

Chapter 1: Theory

1. Introduction

Currently, efforts are being made to improve the quality of education in the sciences in many countries in the world. In Germany, the establishment of education standards is considered an instrument to improve the quality of education at school. Education standards define which kind of knowledge and which competencies students shall develop during secondary level.

The basis of the development of educational standards is the concept of scientific literacy. Scientific literacy enables the individual to actively participate in societal communication and form an opinion about technical development and scientific research, and is therefore an important part of literacy. The aim of scientific literacy is to make explaining and experiencing scientific phenomena possible, to understand the language and history of the sciences, to communicate their results, use their specific methods of acquiring knowledge as well as understand their limitations (KMK 2004 a,b,c).

Education standards for biology teaching are related to four important areas of competency: subject knowledge, scientific inquiry, communication and making normative judgments. The four areas of competency are explained in table 1.1:

Subject knowledge	Biological phenomenon, concepts, principles, know facts and relate basic concepts
Scientific inquiry	Observing, comparing, experimenting, using models and applying work techniques
Communication	Establish and exchange information referring to the subject
Making normative judgments	Recognize and assess the biological situation in various contexts

Table 1.1: Areas of competency in biology teaching (KMK 2004 c)

Scientific inquiry is central to science learning. When engaging in inquiry, students describe objects and events, ask questions, construct explanations, test those explanations against current scientific knowledge, and communicate their ideas to others. They identify their assumptions, use critical and logical thinking, and consider alternative explanations. In this way, students actively develop their understanding of science by combining scientific knowledge with reasoning and thinking skills.

In biology, experimentation as well as criteria-related observations and comparisons are specific forms of scientific knowledge acquisition. They are all characterized by formulating a question and setting up hypotheses, planning and executing an experiment, an observation or a comparison, and assessing the acquired data and their interpretation with reference to the hypotheses.

This work is related to experimentation as one form of the competency of knowledge acquisition. Our study presents an approach to measuring different levels of the competencies of biology students aged 11-12 in experimentation.

2. Theory

2.1. Scientific Discovery as Dual Search (SDDS) model developed by David Klahr

2.1.1. Two main views of the process of scientific reasoning: concept formation view and problem solving view

As David Klahr pointed out there are two main ways of looking at scientific reasoning: The first view is concept formation and the second is problem solving.

Concept formation

Concept formation or concept learning is used to refer to the development of the ability to respond to common features of categories of objects or events. Concept formation allows students to group information by connections and seeing the relationships between items of information. By linking the instances, events or subjects and by explaining their reasoning, students can form their own understanding of the concept. In learning concepts, one must focus on the relevant features and ignore those that are irrelevant (Bourne et al., 1986). Concept formation tends to be the method used by subjects in the laboratory, it might be not appropriate in everyday life.

The hypothesis testing theory was proposed by Bruner, Goodnow and Austin (1956) in a series of investigations on concept learning. In this investigation, they gave participants the available instances, and their study focused on how subjects select instances from a predefined set in order to evaluate hypotheses and how they generate new hypotheses based on the feedback about those instances. Bruner et al. (1956) discovered that subjects use several strategies for gathering information about hypotheses even in a relatively simple context, and they argued that the concept learning task is relevant to real science because it involves two essential components of scientific reasoning: the logic of experimentation and strategies for discovering regularities.

This task is usually used in the laboratory study of scientific reasoning.

Problem solving

Considered the most complex of all intellectual functions, problem solving has been defined as a higher-order cognitive process that requires the modulation and control of more routine or fundamental cognitive skills (McCarthy & Worthington, 1990). It occurs if an organism or an artificial intelligence system does not know how to proceed from a given state to a desired goal state. It is part of the larger problem process that includes problem finding and problem shaping (Simon 1981, 1999). The production of

scientific knowledge was considered by Simon (1981) as solving complex problems, and problem solving was attributed in two problem spaces: Search in the hypothesis space and search in the experiment space. According to Simon (1999), *Problem solving* designates a finding and a moving from an initial state to a goal state. Contrary to a task, the way from the initial state to the goal condition is not known. *Problem solving heuristics* serve searching and finding a way from the beginning to the goal condition. One of the most well-known problem solving heuristics is the *central goal analysis*. Here the current condition and the goal condition are compared and the differences determined. Subsequently, an *operator* is sought, in order to reduce the differences between the current condition and the goal condition. *Operators* are actions which transform a condition into another, for example a starting situation into an intermediate condition. The term *problem space* is defined as the representation of all possible problem states (initial state to goal state) which are produced when all applicable operators are applied.

Higher order cognitive processes are used to understand important aspects of the problem so that an answer or solution can be found. This view is supported by the “2-4-6” rule discovery task invented by Wason (1960). The experimenter tells the participants of the study to discover the rule behind the numerical triads “2-4-6. The subjects propose hypotheses about the rule and give instances to test their hypotheses in order to discover the rule, while the experimenter provides yes/no feedback and thus tells the subjects whether or not their proposed hypotheses are correct. The main outcome of this study is that subjects tend to propose a single hypothesis and seek evidence to confirm – rather than disconfirm – it. The tendency to confirm one’s own hypotheses – the so-called confirmation bias – is very prevalent and has also been found by other researchers (e.g., Mynatt, Doherty & Tweney, 1977, 1978). Chin & Brewer (1998) described eight possible strategies students use concerning data that contradict their own expectations: ignoring the data, rejecting the data, professing uncertainty about the validity of the data, excluding the data from the domain of the current theory, holding the data in abeyance, reinterpreting the data, accepting the data and making peripheral changes to the current theory, and accepting the data and changing the theories.

2.1.2. SDDS model of Klahr

Klahr & Dunbar (1988) introduced two forms of scientific reasoning, one dealing with the two phases of the discovery process (i.e., hypothesis formation and experiment design), and the other with two frameworks for understanding the psychology of these processes (the concept learning view and the problem solving view). Klahr and Dunbar proposed an integrated view to replace both of these forms. “The key to this integration comes from Simon and Lea’s (1974) insight that both concept learning and problem solving are information-gathering tasks and that both employ guided search process.” (Klahr & Dunbar, 1988, p. 5).

The integrated view by Simon and Lea

Simon and Lea (1974) are considered the first people who gave the integrated view. They indicated that both concept learning and problem solving are information-gathering tasks and that both employ guided search processes. They built one model of the processes that involved both concept learning and problem solving - called Generalized Rule Inducer (GRI). In GRI, for both rule induction (concept learning) and problem solving the same general methods are used. However, rule induction requires the search in two problem spaces: a space of instances and a space of rules. In the space of rule, students have to search for hypotheses about the rule and in a space of instances; they have to search for a good instance. However, problem solving search takes place in only a single space: a space of rules.

In rule induction tasks, the proposed rules are never tested directly, but only by applying them to instances. The subject selects an instance and checks to see whether the instance confirms or disconfirms the rule.

The GRI view characterizes some further differences between the previous research on concept formation and problem solving. The concept formation is concerned with rules which are drawn simply from well-defined instances and the full set of permissible instances is predetermined. In problem solving, the task is more complicated, it consists of a series of knowledge states that students can generate. GRI can be considered the departure point for analysis of scientific reasoning. However, according to Klahr and Dunbar (1988), two extensions are required if they are to an effect this proposed integration of the concept learning and problem solving views of scientific reasoning.

Two extensions by Klahr to GRI

The first extension was to study the subjects’ behavior in situations that more closely resemble the scientists’ environment than the traditional laboratory tasks. The second

extension was to extend the GRI to accommodate the added complexity of the new situation (Klahr & Dunbar, 1988).

Klahr et al. (2000) devised a task with a more complicated rule space than that used in most concept formation experiments, and studied the behavior of subjects who attempted to extend their understanding about the device and discover how a new function operates.

Klahr's empirical studies

In order to do this, Klahr and his colleagues used a computer-controlled robot tank (called "Big Trak") that enabled them to track the subjects' behavior through the entire cyclical sequence of stages that comprise the discovery process. There are many function keys and special keys.

First, the user uses the CLR key ("clear the memory", a "special key") to clear the memory, then enters a series of up to 16 instructions ("function keys") and finally enters the Go key. Big Trak executes the program by moving around the floor.

Study 1: Discovering a new function. Twenty adult subjects participated in this study (Dunbar & Klahr, 1989) which included two phases. First, the subjects were informed about how to use all the function keys and special keys, except for the repeat key (RPT). Second, the subjects were told to discover how the RPT key works by proposing hypotheses and testing them.

In the discovery phase, the subjects were asked to speak aloud, to say what they were thinking and which key they were pressing. The subjects' behavior during the phase was videotaped. The subjects had to propose the hypothesis about how RPT worked before designing and running any program. When subjects claimed that they were certain that they had discovered how the RPT key worked or when 45 minutes were finished, the discovery phase was terminated. Statements about how the subject thought the RPT key might work were coded as hypotheses. Statements which the subjects gave of what might happen once the GO key was pressed were coded as predictions. Comments about the behavior of the device once the program had been executed were coded as observations. In this process, they contributed the major processes of scientific reasoning: forming hypotheses, designing experiments and reacting to experimental outcomes.

The results showed that all subjects were successful at the task and found the correct rule for RPT, but that their behavior diverged widely from any normative model of scientific reasoning. All subjects started with the same strategies; they used an initial

hypothesis to guide the search in the experiment space, but then they diverged in the way they searched for new hypotheses once the initial hypotheses were abandoned: One group searched the hypotheses space for a new hypothesis by searching memory for information rather than by further experimentation (the so-called “Theorists”), and the other explored the experiment space to see if they could find some regularities in the experimental outcomes (the so-called “Experimenters”). Some of them conducted many experiments without explicit hypotheses; the others proposed new hypotheses by abstracting from the result of a prior experiment. The findings also indicated that the Theorists took less time than the Experimenters to discover how the RPT key works. The main difference between the two groups, however, is that the Experimenters conducted significantly more experiments than the Theorists, and that their experiments were conducted without explicitly stated hypotheses. In sum, the authors indicated that the important strategy difference between Theorists and Experimenters was that “the Theorists searched the hypothesis space for a new hypothesis, but Experimenters explored the experiment space without an active hypothesis” (Klahr 2000, p. 69). Therefore, Klahr et al. (2000) performed a second study to test the hypothesis that it is possible to think of a correct hypothesis without exploration of the experiment space.

Study 2: Forced search in the hypothesis space. Ten undergraduates participated in the experiment for course credit (Dunbar & Klahr, 1989).

The goal of this study was to investigate whether or not the generation of multiple – rather than single – hypotheses prior to experimentation would change the ways subjects generated and evaluated experiments. Study 2 consisted of two phases. The first phase was the hypothesis-space search phase, when subjects were asked to state multiple ways that RPT might work before doing any experiments. In the second phase, subjects were allowed to conduct experiments as it was the case in study 1. The authors hypothesized that it is possible to generate the correct rule through the hypothesis-space search without any experimentation at all. Furthermore, the authors tested the hypothesis that subjects switch to the experiment space in order to generate new hypotheses if the initially formed hypotheses fail to explain the phenomenon. Thus, Theorists were expected to behave like Experimenters in this study if their hypothesis-space search failed.

The results of phase 1 in study 2 indicated that it is possible for subjects to generate the correct hypothesis without conducting any experiments. This result is consistent with the view that the Theorists in study 1 think of the correct rule by a searching the hypothesis space. In the experimental phase, all subjects who failed to generate the

correct rule in the hypothesis-space phase behaved like the Experimenters in Study 1. This is consistent with the view that when the hypothesis-space search fails, subjects must turn to a search of the experiment space.

The main difference between the results of Study 1 and Study 2 is that the subjects conducted far fewer experiments in Study 2. Furthermore, the subjects in Study 2 conducted experiments that allowed them to distinguish between two hypotheses, whereas the subjects in Study 1 rarely designed hypothesis-discriminating experiments, for they were usually dealing with only one hypothesis at a time.

The influence of prior knowledge is suggested by the finding that all of Theorists, but only one of the Experimenters, had prior programming experience. The authors also found it interesting that the effect of differential prior knowledge effected not only through the initial hypothesis-formulation stage but also the experimental strategies.

In sum, prior exploration of the hypothesis space had two main effects on the experimental phase. First, it allowed subjects to generate an initial set of hypotheses. Second, because subjects were aware that a number of hypotheses could account for their results, they conducted discriminating experiments, and they quickly discovered the correct rule.

So, by means of the results of these two studies, Klahr and Dunbar arrived at three conclusions. First, when people attempt to discover something in a moderately complex environment, they do indeed search in two problem spaces: the hypothesis space contains hypotheses about the aspects of the environment under investigation, and the experiment space contains possible configurations of the environment that might product informative evidence related to those hypotheses. Second, people can use two different strategies for searching these spaces that can be distinguished by which of the two spaces are searched preferentially. Third, during the phase of evidence evaluation when active hypotheses are being reconciled with the empirical results, people often violate the normative canons of “rational” scientific reasoning.

Study 3: Dual search by children. Twenty-two third to sixth graders from a private school participated in the study (Klahr & Dunbar 1989, Klahr et al. 1993).

In study 1 and study 2, the subjects were adults. Thus, in this study with children two sets of questions dealt with searching the hypothesis space and searching the experiment space. Three questions concerned the search in the hypothesis space: 1) Given the same training experience as adults, will children think of the same initial hypotheses as adults? 2) When children’s initial hypotheses are disconfirmed, will the processes that

are used to search the hypothesis space similar in both children and adults? 3) Will children be able to change frames or will they remain in the same frame?

Two questions concerned children's search of the experiment space: 1) Will children search in different areas of experiment space than the adults do? 2) Will children evaluate the results of experiments in a different way than adults?

As in studies 1 and 2, the subjects were taught how to use BigTrak and were then asked to discover how the RPT key works. Here, they used the BigTrak Dragon instead of BigTrak Tank in Studies 1 and 2 (BigTrak Dragon had only a change in instructions was which the "Fire" button now corresponded to the dragon's "breathing fire", instead of firing its "laser cannon").

The results indicated that only two of the twenty-two children correctly concluded how the RPT key works. The children proposed 3.3 different hypotheses during the course of session; however, nearly all of them were incorrect.

There were differences between children and adults in the proportion of experiments performed related to different types of hypotheses. Nearly 10% of the adults' experiments were run under a selector hypothesis (one repeat of last or first n step), but only 1% of children's hypotheses. Moreover, 30% of the children's experiments were conducted under partial hypotheses, whereas adults proposed fully specific hypotheses for all but three percent of their experiments ("common hypothesis" was defined as a fully-specified hypothesis that was proposed by at least two different participants. "Partial hypotheses" were defined as those in which only some of attributes of the common hypotheses were stated by the participant.).

All of the twenty children who failed to discover how PRT works proposed hypotheses that were solely in the n -role: counter frame (hypotheses are classified according to the role that they assign to the parameter (n) that goes with the PRT command. If n counts the number of times that something (program, first step, last step, or subsequent steps) is repeated, these hypotheses are called "counter". If n determines which segment (first or last n steps, or the n^{th} step) of the program will be selected to be repeated again, the hypotheses are called "selector"). These suggested two reasons: First, the children did not have sufficient knowledge available to generate the n -role: selector frame, by searching the hypothesis space. Second, the children did not use the results from their experiment-space search to induce a new frame as adults did. Instead, they used it to induce new slot values for their current frame.

In sum, this study revealed three main differences between adults and children. First, the children proposed hypotheses that were different from those proposed by adults. Second, the children did not abandon their current frame of the search in the hypothesis

space for a new frame, nor did they use the results of the experiment-space search to induce a new frame. Third, the children did not attempt to check whether their hypotheses were consistent with prior data.

So, according to the point made by Simon and Lea (1974) about scientific reasoning the Generalized Rule Inducer (GRI) was used in traditional laboratory studies of problem solving and rule induction. Klahr et al. proposed two extensions to GRI in order to apply the concept of dual-space search, a space of experiments and a space of hypotheses.

The SDDS model

The Scientific Discovery as Dual Search (SDDS) model (Klahr, 2000) was “proposed as a general model of scientific reasoning that can be applied to any context in which hypotheses are proposed and data is collected”. In this view, scientific reasoning processes involve the search in two spaces: “the hypothesis space, consisting of the hypotheses generated during the discovery process, and the experiment space, consisting of all possible experiments that could be conducted. Search in the hypothesis is guided both by prior knowledge and by experiment results. Search in the experiment space may be guided by the current hypothesis and it may be used to generate information to formulate hypotheses” (Klahr & Dunbar, 1988, p. 32).

SDDS involves the three main components that control the entire process, from the initial formulation of hypotheses, through their experimental evaluation, to the decision that there is sufficient evidence to accept a hypothesis. These components are: Search hypothesis space, test hypothesis and evaluate evidence.

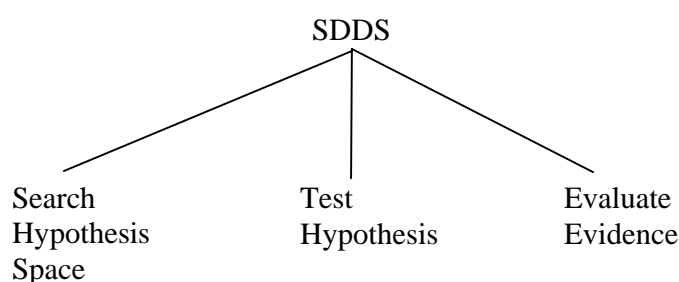


Figure 1.1: The three top-level components of the SDDS model

The SDDS model can be shortly summarized as follows:

- The output of “search hypothesis space” is a fully specified hypothesis which provides the input to “test hypothesis”.

- “Test hypothesis” generates an experiment appropriate to the current hypothesis (E-space move), makes a prediction, and observes the outcome. The output of “Test hypothesis” is a description of evidence for or against the current hypothesis, based on the match of the prediction derived from the current hypothesis and the actual experimental result.
- “Evaluate evidence” decides whether the cumulative evidence - as well as other considerations - warrants acceptance, rejection or continued consideration of the current hypothesis.

Search Hypothesis Space

The search of the hypothesis space is the process of generating new hypotheses. A new hypothesis can be generated by searching in memory, or by revising the outcomes of the experiments. This process includes two components: “Generate Frame” and “Assign Slot Value”. The first component generates a new hypothesis in broad scope, and the second component refines the hypothesis.

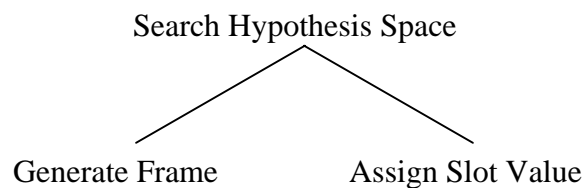


Figure 1.2: The two components of the process “Search Hypothesis Space”

Generate frame: This process has two components corresponding to the way that a frame is generated: “Evoke frame” and “Induce frame”.

- “Evoke frame” is a search of memory for information that could be used to construct a frame. In this way, the prior knowledge plays an important role as the hypothesis can be generated through analogical mapping, priming, reminding, conceptual combination (Gentner, 1983), and heuristic search (Kaplan and Simon, 1990, Klahr & Dunbar, 1988). That means that in this process subjects are able to recall similar situations and use them as the basis for constructing initial frames.
- “Induce frame” is a process of generating a new hypothesis by induction from the outcomes of the experiments. Subjects have to observe some results before producing a frame. Two subcomponents of this process are to generate outcomes and to generalize outcomes to produce a frame.

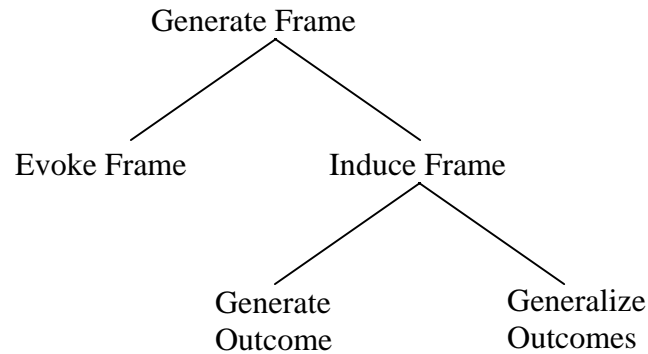


Figure 1.3: The components of the process “Generate Frame”

Generate Outcome

This process contains three sub-processes: Search E-Space, Run experiments and Observe the result.

“Search E-Space” includes two components: “Focus” as well as “Choose & Set.” The most important step is “Focus” on some aspects of the current situation that the experiment is intended to illuminate. If there is a hypothesis, “Focus” determines some aspects of the main reasons for the experiment. If there is a frame with open slot values, then “Focus” will select the most important one of those slots to be solved. If there is neither a frame nor a hypothesis; “Focus” makes a decision about the important aspect of the current situation to focus on. “Choose” sets an evaluation value in the experiment space that will provide information relevant to it, and “Set” determines the evaluation of the remaining, but less important, experimental features necessary to produce a complete experiment.

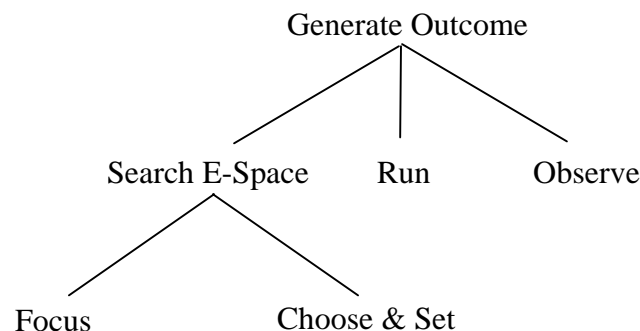


Figure 1.4: The components of the process “Generate Outcome”

The results of “Generate outcomes” are inputs of “Generalize outcomes”. By means of this sub-process a frame is produced.

“Assign slot value” is the process to take a partially instantiated frame and assign specific values to the slots so that a fully specified hypothesis can be generated. It also involves two components like generate frame process: “Use prior knowledge” and “Use experiment outcomes.”

- “Use prior knowledge”: Slot value can be assigned by using prior knowledge. This is usually used in the early phases of the discovery process in order to generate specific slot values.
- “Use experiment outcomes”: The experiment outcomes can be already existent (old outcomes) or they might be produced. If there are already some experimental outcomes, slot value can be determined specifically by using these outcomes (On the other hand, if there are no variable outcomes, or the old outcomes are not appropriate, the system can produce new outcomes by using “Generate Outcome” to assign slot values. The “Generate outcome” process here is as the same as was mentioned in “Generate frame”.

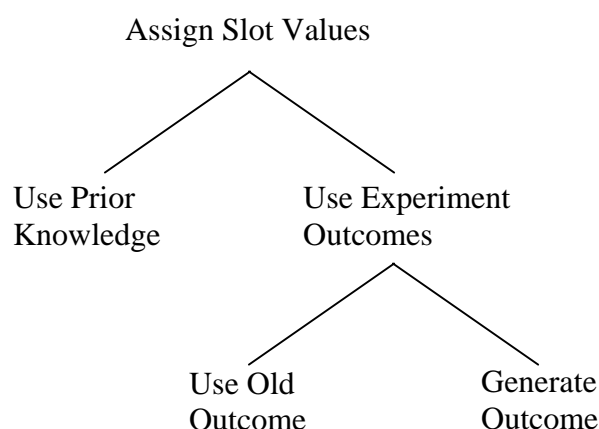


Figure 1.5: The components of the process “Assign Slot Values”

Test hypothesis

A hypothesis refers to a provisional idea which merits evaluation. A hypothesis requires more work by the researcher to either confirm or disprove it. A confirmed hypothesis may become part of a theory or occasionally may grow to become a theory itself. The hypothesis will be evaluated by “Test hypothesis”.

In this process, hypotheses are tested through experiments. In the scientific method, an experiment is a set of actions and observations, performed in the context of solving a particular problem or question, to test a hypothesis and answer a research question concerning a phenomenon. The experiment is a cornerstone in the empirical approach to acquiring knowledge about the world.

This process comprises three sub-processes: searching E-space or formulating an experiment, making a prediction, as well as running the experiment, observing the result, and matching the actual outcomes to expectations.

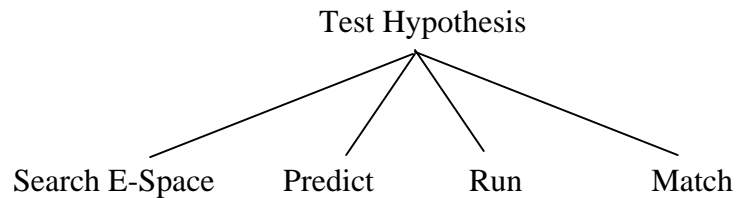


Figure 1.6: The four components of the process “Test Hypothesis”

“Search E-Space” or formulate an experiment: In this process, an experiment is designed. “Make a prediction”: A prediction is a statement or claim that a particular event will occur in the future. In “Test hypothesis”, subjects use the current hypothesis and the experiment designed in order to predict specific results that will occur.

“Run” the experiment, observe the result means that subjects carry out the experiment, manipulate and control the equipment or tools if necessary. Then they observe what happens, note it down on the sheet of papers or make photos. Finally, “Match” the result means that experimental findings are compare to the expectations.

Evaluate evidence

“Evaluate evidence” consists of assessing the fit between the current hypothesis and evidence as well as guiding further search in both the hypothesis space and the experiment space. This process determines whether or not the cumulative evidence about the experiments performed under the current hypothesis is sufficient or if it is to be rejected or accepted. This process consists of two components: “Review outcome” and “Decide” if the current hypothesis can be accepted, rejected or continued in the next experiment: These processes determine whether or not the cumulative evidence about the experiments run under the current hypothesis is sufficient to reject or accept it, or the experiments need to be continued.

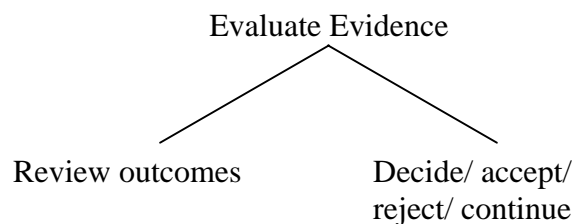


Figure 1.7: The two components of the process “Evaluate Evidence”

Klahr has assembled all components of the three processes in Scientific Discovery as Dual Search.

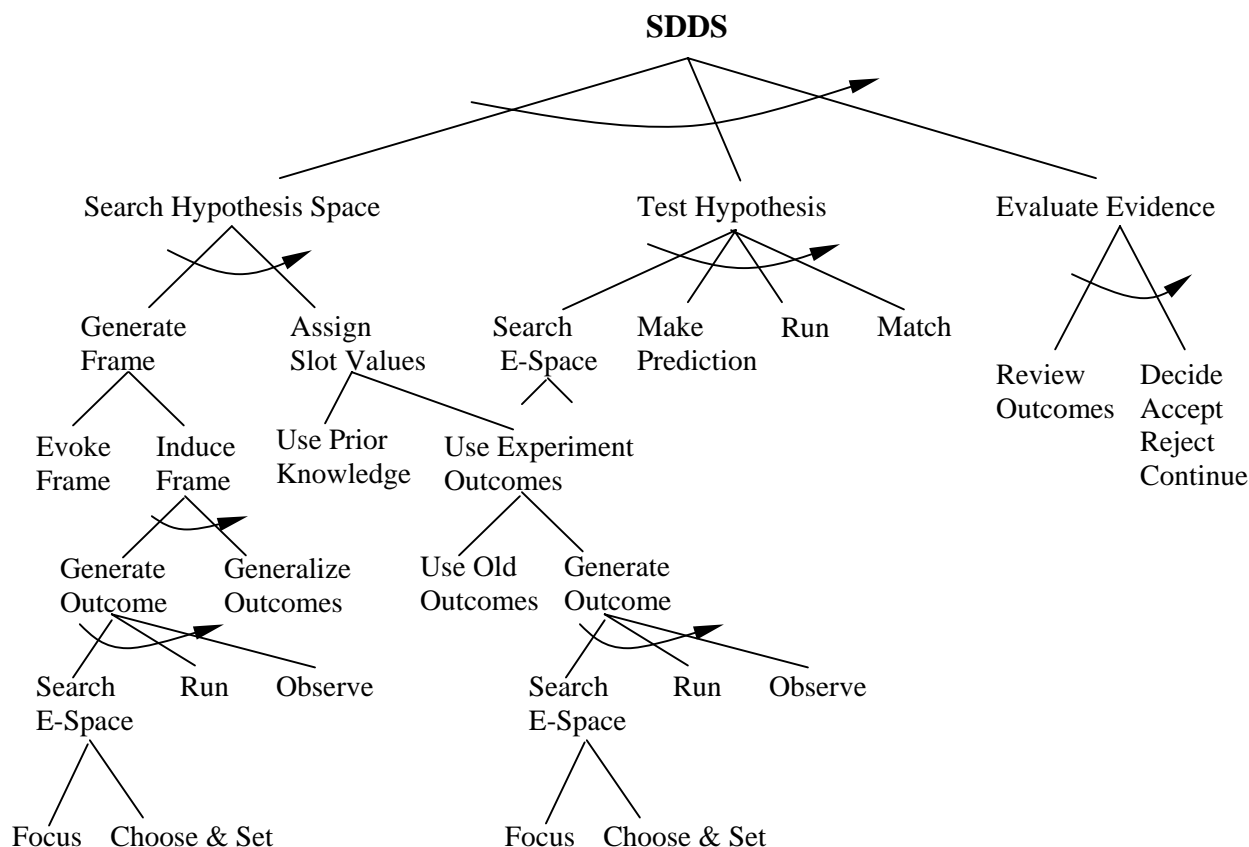


Figure 1.8: Complete SDDS Goal Structure

In summary, as Klahr described in the SDDS model, the complex process of scientific discovery can be considered a system of sub-processes, in which each sub-process plays a role in the system, and they all inter-reacted with each other. One sub-process can be an input to another.

Experimentation as one form of scientific discovery is an important opportunity for fostering student reasoning in the science classroom. Experimentation also affords a rich context for studying the process of theory change, central to learning in general (Schauble, Glaser, 1990; Schauble et al., 1991).

Many researchers studied the correlation between prior-knowledge and the three processes of experimentation. In the next chapter, some particular studies in this area are described.

2.2. The studies about relationship between knowledge and the three processes of experimentation

Many studies indicated that the three processes of experimentation are influenced by the pre-knowledge of subjects (Chi, Feltovich, & Glaser, 1981; Lord, Ross & Lepper, 1979; Kuhn, Amsel & O'Loughlin, 1988; Schauble, 1990; Carey, 1985; Dunbar, Klahr, 1989). The relationships between the three sub-processes in experimentation and the domain-specific and domain-general knowledge were systematically characterized by Klahr (2000) as indicated in the following table.

	Hypothesis Generation	Experiment Design	Evidence Evaluation
Domain-specific knowledge	A	B	C
Domain-general knowledge	D	E	F

Table 1.2: Types of foci in psychological studies of scientific reasoning processes (Klahr, 2000, p. 13)

Klahr distinguishes between domain-specific knowledge and domain-general knowledge, and proposes six cells for investigating the relationship between knowledge and the three processes of experimentation.

In the following such studies are referred to in which a correlation between domain specific pre-knowledge and the three processes of experimentation is analysed.

2.2.1. Domain-specific pre-knowledge and hypothesis generation

Mike McClosky (1983) investigated the subjects' understanding about motions in physics. The participants were asked to depict the motion of an object that is dropped from an airplane, or the motion of a ball when it exits a curved tube. He found that many college students believed that the curved tube imparts the ball a with curvilinear impetus, and when the ball exits the tube, it continues in a curved trajectory that eventually straightens out. In this study, the participants did not perform any experiments, but they had to use their specific knowledge concerning the area and form hypotheses about the motion of objects in order to answer the investigator's questions.

Carey (1985) studied children's conceptions of animals and other living things and argued that children's biological conceptions are radically reorganized between ages 4 and 10. For example, 10-year-old children represent relations among the processes of living as eating, breathing, growing, dying, and having babies that the 4-year-old children do not.

She also compared the activities of children and adult scientists when they search the hypothesis space. She was able to confirm that - among others - the development of related conceptual systems (physical object, living thing, animal, plant and person) and classes of phenomena are a pre-condition of formulating and evaluation hypotheses (Carey, 1985a). Carey showed that the most important factor differentiating the formulation and evaluation of the hypotheses of children and adults is the adequacy of domain-specific knowledge.

2.2.2. Domain-specific pre-knowledge and experiment design

In some studies, subjects were asked to design experiments or choose the correct experiment in available experimental sets without hypothesis generation and evidence evaluation. Tschirgi (1980) investigated the influence of the subjects' pre-knowledge on experiment design. She used eight different multivariable stories about common everyday situations. Each story had a different outcome. In the study, items were used, in which a hypothesis was stated. The participants were then asked to choose an experiment, with which it is possible to test the hypothesis. Thus, Tschirgi was able to investigate the participants' skill to test hypotheses without searching in the hypothesis space. For example, in the story of baking a cake, a good cake was made by using three ingredients (margarine, brown whole-wheat flour, and honey). The hypothesis was that it was the honey that was responsible for the good cake. She gave the participants three different experiments to test this hypothesis. The first experiment was called "vary-one-thing-at-a-time" (VOTAT), in which only the test variable was changed and the others were kept constant (sugar was used instead of honey, and margarine and brown whole-wheat flour). The second experiment was called "hold-one-thing-at-a-time" (HOTAT), in which the test variable was kept constant, but other variables were changed (the cake was baked again, still using honey but butter was used instead of margarine and white flour instead of whole-wheat flour). The third experiment was called "change-all" (CA) in which all elements were changed (the cake was be made again, but this time sugar, butter and white flour were used).

In order to choose the correct answer, the participants had to be able to design an unconfounded experiment. Tschirgi found that the domain-specific knowledge influenced the ability of people to design an experiment. In particular, she found that the participants chose unconfounded experiments if the initial outcome was good because the participants tended to keep the hypothesized cause of the good outcome. They tended to form confounded experiments, however, if the initial outcome was bad because the participants tended to change the hypothesized cause of the bad outcome.

Sodian et al. (1991) presented two studies to test whether young elementary school children are, in fact, fundamentally unable to distinguish between hypotheses and evidence. They asked two questions: (1) Do children have the notion of testing a hypothesis as opposed to producing an effect? And (2) Given a choice between conflicting hypotheses, can they distinguish between experiments that would produce conclusive as opposed to inconclusive evidence?

Study 1 explored children's ability to spontaneously generate, and/or to choose from two alternatives, an adequate test for a genuine scientific hypothesis.

Study 2 is designed to assess the children's ability to distinguish between testing a hypothesis and producing an effect. The children were asked to choose an adequate test for a hypothesis required to distinguish between a conclusive and a inconclusive test.

Study 1: A story about two brothers who know that there is a mouse in their house, though they had never observed it because the mouse is only active at night. They want to know if the mouse is big or small. There are two boxes, one with a big opening, and the other with a small one. Which box should they choose to put food in, in order to be sure about the size of the mouse? Which box should they choose in order to be certain that the mouse will get the food?

All children answered the control questions correctly, indicating their understanding that both a big and a small mouse could get into the house with the large opening, whereas only a small mouse could get into the house with the small opening.

Study 2: Tom and Mike got a new pet animal, an armadillo (*Dasypus novboracensis*; "antbear"). They want to know how well this animal can smell. Tom thinks that an armadillo has a very sensitive nose, but Mike thinks that the armadillo does not smell very well. They want to find out who is right. What could they do to find this out?

They plan to put a piece of food in the animal's box and cover it up with sand so that the armadillo can not see it, and they want to see whether the armadillo finds it or not. They have two different kinds of food, one has a very strong smell and the other has very weak smell. Children were asked to predict the outcome of the experiment for each of the two pieces of food: "To find out whether the armadillo has a good nose or a bad nose, which one of two pieces of food should they hide? Why?"

This study indicated that young elementary school children were able to distinguish between testing hypothesis and producing a positive effect. Furthermore, they were able

to distinguish between a conclusive and an inconclusive test of simple hypotheses and understood the inferences that could be made by the outcome of a conclusive test.

In this study, children were asked to choose a test to decide between two alternative hypotheses, they did not have to generate an alternative and seek for evidence that would support this alternative hypothesis.

Brown (1990) argued that one can not study learning and transfer in a vacuum and that children's ability to learn is intimately dependent on what they are required to learn and the context in which they must learn it. She proposed a series of studies with children from 1 to 3 years of age learning about the simple mechanism of physical causality. In her study, the children and the mothers sat side by side, and the children restrained by a "sassy seat". She gave them toys and a set of tools. The tools were put in front of the child; it was at least 6 inches from making contact with the toy. In order to reach the toy, the child would need to select a "means for bringing". She found out, there were clear age differences on the learning trials. Most of the children below 24 months needed the mother to demonstrate the solution, only 21% succeeded with no help, while in the older group almost all children solved the learning trial unaided (92%). Furthermore, children below 24 months had special difficulty of learning the solution to solve difficult problems, only 17% solved, while 42% of the children of aged 24-36 months could solve the problems. This result shows the important influence of domain-specific knowledge on children's ability to choose a tool to do an experiment.

Chen and Klahr (1999) confirmed that the ability to design unconfounded experiments and make valid inferences based on outcomes of experiments is an essential skill in scientific reasoning. They wanted to know if early elementary school children are able to understand, create and interpret unconfounded experiments and how they learn and generalize this strategy across various domains. The domain-general strategy they focused on in this study they called the "Control of Variables Strategy" (CVS). CVS is the method for creating experiments in which a single contrast is made between experimental conditions. This strategy is able to allow distinguishing between confounded and unconfounded experiments.

Their study consisted of two parts. Part I included hands-on designs of experiments. Children were asked to set up experiments to test the possible effects of different variables. Part II was a paper-and-pencil post-test given seven months after part I. The post-test examined children's ability to transfer the strategy to remote situations.

The participants were second, third and fourth grade children.

In part I, children were asked to make a series of paired comparisons to test particular variables of each problem in four phases of the study: Exploration, Assessment, Transfer I and Transfer II. In each phase, children were asked to make comparisons in one task to find out whether or not a variable made a difference in the outcome. Three similar tasks were used: Spring, Slope and Sinking.

In the Training Probe condition, children were given explicit instruction regarding CVS, it included an explanation of the rationale behind controlling variables as well as examples of how to make unconfounded comparisons. Besides, children in this condition also received probe questions about each comparison they made.

In the No Training-Probe condition, children received no explicit training but they did receive the same series of probe questions about each comparison as were used in the Training Probe condition.

Children in the No Training No Probe condition received neither training nor probes.

In each of three tasks, there were four variables that could assume either of two values. In each task, participants were asked to focus on a single outcome that was affected by all four variables.

For example, in spring task, children had to make comparisons to determine the effects of different variables on how far springs stretch. Materials consisted of 8 springs varying in length (long and short), coil width (wide and narrow), and wire diameter (thick and thin). A pair of “heavy” and a pair of “light” weights was also used, heavy and light weights differed in shape as well as weight, so that they could be easily distinguish. To set up a comparison, children selected two springs to compare and hang on hooks on a frame and then selected a weight to hang on each spring. Then they observed the stretched springs.

The procedure was divided into four phases spread over two days. The phases Exploration and training (for the Training Probe condition only) and Assessment took place on day 1, and phases Transfer 1 and Transfer 2 took place on day 2, day 2 was separated from day 1 by approximately one week. Participants were interviewed individually (in a quiet place) in their school.

In Part 2: The post-test was designed to examine children’s ability to transfer the CVS strategy to relative remote situations. The post-test consisted of a 15-page packet containing three problems in each of five domains: Plant growth, cookie baking, model airplanes, drink sales and running speed. Each domain involved three two-level variables. For example, in the plant growth domain, plants could get a little or a lot of water, a little or a lot of plant food, and a little or a lot of sunlight. Children were asked to evaluate comparisons that tested the effect of one target variable. The comparisons

comprised four types: unconfounded comparisons, comparisons with a single confound, comparisons in which all three variables had different values, and noncontrastive comparisons in which the target variable was the same in both items in the pair. With each domain, one of three comparisons shown was unconfounded, whereas the other two items were chosen from the three types of poor comparisons.

Comparison pairs were presented both in text and as pictures. Children were asked to circle “good test” if they felt the pictures showed a good way to find out about that variable and to circle “bad test” if they felt it was a bad way.

Chen and Klahr measured four major dependent variables: The CVS score was measured based on the children’s use of CVS in designing tests. Robust use of CVS was based on both performance and verbal justifications about why children designed their experiments as they did. Strategy similarity awareness was based on children’s responses to questions about the similarity across tasks. Domain knowledge was based on children’s responses to questions about the effects of different causal variables in the domain.

They found out that with appropriate instruction, elementary school children are capable of understanding, learning and transferring the basic strategy when designing and evaluating simple tests. The results also showed that explicit training with domains, combined with probe questions, was the most effective way to teach CVS.

In the case of CVS, the relevant processes require the acquisition of a strategy in a specific domain and then implementation of that strategy from one context to another. Third and fourth graders proved capable of transferring a newly learned strategy to other tasks in the same general domain. Second graders, however, showed difficulty in mapping the original task and the newly encountered task, and in implementing the strategy in designing tests. On the other hand, third and fourth graders successfully applied CVS across problems, whereas only fourth graders used the learners, CVS in solving problems with different formats and in different domains after a long delay. In contrast, second graders proved able to use CVS only within the original problem. In brief, second graders transferred CVS only to very close situations. Third graders were able to transfer CVS to both very close and close situations; and fourth graders were successful in transferring the strategy to remote situations.

These findings showed that the effects of domain-specific knowledge play a role in CVS.

In contrast, the authors indicated that strategy training also facilitated the acquisition of domain specific knowledge. In their study, children in the Training Probe condition

improved their domain-specific knowledge, whereas children in other conditions did not.

Lawson and Wollman (1976) did experiments with fifth-grade and seventh-grade students. One of their purposes was to investigate if instructional procedures can be designed and employed to successfully affect the ability to isolate and control variables.

The participants were 32 fifth-grade students and 32 seventh-grade students. In the first experiment, the fifth-grade students were randomly placed into two groups of 16 students each: an experiment group which received training concerning the concept of controlling variables, and a control group which not received training. Both groups were pre-tested in individual interviews with a battery of Piagetian tasks. The experimental-group students then participated in four sessions of individual training; each session lasted about 30 minutes. The control-group students attended their regularly scheduled classes during this time.

Post-testing of all 32 fifth grade students was conducted in two phases. The first phase consisted of individual interviews conducted by two trained experimenters. Three Piagetian manipulative tasks (bending rods, the pendulum and the balance beam) were administered. In the second phase of the post-testing, all 32 students were grouped together and two pencil-and-paper examinations were administered. Students had to respond to a spheres task involving the control of variables, a logic question involving the logical fallacy known as affirming the consequence and finally a combination question.

The experimental design described above was also used for the 32 seventh-grade students, whereby the only changes were made on the post-test. It was decided to use a shortened version of the Longeot examination (Longeot 1962, 1965).

In the pre-test, four Piagetian-styled tasks were administered. The students were asked why they responded as they did.

On the basis of the responses given to three tasks (conservation of weight, conservation of volume and volume displacement) fifth grade and seventh grade students were classified into developmental levels as follows: Concrete-IIA: Nonconservation responses on all three tasks. Concrete-IIB: Conservation of weight and nonconservation of volume and volume displacement. Postconcrete: Conservation of weight and conservation of volume or volume displacement. Formal-IIIA: Conservation responses on all three tasks.

The bending rods task (Inhelder & Piaget, 1958) was used to test students' ability to identify and control variables. Given six flexible metal rods of varying length; diameter, shape and material which were fastened to a stationary block of wood. The students were asked to identify variables and demonstrate proof of the effect of each variable in bending the rods.

Each student in the experimental group was given four 30-minute individual training sessions.

In the first session, students were told that a number of different kinds of materials would be used to teach how to perform "fair tests". The materials used in this session were very familiar to the children, for example, three tennis balls with a different bounce, two square pieces of cardboard. Each student was told that the first problem was to find out which of the tennis balls was the bounciest. The student was to instruct experimenter in how to perform the experiment and the experimenter would carry out the student's instructions. The students were then told that a test was called a "fair test" if all things (variables) that might make a difference were the same in both balls (except for the difference in the balls themselves). A test in which these variables were not the same was called an "unfair test".

In the second session, the materials used were new ones. These materials were used for the bending rods task administered during the post-test. Six metal rods of varying size, shape and material were placed on the table and student was asked to classify them in as many ways as possible. The rods were then placed into a stationary block of wood. Students was asked to perform a "fair test" to find out if the variables of length, thickness, shape and material of the rods, as well as the amount of weight affects the amount of bending of the rods. Whenever a student performed a test, he was asked the following question: "Is this a fair test?" "Why is it a fair test?" "Can you be sure that this rod bends more than that one only because it is thinner?" "Is there any other reason why it might be bending more?" These question were used to focus students' attention on all relevant variables, they helped them to understand the necessity for keeping all factors the same except the one being tested to determine causal relationship.

In the third session, students were asked to experiment with an apparatus called a Whirly Bird (Science Curriculum Improvement Study, 1970). The Whirly Bird consisted of a base which holds a post. An arm is attached to the end of the post. When pushed or propelled by a wound rubber band, the arm will spin around like the rotor on a helicopter. Metal weights can be placed at a various positions along the arm. Students were briefly shown how the Whirly Bird worked and were given the task of finding out everything they thought might make a difference in the number of times the arm would

spin before it came to rest. Afterwards, the students were asked to perform “fair tests” to prove that the independent variables mentioned actually did make a difference in the number of times the arm would spin. Again, whenever, a test was performed students were asked questions such as these: “Is this a fair test?” “Why is it a fair test?” “Does it prove that it makes a difference?”.

In the fourth session, they were given written problems instead of concrete materials. Probing questions relative to students’ understanding of the written situations were asked as was done in the previous sessions. However, in this session, learning by doing was replaced by learning by discussion. The following two written problems were presented: Written problem 1: Five pieces of various parts of plants were placed in each of five sealed jars of equal size under different conditions of color of light and temperature. At the start of the experiment each jar contained 250 units of carbon dioxide. The amount of carbon dioxide in each jar at the end of the experiment was changed. Students were asked to choose two of the jars to make a fair comparison to find out if temperature makes a difference in the amount of carbon dioxide used.

Written problem 2: An experimenter wanted to test the response of mealworms to light and moisture. He set up four boxes. Box A: light but no moisture; Box B: moisture but no light; Box C: both light and moisture; and Box D: neither light nor moisture. He used lamps for light sources and watered pieces of paper in the boxes for moisture. In the center of each box he placed 20 mealworms. One day later he returned to count the number of mealworms that crawled to the different ends of the boxes.

In the post-test, they used the bending rods as described above for the pre-test in addition to some other tasks. The pendulum task (Inhelder & Piaget, 1958) tested the children’s ability to control and exclude irrelevant variables using a simple pendulum. The balance beam task (Inhelder & Piaget, 1958) tested students’ ability to balance various combinations of weights at various locations along the beam. The Peel questions (Peel, 1971) were used to assess a degree of consistency of level of judgement among series of similar passages.

In the second phase of post-testing, the following measures were administered: The spheres task (Wollman, 1975) consisted of three written questions requiring understanding of the necessity for the control of variables in the context of rolling spheres down inclined planes. The Longeot examination (Longeot, 1962, 1965) is a subject matter-free examination, consisting of 28 problems requiring either concrete, transitional, or formal operational thinking for a successful solution. However, as the time was limited, the examination consisted of 8 problems. A combinatorial question

involving combinatorial analysis was given to the sample of fifth-grade students. Finally, a logic question was also used.

Lawson and Wollman indicated that instruction incorporating the described procedures can affect the transition from concrete to formal cognitive functioning in these fifth- and seventh-grade students with respect to the ability to control variables. Both fifth- and seventh-grade experimental group students performed at the formal level on the post-test; however, the seventh-grade students performed at slightly more formal level than the fifth-grade students. The experimental groups also performed significantly better than control groups on the specific transfer tasks, i.e. the training was generalizable to tasks involving novel materials.

In summary, Lawson and Wollman performed many studies with fifth-grade and seventh-grade students about their ability to control variables, to design and evaluate experiments (fair or unfair). They indicated that these abilities of students were influenced not only by differences in grade, but also by their knowledge about content (familiar or unfamiliar knowledge) as well as by training that helped students have a knowledge about strategy for doing experiments.

2.2.3. Domain-specific Pre-knowledge and Evidence Evaluation

Schauble (1990) used the race cars microworld in order to investigate the role of prior knowledge and strategies for evaluation evidence. Children attended eight weekly sessions to explore a race cars microworld, they had to control experiments in order to find out which of five design features (engine, wheels, tailfin, muffler and color) had a causal relation to car speed. Before doing experiments with the race cars microworld, the children were interviewed to find out their beliefs about the domain, and after the final session, they were interviewed again in order to summarize their current beliefs about the causal status of each of the features. Through this study, Schauble wanted to investigate changes in the children's domain knowledge and reasoning processes over an extended period, and to identify interactions between knowledge and reasoning. She found out that, when children's pre-knowledge about the features were disconfirmed by evidence, they made judgements that showed some influence of evidence, but also used invalid judgements in order to cling to their own theories. Even when their belief was completely disconfirmed by evidence, children did not necessarily abandon it altogether.

These results show an influence of pre-knowledge on evidence evaluation.

Schauble, Klopfer and Raghavan (1991) studied the understanding of experimentation. They formulated the hypothesis that children use the engineering model of experimentation to produce a desired outcome.

They presented 16 fifth- and sixth-grade children two problems designed to elicit an engineering model and a science model of experimentation, and compared their processes of experimentation.

The structure of the engineering task is consistent with the goal of trying to reproduce a desired outcome. They based the experiment on that of Franklin (1768, cited in Goodman, 1931) to learn how water canals should be designed to optimise boat speed. In the system used by Schauble et al., one can vary the depth of the canal (shallow or deep) by moving a portable floor, the shape of the boats (circle, square, or diamond cross section), the boat size (large or small), and the boat weight (light, or unloaded, versus heavy, i.e. loaded with a small barrel). The children designed the experiments by setting up the variables, running the boat and recording the travel time. The children were told in advance that they have to find out how to design the canal and the boat system so that people could travel on the canal as quickly as possible.

In contrast to an engineering task, a science task does not include results that can be interpreted as more or less desired, and thus does not support the reinterpretation of the goal as does an engineering objective. This task has been used to study children's conceptions about buoyancy and water pressure (Champagne, Klopfer, & Chaiklin, 1984). In this version, an object was suspended on a spring into a fluid to observe the effects of buoyancy forces on objects of different mass (large, intermediate and small) and volume (small, medium and large) as well as four standard positions that the object could take with respect to the water (outside of the water altogether, immersed just beneath the surface, halfway down, close to the bottom). Children have to explain how scientists perform experiments to learn what features make a difference in how a system works (which features make a difference in the length of the string).

This task differs from the engineering task in two different ways: First, it needs not be addressed by the try-and-see method; second, there is no desirable outcome to distract the subjects from the objective of understanding how the system works.

The children spent three 40-minute sessions working on each of the two tasks in counterbalanced order. During each session, the children designed and ran experiments until time expired; the numbers of experiments were recorded.

In each case, the children had to design the experiment, to record it, to make a prediction, to run the experiment and to interpret the outcome.

Before beginning the experimentation with each of the two tasks, and again after the six sessions, the children were interviewed about their understanding of the system, including the role of each variable.

Their findings indicated that the subjects in the science problem context generated experiments that covered a greater proportion of the Experiment space than did those children assigned to the engineering context. On the other hand, with the exception of one variable, the distributions of inferences made by subjects in the science context and by those in the engineering context were equivalent. However, within the spring problem, children in the science context devoted a significantly greater percentage of their inferences to the variable object size than did children in the engineering context. Children were confident before beginning experiments that size was irrelevant, and only two children changed their minds during the time exploring the spring problem. Only 7% of the inferences of the children in the engineering context concerned size, whereas 22.6% of the inferences of the children in the science context were about size.

They also found that children's inferences (judgements made in response to the interpretation question asked after each experiment) were influenced by their pre-knowledge. They made greater proportions of causal inferences than noncausal inferences, and more inferences were devoted to conclusions about the variables they originally believed were causal. Reliance on knowledge certainly is also reasonable, one can make inferences based on correct pre-knowledge, and experimentation is usually driven by prior theories. However, in some cases, the beliefs were incorrect, but students tended to accept evidence that was consistent with their prior beliefs, but to either distort or fail to generate evidence inconsistent with their prior beliefs. For example, the children who initially believed that boat weight related to speed did not change their minds during experimentation, in spite of the fact that changing the weight affected the speed of small but not large boats.

In sum, Schauble et al. confirmed that knowledge and experimentation strategy interact, effective experimentation entails noting the relations between one's evolving theories and one's interpretation of the evidence.

Shaklee and Paszek (1985) used two sets of 12 covariation problems, each of them structured to produce a distinctive pattern of solution accuracy, and a rule analysis methodology was used in order to investigate covariation judgments of different aged children. Children were given pictures which illustrated information about combinations of alternative event states of two potentially related events (e.g., plant healthy or not healthy; plant food present or absent) and were asked to identify the direction of the

relationship. The students had to compare the illustrated pictures and events in order to recognize which picture can be related to any event, and the pictures and events belonged to a set of problems (e.g., the picture of healthy plant is related to the picture plant food present, and the picture of weak plants is related to plant food absent). In order to solve these problems, children need knowledge about strategies in analysing data, evaluating evidence or mathematic strategies.

In some studies, researchers examined children's judgements in theory choice tasks. Samarapungavan (1992) performed two experiments about two areas of astronomy and chemistry for first, third and fifth grade children. He wanted to know what kinds of criteria children use to evaluate the adequacy of their knowledge.

In the domain of astronomy, each child was given a set of questions that were developed by Vosniadou and Brewer (1994). Using the questions the children's beliefs about the shape, relative locations and movements of solar objects were tested both verbally and with pictures and a physical model. Based on their responses, the children were scored as having geocentric or heliocentric frameworks. If children indicated that the earth and moon moved while the sun was stationary, and that the sun is located in the center of the solar system, they were designated heliocentric. Whereas, if they said that the sun and the moon moved relatively to the earth and the earth is located in the center, they were designated geocentric.

Three sets of materials for theory choice tasks were developed; these were related to geocentric astronomy, heliocentric astronomy and chemistry.

Geocentric astronomy: The children watched while a rod and a wooden square were released from a small distance above the ground and allowed to fall, and then a small helium balloon was released and allowed to rise. Inconsistent theory (T1): the rod and the square fell to the ground because the gravitation of the earth pulled things down toward it. Consistent theory (T2): The air is like water. Some things are "heavy" and fall to the bottom, but other things are "light" and float on top.

Heliocentric astronomy: The children observed four pictures of the night sky depicting the phases of the moon (full moon, half moon, crescent moon and no moon). T1 explained the phases of the moon in terms of occlusion by clouds. T2 explained the phases of the moon by proposing that the moon had a light side and a dark side and that the moon rotated, so that its spinning caused the appearance of the phase on earth.

Chemistry task: The children were shown five jars mounted on boxes that were labelled either "hot" or "cold". The jars contained either blue, red, or colorless liquids. A pH indicator stick was dipped into each of the liquids and any color changes in the indicator

stick were noted. The stick changed its color to blue after being immersed in the blue and colorless liquids and it changed its color to red after being immersed in the red liquids. T1, the narrower theory, proposed that the liquids in the jars were dyes that coated the sticks with their color (it could not explain why the stick changed color in the colorless liquid). T2, the broader theory, proposed the function of the stick was to indicate the temperature of substances and that it turned red in hot substances and blue in cold ones.

In each task, children were asked to justify their choice of theory. The justifications were assigned to three categories. The first was the “criterion-based” justification that represented the theory choice based on the criterion being tested. The second was “content-based”, i.e. based on the “truth” or “falsehood” of the conceptual content of the theory; and the third one was “no justification”.

The findings indicated that 86% of all justifications of correct theory choice were criterion-based. Only 10% of the correct justifications were content-based.

All children did better in the knowledge-consistent condition than in the knowledge-inconsistent condition, and this was especially true for the fifth grade. In grade one, 78% children chose the theory of wider range in the knowledge-consistent condition, but only 20% did so in the knowledge-inconsistent condition. In third grade, 94% of the children chose the theory of wider range in the knowledge-consistent condition. In the fifth grade, this rate was 98%. However, the effect of the grade was not significant in the chemistry domain.

In summary, primary school children can use metaconceptual criteria to evaluate ideas as scientists do in selecting among competing theories; this is especially true when the conceptual content of the theories to be evaluated is consistent with children’s prior beliefs.

2.2.4. Domain-specific pre-knowledge and combination of hypothesis generation & evidence evaluation

In this process, participants were asked to produce hypotheses which must be consistent with evidence. They do not have to design new experiments and need not produce new evidence.

The investigation by Stella Vosniadou and William Brewer (1992) is an example. In their study, the development of children's conceptual knowledge about the earth's shape was investigated. Children assumed that the earth is flat, that is supported by everyday experience, but adults claim that the earth is a huge sphere, and children have difficulty to understand and reconcile these views. Vosniadou and Brewer asked children two kinds of questions, factual and generative ones. Factual questions were related to certain theoretically important facts, and generative questions to children's underlying conceptual structure. Based on the children's responses about the earth's shape, Vosniadou and Brewer analyzed and described the process of changing from the initial mental model to the mental model of a spherical earth, which included a rectangular earth, a disc earth, a dual earth, a hollow sphere earth, a flattened sphere earth, and finally, a sphere earth. The children answered the questions and generated a hypothesis, then they evaluated the available evidence in order to test their hypothesis. By means of this process, the conception of the children about the earth's shape changed. The children had to answer the questions based on their pre-knowledge; they did not perform any experiments.

If in a problem solving process, the cause of an event is to be identified, a first step in achieving the solution consists in deciding upon which factors that are merely potentially causal are likely to be actually causal (Koslowski and Okagaki, 1986). Koslowski and Okagaki performed a corresponding study with college students, as well as with college-bound 14- and 11-year-olds. The participants were presented a series of stories. Each story consisted of an initial description of the problem situation followed by a more recent report about causal mechanisms, analogous effects, sampling procedures, and alternative hypotheses. Subjects then were asked to make a decision about a potential cause factor.

An example of such story is described in the following: A man who makes pottery notices that some of his pottery is likely to crumble easily. Known fact: The potter baked his pottery in a low-heat oven. Nothing is known about the other pottery which does not crumble. Usual causes: Pottery crumbles if it is made with the wrong kind of clay, if it is made with the right clay, but the clay has impurities in it, and if there is

something wrong with the glaze. The potter had not yet had time to check and rule out the actual cause. So he makes an estimate about how likely it is that the low-heat oven has something to do with the potteries crumbling. He gives this cause an average score on a scale of 1-10.

Afterwards, the participants were asked to indicate whether they expect the potter to think more or less the same as last time, or not.

Eight story problems were constructed with the same format.

Subjects had to use their specific knowledge in order to evaluate the hypothesis and they had to make decisions. They did not plan any experiments.

The findings indicated that adults as well as adolescents judge the factor to be causal when they learn the factor covaries with the effect. Subjects' tendency to judge the target factor to be causal becomes exaggerated when subjects learn that the usual causes of the effect have been ruled out.

2.2.5. Domain-specific pre-knowledge and combination of experiment design & evidence evaluation

Some researchers wanted to investigate the development of subjects in scientific reasoning by studying the combination of experiment design and evidence evaluation.

Kuhn and Angelev (1976) explored children's abilities to design experiments and evaluate data in chemical problems and in a pendulum problem. They wanted to test the hypothesis that preadolescent subjects would begin to develop formal operations if they were given opportunities to work on situations which require formal thought over a period of time.

In the pendulum problem, the subjects were presented a pendulum and a set of 4 weights which could be fastened to the end of string by means of a hook. The string could be shortened to various lengths. The other variables could be the height of the release point or the amount of force applied. The subjects were asked to do experiments and try to find out what makes the pendulum go faster or more slowly.

In the chemical problem, they showed children five beakers containing colorless, odourless liquids. The students' task was to try to obtain the yellow color by mixing different combinations of the chemicals. They had to make a plan of all mixtures and to write them on the sheet, and then run the experiments.

Both of the above problems focused on designing experiments, evaluating outcome and drawing the causal factors. The subjects did not need content knowledge, but the strategy knowledge in doing experiments and evaluating evidence.

2.3. Competency model and mistakes in experimentation

2.3.1. Models for developing competencies in experimentation developed by Hammann (2004)

Focusing on the experimentation area, Hammann (2004) developed “Models for developing competencies” (Kompetenzentwicklungsmodelle) for the three dimensions of experimentation: Search in the hypothesis space, search in the experiment space and data analysis. In each dimension of experimentation four levels of competency were described.

Search in the hypothesis space. *Level 1:* Experimenting without hypotheses. Students experiment without hypotheses. That is, they do experiments without having any idea about cause-and-effect relationships; they try to achieve a certain effect. *Level 2:* Unsystematic search for hypotheses. Students search for hypotheses but do not look for all hypotheses that are necessary for answering a question or while searching for hypotheses, students do not logically relate hypotheses to each other. *Level 3:* Systematic search for hypotheses. Students form multiple hypotheses that are logically related. Problems, however, occur when hypotheses need to be revised. *Level 4:* Systematic search for hypotheses and successful revision of hypotheses. Like level 3, however, in contrast to level 3, revision of hypotheses also is successful in situations in which all hypotheses originally formed have been falsified.

Search in the experiment space. *Level 1:* Unsystematic use of variables. Learners unsystematically change one or several variables so that no logical statements can be made (confounded variables, “change all”, “no plan”, “intuitive”). *Level 2:* Partially systematic use of variables. More systematic use of variables (local chaining HOTAT: hold-one-thing-at-a-time), but still deficits concerning the systematic variation and control of variables. *Level 3:* Systematic use of variables in known domains. Learners only vary the testing variable while keeping the other variables constant. This procedure allows for explaining the effect of one variable. *Level 4:* Systematic use of variables in unknown domains. Like level 3, but learners also use this strategy successfully in unknown domains.

Data analysis. *Level 1:* Data are not related to hypotheses. Observed effects are described, but causes are not explained. There are deficits because of a lack of understanding about the aims of gathering data when experimenting. *Level 2:* Illogical

analysis of data. Learners relate data to hypotheses. However, their conclusions are illogical, i.e. students neglect relevant contrasts between the experimental tests and the control. Explanations of experimental results are contradictory. Hypotheses are changed or kept up in the face of data that do not allow doing so. *Level 3*: Mostly logical analysis of data, but problems with data that contradict students' expectations. Students explain data in a logically consistent way in most experiments. Difficulties exist when it comes to anomalous data. Data that contradict one's own expectations are often ignored or misinterpreted. *Level 4*: Logical analysis of data. Students successfully analyze data even when the data are difficult to interpret due to the students' expectations or due to the conditions of data gathering (for example, continuous variables with small measuring differences or measuring errors).

2.3.2. Mistakes in experimentation

Experimentation plays a role in scientific reasoning process, it is an important opportunity for fostering student reasoning in the science classroom and experimentation also affords a rich context for studying the process of theory change, central to learning in general (Schauble, Glaser, Duschl, Schuze and John, 1995; Schauble et al., 1991).

Thus, it is necessary to improve the ability of students in experimentation. However, many studies indicated that students usually make many mistakes in doing experiments. Basing on the competency model in experimentation by Hammann (2004) we studied the mistakes students usually encounter in doing experiments.

International school achievement studies prove German students' weakness in the development of experimental tasks. The problem is where the causes of these deficits lie. Frequently, mistakes are made, because students show conceptions about the experimental method, which differ from the scientific conceptions. For example, students try to obtain an effect or a particularly good result by planning and executing an experiment instead of explaining cause-effect relations systematically. An important difference to the scientifically correct application of the experimental method is that natural conditions are changed in such a way that it is possible to explain causes and effects. Mistakes in experimentation frequently result from such methodical student conceptions. Therefore, a priority is to change student conceptions about the experimental method in experimental instruction of areas in biology, chemistry and physics and to develop an adequate method understanding.

2.3.2.1. Student's conceptions about experimentation

A set of investigations was concerned with the question which conceptions students possess about the experimental method (Schauble, Klopfer, Raghavan, 1991; Carey, Evans, Honda, Jay, Unger, 1989). They showed that methodical knowledge cannot be presupposed in experimentation. Typical conceptions about the experimental method were clarified on the basis of an experiment about seed germination (Hammann et al. 2006). In this study 5th grade students (High School) had to plan an experiment about seed germination. The students set up the statement that seeds need water, air and nutrients from the soil to germinate. But this statement remains unproved, because the students performed their experiment without a control approach. Furthermore, the students did not manipulate variables, thus substantial characteristics of an experiment were missing. Causes for these deficits become obvious by critical analysis of the argumentation of one student. He says in a general manner: "The seeds need the factor X, therefore I add this so that the seeds germinate". This refers to a fundamental misunderstanding about experimentation which can frequently be found. A substantial conception of students about the experimental method is that one must obtain an effect, not test cause-effect relations under controlled conditions. The above mentioned student wanted to reach that the seeds germinate. Thus, he creates the "correct" conditions for it, however, the conditions that he considers correct, are not proven. He argues: "I put seeds three cm under the soil. The seed needs water to grow, therefore I add water. The seed takes up air and nutrients and grows".

For this proceeding the terms "confirmation bias" (Wason, 1960) and "failures to seek disconfirmation" (Klayman & Ha, 1989) are formed. The tendency is to raise merely confirming data and evaluate the predicted combinations as evidence. This strategy of "positive testing" is, however, problematic and masks the risk of wrong conclusions. This becomes particularly clear in the present case by the absence of an experimental control. Only by the comparison between the experimental approach and the control approach could the pupil have found out that it is against his statement, seeds do not need to take up nutrients from the soil for germination. The planning by the pupil can be compared further with procedures of engineers (Schauble, Klopfer, Raghavan, 1991), because they want to obtain frequently determined effects by the optimization of technical products, it is not necessary to fathom the influences of the concerned factors systematically like scientists. Instruction approaches for the training of experimental competency see, therefore, in the change of student's conceptions about the experimental method an effective instrument that bring pupils a systematic and reflected

procedure: They should learn to do experiments like scientists and not like engineers (Carey, Evans, Honda, Jay, Unger, 1989).

2.3.2.2. Deficits when planning experiments

Unsystematical handling of variables

Systematic handling of variables is one of the most important characteristics of the experimental method (Puthz, 1988). It is necessary to differentiate between two variable types: the test variable and the variables which can be controlled. On condition that only the test variable is varied and the variables which can be controlled are kept constant, it is possible to test the effects of the test variable. The distinction between test variable and controlled variables is not always easy and forms an important cause of confounded experiments. Which variable must be changed and/or kept constant is to be decided by the learners. In the Third TIMSS study, in a task on seed germination, a suitable control approach must be selected for one given experimental approach (a plant is put in sand with the mineral nutrients; it is also watered and put in the sunlight). The hypothesis is “the plant needs mineral nutrients from soil to grow healthy”. The solution of the task is simple, if the principle of the control approach is understood. In addition, it must be recognized that the variable “mineral nutrients” can be changed, whereas the others which can be controlled must be kept constant. Nevertheless, the task was incorrectly solved by the majority of the students, because 60% of the seventh grade students and 58% of the eighth grade students confounded the two variable types and selected a wrong approach as control approach.

A similar task was set in our own investigation (Phan, Hammann, Bayrhuber, 2004). The task was carried out by 490 students of age between 10 and 12 years (fifth and sixth grade students). The students were asked to choose an appropriate control approach for the given experimental approach (seeds are put in soil, 22°C and light). The hypothesis here was “Seeds germinate better if it is warm”. 55.5% of the students selected the correct control approach (answer A: soil, light, 10°C), while over a third of the students changed another variable in addition to the test variable and thus selected a confounded experiment. 7% of the sample changed both the test variable and the variables which can be controlled (pot D). This strategy, in which all variables were changed, can be called “change all” (Tschirgi, 1980).

Illogical relation of approaches in test series

The deficits described in the preceding sections prove particularly impedimental, when in a test series of a control group is to be planned. Students without method training tend to an unsystematical procedure and confounded variables. Two studies showed this, by which it was investigated, how students independently plan experiments: In one study, the students were asked to determine which chemical substances have to be combined in order to produce a solution with a certain color (Kuhn, Angelev, 1976). In the other study, subjects had to explore which factors affect the speed of boats in a channel filled with water (Schauble, Klopfer, Raghavan, 1991). The studies showed typical deficits in changing the characteristics of variables and different test approaches were illogically related to each other. For example, the students conducted the necessary changes of the test variable (e.g. the boat size), but the controlled variable was also simultaneously changed (e.g. weight of the charge). Of course, no valid conclusions can be derived from such experiments.

The students showed similar deficits when they had to decide which test series are suitable to test a given hypothesis. For example, in TIMSS a test series is to be selected to test the prediction: “The heavier the car is, the faster is its speed at the foot of the ramp”. In this task, different development levels of three variables must be considered, i.e. the angle of inclination of the ramp, the weight of the charge and the wheel size. The correct solution was chosen by only 28% of the seventh grade students and 35% of the 8th grade students. About two thirds of the students of the lower secondary level did not recognize the influence of one or several disturbing variables and thus committed typical mistakes in construction of a test series.

2.3.2.3. Deficits in the data analysis

Important aspects of the data analysis are the appropriate consideration of all data, their critical interpretation and the evaluation of the disposed hypotheses. Reasoning plays an important role, because different information must be united and linked logically (convergent thinking). In particular, conclusions must be drawn from the data regarding the underlying hypotheses of the experiment, so that the hypotheses can be either confirmed or revised partly or completely. Students, however, show serious deficiencies.

No proven causality

Cause-effect relations can be clearly determined only when test approaches which only differ from the application of the test variable, are compared. First, if an effect arises by the present of the test variables, but is missing by omitting the test variables, these variables can be evaluated as causal for the observed effect. On the basis of only one test approach, no statements about cause-effect relations can be met. Students consider this frequently, but do not draw causality, although this cannot be derived from the results of an experiment without a control approach.

This became clear in an investigation by Hammann and his colleagues (2006), in which students had to evaluate whether a given experimental arrangement is suitable in order to prove causal connections. In one experiment, butter was produced by addition of lemon juice. In the experimental arrangement “lemon juice” was used as test variable. The observed effect is that butter is produced. It is not tested whether butter can be also made without the addition of lemon juice (control approach).

In the study, 13 students (fourth and fifth grades) accomplished the experiment for butter production first. They observed that butter is produced. The students were asked directly referring to the investigational procedure, whether one could say clearly by means of the experiment that one needs lemon juice to produce butter. The students had to justify their decision. The answers of the students were specified as follows: Eight students (62%) indicated that the experiment proves clearly that lemon juice is necessary in order to produce butter. Five students (38%) decided for the opposite. These five students pointed out that a control approach was absent. One student judged that “we did not try whether it is also possible to make butter without lemon”. The other 4 students referred to content reasons, like e.g. “No, one does not need lemon, because one can also make butter in such a way”. Although in these reasons the absence of the control approach is not explicitly referred to, the argumentation of some these students shows implicitly the idea of the control approach.

Illogical conclusions

Students do not only plan confounded experiments, they also draw illogical conclusions from methodically correctly planned experiments. These was shown by the results of our own investigation (Phan, Hammann, Bayrhuber, 2004). In the study, students were shown an experiment by pictures and text about seed germination (pot 1: soil, water, light and 22°C; pot 2: no water; pot 3: cotton wool instead of soil, seeds germinated in pot 1 and pot 3, but not in pot 2). Altogether nearly half of the students (n = 490, age

10-12) referred to the variables soil, warmth and light, while from the results only the factor water is to be determined as condition for seed germination. Because the factors light and warmth were not varied at all in the experimental arrangements and a change of the variables soil or cotton wool in both cases led to seed germination, only the omission of the factor water brought a critical change of the result. However, 33.6% of the students chose the disturbing variable soil and 14.3% of the students evaluated warmth and light as crucial conditions for the seed germination. 8.4% of the students selected the answer that stated that the experiment did not obtain the desired effect and did not succeed.

Why did less than half of the students solve this task? Certainly, there is not a lack of logical thinking abilities of the students in this age group. Rather the conclusions of the students could have been guided by their initial beliefs. The fact is well documented that content beliefs affect the interpretation of test results, in particular, when the experimentally obtained data do not confirm the initial beliefs. This aspect is deepened in the following section.

Missing recognition of anomalous results

The lack of recognition of anomalous results is caused by two incorrect ways of thinking. On the one hand, learners plan experiments to confirm their own expectations as initially illustrated in section 2. On the other hand, learners in the data analysis attempt to confirm the initial hypothesis (“confirmation bias”) (Wason, 1960), even if the data require an alteration or a distortion of the hypothesis. Mayer (1999) shows that, experimental data serve only as means of the confirmation of hypotheses and not of their refutation. He demands a stronger emphasis of the hypothesis revision in instruction of academics. Frequently, the positioned hypotheses originate from strong content beliefs which are proved in the daily life. An inadequate analysis of experimental data in this case is not caused by a lack of logical thinking of the students, but by the tendency to confirm proven concepts in everyday life. Students show thereby a set of strategies in order to hold to their beliefs, even if these are disproved by the experimental results. These strategies were explored in detail (Chinn and Brewer, 1998). For example, students predict that a heavier object falls faster from the same height than a light one. If the observation does not confirm this assumption, so that between both objects no temporal flight time difference can be recognized, the students conclude that the objects - against the indication of the instructor - must have the same weight. Thus the students ignored the unexpected test results and/or interpret them according to their own beliefs (Champagne, Gunstone, Klopfer, 1985).

If the test approaches and the experimental results are arranged more complexly, not all data, but only the contradictory ones are possibly rejected. Also, in this way one's own expectations can be confirmed with data, although the results of the experiments demand another interpretation (Schauble, Klopfer, Raghavan, 1991; Chinn, Brewer, 1998; and Metz, 1998).

Another reaction to unexpected results is the conclusion that clear statements are not yet possible due to the available data situation (Schauble, Klopfer, Raghavan, 1991). Instead, further experiments were accomplished in hoping to receive finally confirming data. Though such a repetition is a probate means for the test of the reproducibility of the obtained results, no fundamentally different results can be created. Likewise, it is impossible to yield the expected results by a change of the experiment (Carey, Evans, Honda, Jay, Unger, 1989).

2.3.2.4. Deficits when setting up and testing hypotheses

Substantial characteristics of the experimental method consist in setting up and testing hypotheses. Systematic theory-conducted experimentation is impossible without the consideration of hypotheses, because hypotheses guide the planning of the experiment and the data analysis. The goal of each experiment is the test of the initial hypotheses.

Making mistakes when recognizing the hypotheses which were tested

Students frequently completely experiment without hypotheses. We (Phan, Hammann, Bayrhuber, 2004) tested, whether students can recognize those hypotheses which were tested by an experiment about seed germination. In this experiment the combination of the variables soil and water were tested and possible hypotheses to the selection were given. 490 students aged between 10 and 12 years worked on this task. The majority of the students took the wrong hypothesis combination of the experimental arrangement and selected too many (51.9%) and/or not relevant hypotheses (14.7%). They did not recognize the combination between the variables soil and water. In addition, only 23.9% of the students made a correct choice. Furthermore, 9.5% of the participating students regarded the reaching of the effect, namely germination of seeds as the goal of the experiment. They understood the experimentation, thus, like an activity of an engineer, in which no hypotheses are tested.

Missing reference to subsequent hypotheses

It is often hard for students to control a planned test series with reference to subsequent hypotheses. They often conduct series of tests completely unrelated to each other or

compare only directly successive pairs of experiments. This procedure is called “local chaining” (Schauble, Klopfer and Raghavan, 1991). Thus, hypotheses were not systematically tested, but only regarded in parts. This procedure frequently accompanies an unsystematical use of variables.

This deficit in proceeding shall be illustrated by means of an example (speed of boat in a channel) (Schauble, Klopfer, Raghavan, 1991). In the experiment, students are asked to determine which factors affect the speed of a boat in a channel. The boat size and boat form, weight of the charge and height of the water level in the channel can be varied.

A typical student in this study formulated the following hypothesis: “The weight determines the speed: light boats drive faster than heavy ones”. He compared a large rectangular heavy boat in a channel with low water level (experiment 1) with a small round light boat in a channel with high water level (experiment 2). He showed that unsystematical handling of variables already described when he changed more than only the test variable. He did not accomplish further attempts to test this hypothesis, although still no valid conclusions were possible from the connection between weight and speed. Instead, in the following attempt he tested the hypothesis “the boat form determines the speed: round boats drive faster than square ones”. In addition, he compared the result of experiment 2 with the results of the experiment with a small light boat in diamond form in a channel with low water level (experiment 3). In the following attempts no systematic approach between subsequent hypotheses were revealed. Later experiments which served to the investigation of the same hypothesis, were not related with earlier results.

Making mistakes by strong containment of hypotheses

Test results of an experiment can be contrary to the initially positioned hypotheses, so they must be revised. In this case, it could be documented that students perform the search for new hypotheses too intensely, so that these count only for a part of the results (Schauble, Klopfer, Raghavan, 1991). A further investigation clarifies this too. In this study, the test participants were asked to find out which rule the numerical series 2-4-6 follows (Wason, 1960). They could test on own whether the series corresponded to the assumed rule or not.

The participants named mainly number sequences that corresponded to the rule “2 rising numbers” (e.g. 7-9-11 or 22-24-26) and not ones which contradicted the assumed rule (e.g. 7-8-9). However, this confirmation strategy led to wrong conclusions, because the

correct rule is: “rising numbers” and not “2 rising numbers”, as guessed by most test participants. The assumed hypothesis was formulated too close and thus tested also only within its close borders. The test subjects did not consider the possibility that the series mentioned above could apply also to another rule.

This procedure of hypothesis formation is characterized as “the positive capture” problem. This refers to that learners are frequently not able to consider all potential hypotheses at all but only consider a certain part of the hypothesis space, indeed that which contains the most obvious hypotheses (Fay and Klahr, 1996). In the example mentioned above the continuous distance of the series 2-4-6 is considered as dominant characteristic. This is tested as only characteristic, although the series 2-4-6 also still could follow different rules, e.g. the rules “natural numbers”, “even numbers”, “numbers which are smaller than 10”. As in the case of an iceberg, in which only 10% are over the water surface, parts of the hypothesis remain space hidden, if only the rule “2 rising numbers” is tested. If one wants to determine the size of an iceberg, one must more exactly investigate its delimitations and may not settle for the initial impression. By skilful choice of the number sequence (e.g. 7-8-9, from which one expects that it does not belong to the underlying rule) the alleged hypothesis can be also more exactly tested. Thus, one receives more comprehensive information which can be used for a more reliable evaluation of the validity in hypotheses. Only under these conditions is the search in the hypothesis space comprehensively exhausted and arranged effective.

Summary

Important sources of mistakes in experimentation are related to the three fundamental processes of hypothesis formation, planning experiments and data analysis. Substantial deficits when setting up and testing hypotheses concern the problem that not all relevant hypotheses are considered and only those hypotheses which correspond to one’s own expectations are tested. When planning experiments the control group frequently fail. Variables are unsystematically varied and parts of a test series are illogically related to each other. In the data analysis students give statements about the causes of effects. Although these are not proven, they draw illogical conclusions and reinterpret experimental findings, in particular when these contradict their own expectations. These important types of mistakes are used for a theory-led promotion of competency in experimentation (Hammann 2004).

2.4. Assessment of levels of students' competency in scientific literacy and in experimentation

2.4.1. Assessment of levels of students' competency in scientific literacy

One of the important steps to improve the quality in science education is the evaluation of students' achievement. In recent years, many researchers focused on the assessment levels of the competency of students instead of the assessment in general. By means of this method of evaluation, one can assign students to different groups, and for each group one can determine which level of competency the students reached. In different groups different methods of improving the student abilities in learning can be used.

PISA (The Programme for International Student Assessment) is designed to assess the achievement of 15-year-olds in reading, mathematical and scientific literacy through a common international test.

In PISA, students were assigned to a proficiency level based on their probability of answering correctly the majority of the items in that range of difficulty. A student at a given level could be assumed to be able to correctly answer questions at all lower levels. To help in interpretation, these levels were linked to specific score ranges on the original scale (Source: Organisation for Economic Cooperation and Development, Programme for International Student Assessment, PISA, 2003).

PISA was launched in 2000. In that year, five proficiency levels in scientific literacy were defined as follows:

Level 5 (above 661 points): Analyze scientific investigations concerning design and tested predictions. Use data as evidence in order to evaluate alternative aspects or different perspectives. Communicate scientific arguments and/or descriptions in detail and accurately. Develop and apply simple conceptual models in order to give predictions or explanations.

Level 4 (from 553 to 660 points): Identify or formulate information that one needs for investigation additionally in order to be able to draw valid conclusions. Apply data systematically to statements about probable conclusions and develop a chain of arguments. Communicate simple scientific arguments and/or descriptions. Apply elaborated scientific concepts in order to give predictions or explanations.

Level 3 (from 497 to 552 points): Identify detail of scientific investigations, recognize questions that can be answered by a scientific investigation. Distinguish between relevant and irrelevant data when drawing and evaluating conclusions or choose a chain of arguments. Apply scientific concepts in order to give predictions or explanations.

Level 2 (from 421 to 496): Determine variables that one has to control in investigations in simplified context, name questions that can be answered scientifically. Draw and

evaluate conclusions with reference from data or scientific information. Apply scientific everyday knowledge in order to give predictions or explanations.

Level 1 (less than 420): Draw or evaluate conclusions based on scientific everyday knowledge. Reproduce simple factual knowledge (eg. terms, expressions, facts, simple rules).

Studying Scientific Literacy, Bybee (2002) proposed four levels of competency based on the understanding of basic science concepts, principles, processes and relationships between individual components in science. The lowest level was nominal scientific literacy and the highest level was multidimensional scientific literacy.

Nominal Scientific Literacy: Students identify terms and questions as scientific. However, students show deficiencies concerning themes, problems, information, knowledge and understanding. Misconceptions about science concepts and processes. Inadequate explanations of phenomena. Naive statements about science topics.

Functional Scientific Literacy: *Uses scientific terms*. Students define scientific terms correctly, learn technical terms by heart.

Conceptual and procedural scientific literacy: *Understanding of science concepts*. Understanding of procedural knowledge and skills in science. Understanding the relationships between individual components of a science discipline and their conceptual structure. Understanding of basic science principles and processes.

Multidimensional Scientific Literacy: Understanding of what is particular about the sciences. Distinguishing between the sciences and other disciplines. Knowledge about the history and the nature of science. Understanding of the sciences in their social contexts.

Bybee's point for developing scientific literacy can be understood as the development of competencies.

2.4.2. Assessment of levels of competency of students in experimentation

Carey (1989) gave students a two-week series of yeast lessons which were developed to introduce the constructivist view of science and interviewed them prior and after practising in order to assess students' levels of understanding about the nature of scientific knowledge and inquiry. She indicated three levels of understanding in each of the six sections: Nature/purpose of science and scientific ideas; nature of a hypothesis; nature/purpose of an experiment; guiding ideas and questions; results and evaluation; relationship. In each section, three levels of competency were characterized.

Nature/purpose of science and scientific ideas. In *level 1* answers focus on activities themselves: scientists have ideas about how to carry out these activities. In *level 2* answers focus on the development of a mechanistic understanding of the world: scientists have ideas, questions and predictions about how things work and predictions about the outcomes of experiments. In *level 3* answers focus on the development of an explanatory understanding of the world.

Nature of a hypothesis. In *level 1* a hypothesis is an idea or a guess. In *level 2* a hypothesis is also an idea or a guess but it is clearly related to an experiment or a phenomenon and it is explicitly something that can be tested. In *level 3* the hypothesis is not only related to an experiment, but aids in interpreting the results of an experiment and is evaluated and developed in terms of the results.

The nature/purpose of an experiment. In *level 1* there is no clear distinction between experiments and ideas (“a scientist tries something to see if it works or reacts”). In *level 2* the distinction between the idea and activity is clear (“the experiment is a test of a scientist’s idea, scientists do experiments to test to see if their idea is right”). In *level 3* the distinction is the same as level 2 but additionally the relationship between the results and the idea being tested is clearly articulated.

Guiding ideas and questions. In *level 1* answers focus on activities such as thinking, observing, exploring and the goal of these activities is to gather information. In *level 2* an exploration is guided by a particular idea, question and object of phenomenon. In *level 3* answers which guided exploration in level 2 are elaborated to include reflection on prior knowledge and experience, or there is an understanding of evaluation and development of ideas.

Results and evaluation. In *level 1*, if the outcomes of the experiment are unexpected, the answer is that something is not working and should be checked or changed; the “something is not clearly specified” as an idea. In *level 2* idea and experiment are clearly distinguished. In *level 3* there is an understanding that an idea is to be modified because of a conflict between the idea itself and experimental outcomes or other evidence and the modified idea takes these data into account.

Relationship. In *level 1* there is still no clear distinction between ideas and experiments (“a scientist tries out an idea to see if it works”). In *level 2* there is a clear distinction

between ideas and experiments (“the idea is tested, to see if it is right, or the idea is used to predict the outcome of an experiment”). In *level 3* it is clear that ideas are tested in experiments to include the understanding that they are evaluated or developed in accordance with the results of the tests. Carey found out that after her lessons many students clearly understood that inquiry is guided by particular ideas and questions and that experiments are tests of ideas.

Tamir (1989) investigated the levels of competency of subjects in doing experiments. He asked subjects to do experiments by using some of the prepared materials. Furthermore he asked them some questions about experiments, for example, what problem they are going to investigate, what hypothesis they tend to test and how to design an experiment to test a hypothesis. Afterwards, he assessed the levels of students in reporting and analysing the results.

Lowest level: pupils report on the results of one simple experiment using units of measurement specified by the teacher.

Medium level: pupils report on the results of several treatments and replications of the experiment, choose themselves the units of measurement and justify their choice.

High level: pupils report on the results of several treatments and replications, determine not only the units of measurement but also the most efficient and visually expressive organization and presentation of the complex results.

Similarly, different difficulties can be worked out for other phases of the investigation (problem and hypothesis formulation, experiment design, etc.).

3. Research questions

In the international scientific literacy test for PISA 2000, five levels of competencies were proposed. Items were mapped onto the levels of competencies by dividing up the maximum total sum score into five segments and by assigning items with a low/high difficulty to a low/high level of competency.

Carey (1989) investigated levels of understanding about the nature of scientific knowledge and inquiry in young students by interviews. She identified three general levels of response. The levels were determined based on making the distinction between ideas and activities and understanding the motivation for experimentation.

This study presents an approach to measuring achievement of students and focuses on levels of competencies in experimentation. For this, a paper and pencil test is developed, but it differs from the PISA test and Carey's investigation, the answers of this test are closed-end items, with response categories that can be directly related to specific levels of competency. By using this test it shall be analysed if a paper and pencil test with the closed-end items can be used to assess levels of competencies in experimentation (*research question 1*).

The test items are designed to assess the levels of students' competency in all three dimensions of experimentation (search in the hypothesis space, search in the experiment space and data analysis), each of which forms a single dimension and will be crossed with other dimensions. This design of the test may allow looking at how the three dimensions in experimentation interact (*research question 2*).

Besides, a content knowledge test is also developed in order to assess the pre-knowledge of students about the content corresponding to experiments. Because many researchers asserted that pre-knowledge affects the processes of experimentation (Chi, Feltovich, & Glaser, 1981; Lord, Ross & Lepper, 1979; Kuhn, Amsel & O'Loughlin, 1988), both a knowledge test and an experiment test are used to look at which correlations can be found between biological content knowledge and levels of competencies in experimentation (*research question 3*).

4. Hypotheses

In PISA students are asked to answer the questions in writing, then the investigators sum the total points and map the points on a scale with five levels of competency. Some researchers assess levels of understanding of students in experiments by interview (Carey, 1989), or they prepare the materials for an experiment and ask some questions about the hypothesis and then ask the students to use the materials to design the experiment (Tamir, 1989). In this study a paper-and-pencil test with closed-end items, with response categories that can be directly related to a specific levels of competency is developed too. The test includes tasks related to all three processes of experimentation. This test can presumably be used in order to assess levels of competency of students in three dimensions of experimentation.

However, both structures of knowledge and strategies of experimentation are fundamental to scientific reasoning (Schauble, Glaser, Raghavan and Reiner, 1991). Thus, exploring the relations between the pre-knowledge of students and competencies in experimentation was done by many researchers.

Klahr and Dunbar (1988) in SDDS implicated that subjects have to use two strategies to formulate hypothesis, one is to search memory and the other is to generalize from the results of previous experiments, thus prior knowledge of subjects plays an important role and influences directly formulating the hypothesis. In the study of participants working on a computer-controlled robot tank (called BigTrak tank), Klahr (2000) indicated that the three categories of prior knowledge that may influence participants' hypotheses are linguistic knowledge about the meaning of "Repeat", programming knowledge about Iteration and specific knowledge about BigTrak. Besides, in the investigation about the difference between adults and children in formulation and evaluation of hypotheses Carey (1985a) showed that the most important differentiating factor is the adequacy of domain-specific knowledge. Adults can solve the task better than children because they have good specific knowledge in the area.

Hypothesis one is as follows:

If students possess good prior knowledge, they can also gain a high level in the search in the hypothesis space.

In the search in the experiment space, many researchers indicated the influence of prior knowledge in designing and carrying out the experiment. For example, older children have better ability in choosing a tool to do an experiment than younger children (Brown, 1990). However, some studies indicated that it is not the abilities of controlling variables and designing an experiment that are influenced by prior knowledge but the strategy of doing experiments. Lawson and Wollman (1976) investigated students' ability to control variables, to design and evaluate experiments and they indicated that these abilities of students were influenced not only by difference in grades, but by their knowledge about content (familiar or unfamiliar knowledge), by training that helped students to have the knowledge about strategy of doing experiments. In addition, Chen and Klahr (1999) who studied students' ability in "Control of Variables Strategy" (CVS) found that the effects of domain-specific knowledge play a role in CVS. Besides, children in the Training Probe condition can solve the task better and improve their domain-specific knowledge, whereas children in other conditions did not, this concerns the training strategy in doing experiment. Tschirgi (1980) investigated the differences in reasoning between adults and children in second, fourth and sixth grades. He showed that the domain-specific knowledge influenced the ability of people to design an experiment. Thus, hypothesis two is as follows:

If students have good content knowledge, they gain high levels in the search in the experiment space.

However, there are many ideas about the relationship between prior knowledge and evaluation evidence. In evaluating the evidence process, it was discovered that children usually seek the evidence consistent with their belief and ignore the disconfirmed evidence (Dunbar & Klahr, 1989; Schauble, Glaser 1990; Champagne, Gunstone, Klopfer, 1985). In other words, students usually show the attempt in the data analysis to confirm the positioned initial hypothesis ("confirmation bias") (Wason, 1960), even if the data require an alteration or a distortion of the hypothesis. That means, they frequently evaluate evidence based on their initial belief. Thus, they sometimes show an inadequate analysis of experimental data, in particular, when the data of the experiment are unexpected. In this case, the cause is not the lack of logical thinking of the students, but the tendency is to confirm proven concepts in everyday life (Mayer, 1999). Also, expectations can be confirmed with the data analysis, although the burden of proof from experiments contradicts (Schauble, Klopfer, Raghavan, 1991; Chinn, Brewer, 1998; and Metz, 1998).

Another reaction to unexpected results is the conclusion that clear statements are not yet possible due to the available data situation (Schauble, Klopfer, Raghavan, 1991).

In brief, the data analysis is influenced deeply by prior knowledge of subjects, however this effect is not always positive. If their belief is wrong, the result is the opposite. Therefore, hypothesis three is as follows:

If students have good pre-knowledge about the area, they can also gain high levels in data analysis.

Pre-knowledge influences all three processes of experimentation. However, the level of effect is different in different dimensions: If students possess good prior knowledge, they can reach high levels in data analysis and in the search in the hypothesis space. For the latter depends on their knowledge about methodology or strategy in doing experiments.

Beside the effects of prior knowledge to levels of competency in experimentation, three processes of experimentation themselves can be correlated.

Dunbar and Klahr (1989) indicated that “search in the hypothesis space is guided both by prior knowledge and by experimental results and search in the experiment space may be guided by the current hypothesis and it may be used to generate information to formulate hypotheses”. Simultaneously, as we argued above, students can gain high levels of competency in both search in the hypothesis and data analysis if they have good content knowledge; thus, it is possible, that data analysis and search in the hypothesis relate closely. So hypothesis four is as follows:

If student have high levels of competency in search in the hypothesis space, they might gain high levels in data analysis and vice versa.

However, the correlations between search in the hypothesis space and search in the experiment space and between data analysis and search in the experiment space are not specific. Because to attain high levels of competency in search in the experiment space not only depend on good content knowledge, but also on good skills of doing experiments. Thus, you can not be sure that students who achieve the high levels in search in hypothesis space and data analysis gain high levels in search in the experiment.

In contrast, if students possess good knowledge about methodology in doing experiments they can achieve high levels in all three dimensions of experimentation search in the experiment space, data analysis and search in the hypothesis space.

Hypothesis five is then as follows:

If they gain high levels in the search in the experiment space, they will achieve high levels in both dimensions “search in the hypothesis space” and “data analysis”.

In sum, we expected high correlations between pre-knowledge and the three dimensions in experimentation, especially, high correlation between pre-knowledge and the dimensions search in the hypothesis space and data analysis. On the other hand, in the relationships between the three dimensions of experimentation, we also expected high correlations in all three combinations (search in the hypothesis space * data analysis; search in the hypothesis space * search in the experiment space; and data analysis * search in the experiment space), especially, as we assumed that there is higher correlation between the dimensions “search in the hypothesis space” and “data analysis” than between the dimensions “search in the hypothesis space” and “search in the experiment space”. We also assumed that there are higher correlations between the dimensions “search in the hypothesis space” and “data analysis” than between the dimensions “data analysis” and “search in the experiment space”. These hypotheses were based on the assumption that the dimensions “search in the hypothesis space” and “data analysis” are driven by the students’ pre-knowledge about the science contents of the experiment while the dimension “search in the experiment space” should prove more dependent on the students’ methodological knowledge.

Part II

Empirical part

Chapter 2: Cognitive Laboratory

1. Method

This study focuses on the development of two independent paper-and-pencil tests for assessing students' content knowledge and levels of competencies in experimentation. The competency test has three dimensions: "search in the hypothesis space", "search in the experiment space" and "data analysis". Prior to item development for each of these dimensions, levels of competencies had been described which are based on empirical evidence (Hammann 2004). These levels of competencies take into consideration major empirical findings concerning student performance in tasks that involve forming hypotheses, planning experiments and analysing data. The answering format consists of simple multiple-choice with four options which can be directly related to specific levels of competency. If students solve the item correctly, they will gain level 2 (full credit). If they choose the incorrect option they will achieve level 0 (no credit) and if they choose one of the two partly correct options they will obtain level 1 (partial credit).

Our test items were designed to test the students' knowledge about the method of designing controlled experiments. In order to test a cause-and-effect hypothesis, an experiment must investigate if, for example, a phenomenon occurs when a specific factor is present and that the phenomenon does not occur in the absence of the factor. In controlled experiments, the results obtained from an experimental test are compared with the results obtained from an experimental control which differs from the experimental test in exactly the factor whose effect is being tested. In the sciences, the word "variable" is used to refer to a measurable factor or characteristics. In a scientific experiment, so called "independent variables" are factors that can be altered or chosen by the scientist. For example, temperature is a common environmental factor that can be controlled in laboratory experiments. "Dependent variables" or "response variables" are those that are measured and collected as data. An independent variable is presumed to affect a dependent one. Most often, tests are done in duplicate or triplicate in order to rule out chance effects and measurement errors.

The test developed for this study consists of texts and pictures that depict experimental designs. The test has three dimensions. In the dimension "search in the hypothesis space", the students are asked to identify the hypothesis that can be tested with the experiment depicted. In the experiment, there are four or five variables, one or two of

which was/were tested. Four options are given, but only one of them is correct, because the hypothesis stated relates to the experiment depicted. In the other options, only one hypothesis, but not the other is correctly identified (intermediate level) or no hypothesis is identified at all (lowest level).

Sample Item: A student named Jan did an experiment about seed germination. He used two pots with soil (pot 1 and pot 2) and one pot with cotton wool instead of soil (pot 3). He sowed bean seeds in the three pots and put all three pots in the sunlight at a temperature of 22°C. He watered pot 1 and pot 3, but he did not water pot 2.

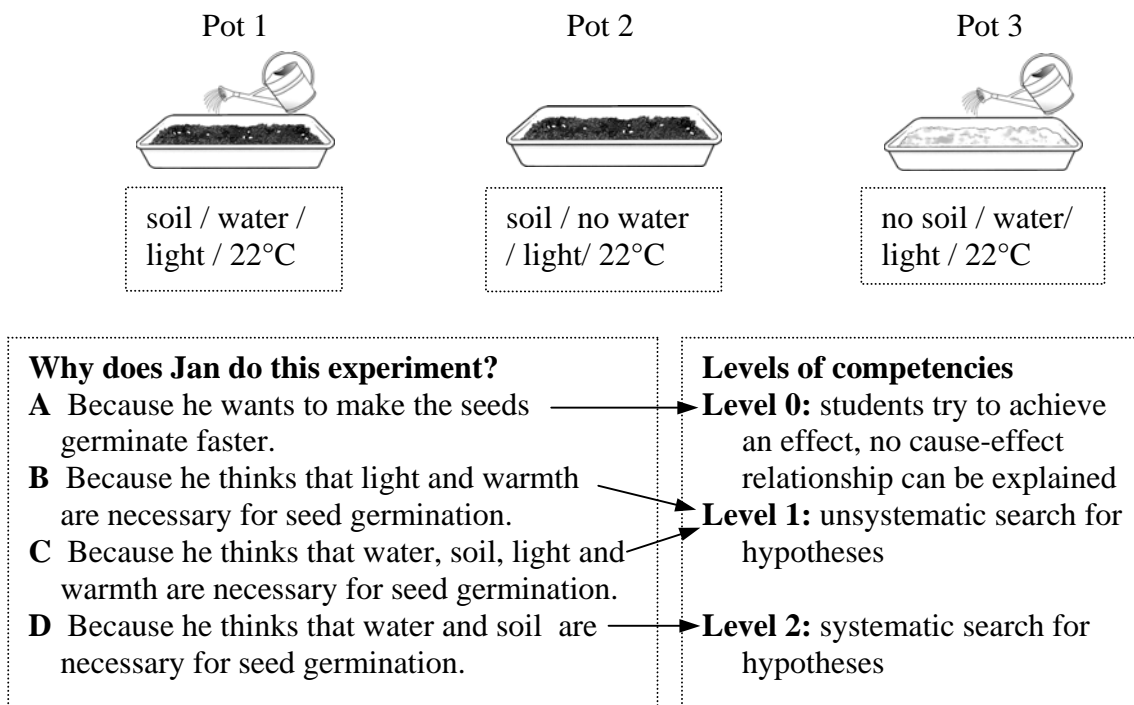


Figure 2.1: Sample item: Search in the hypothesis space (Version 2)

In the dimension “data analysis”, the findings of the same experiment are presented to the students. Four explanations are given, one of which is correct because it relates to the tested hypothesis and correctly explains the cause-and-effect relationship between the variable that has been tested and the findings. One of the four explanations does not relate to the hypothesis at all (lowest level). The other two explanations relate to a specific hypothesis, but not to the hypothesis tested (intermediate level). For the intermediate level, it is also possible that the answer relates only partly to the tested hypotheses or that the explanation is not based on the experimental findings but relies on the pre-knowledge.

Sample: After a few days Jan obtained the following results: The seeds in pot 1 and pot 3 germinated, whereas they did not germinate in pot 2.




<p>Pot 1</p>  <div style="border: 1px dashed black; padding: 5px; width: fit-content; margin: 0 auto;"> soil / water / light / 22°C </div>	<p>Pot 2</p>  <div style="border: 1px dashed black; padding: 5px; width: fit-content; margin: 0 auto;"> soil / no water / light / 22°C </div>	<p>Pot 3</p>  <div style="border: 1px dashed black; padding: 5px; width: fit-content; margin: 0 auto;"> no soil / water / light / 22°C </div>
<p>Which one is the best explanation of the findings?</p> <p>A The experiment did not work because the seeds in pot 2 did not germinate</p> <p>B The experiment showed that seeds need light and warmth to germinate</p> <p>C The experiment showed that seeds need soil and water to germinate</p> <p>D The experiment showed that seeds need no soil, but water to germinate</p>		<p>Levels of competencies</p> <p>Level 0: data are not related to the hypothesis tested</p> <p>Level 1: data are related to a hypothesis, but not to the hypothesis tested</p> <p>Level 2: data are related to the hypothesis tested</p>

Figure 2.2: Sample item: Data analysis (Version 2)

In the dimension “search in the experiment space”, the hypothesis is clearly stated, but the experimental design presented in the item is incomplete. The students are asked to complete the experimental design by choosing an experiment that can be compared to the one already described. Again, the students have to choose among four options, only one of which is fully correct because the test variable is changed and the others are kept constant. One of the four experiments is completely incorrect because in this experiment either all variables (the test variable and all other variables) are changed or no variable is changed (lowest level). This option was provided because students were found to show a strategy called “change all” (Tschirgi 1980). For the intermediate level, either the test variable and one other variable are changed, so that the students confused the two different types of variables, or not all variables that need to be kept constant are controlled so that the experiment is confounded.

Sample Item: Jan thinks **bean seeds will germinate faster if they are kept in a warm place.**

He plans an experiment to test this idea.

This is Jan’s experiment. He puts the bean seeds in a pot with soil (pot 1), waters them and keeps the pot at a temperature of 22°C in the sunlight.

Jan needs another pot in order to compare with pot 1

Which one in the following pots (A-D) should he do?

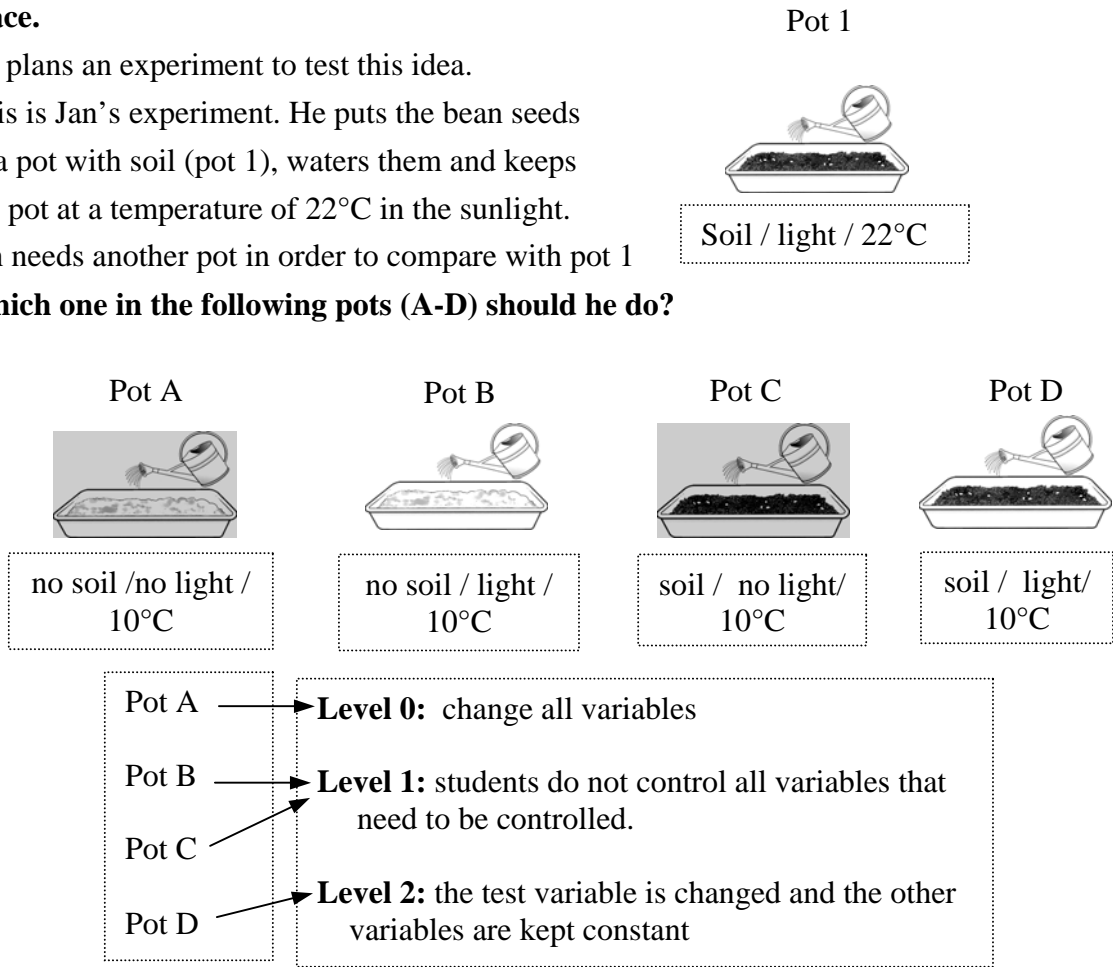


Figure 2.3: Sample item: Search in the experiment space (Version 2)

In the competency test, we designed two versions that we expected to have different levels of difficulty. The items presented above are examples of the more difficult version 2. In version 1, for example, the students were asked to compare two pots with seedlings – not three (see appendix, p. 206-208). The aim of developing two versions was to test which version is more appropriate for fifth and sixth grade students.

In version 1, only one variable is tested. In version 2, the experimental design is more complex because two variables are tested.

However, the tasks in “Search in the experiment space” were designed similarly in the two versions.

Time and aim of the cognitive laboratory

In May 2004, we performed the cognitive laboratory in a comprehensive school (Gesamtschule) in Schleswig-Holstein.

The idea of the cognitive laboratory, typically, is to receive qualitative feedback to the items. Thus, the major aim of this study was to investigate the students' responses to the units in the test by asking them questions like: Did you understand what you were expected to do? Did you understand the text? Did you find the pictures helpful? Did you think that the test was easy or difficult? Item profiles convey a general idea of individual responses to the test.

Sample

The sample consisted of six students – three girls and three boys – who came from the fifth grade (aged from 11 to 12) at the Klaus-Groth school in Tornesch, Germany, and volunteered to take part in the study.

Design

The term “unit” refers to a sequence of items that are thematically linked. The five units used in this study are: Unit 1: Seed germination, Unit 2: Chicken eggs, Unit 3: Potatoes, Unit 4: Baking bread and Unit 5: The growth of bean plants. Each unit exists in two versions. In version 1, each unit contains 5 items: two items (items 1 and 2) belong to the dimension “search in the hypothesis space”, two items (items 3 and 4) belong to the dimension “data analysis” and one item belongs to the dimension “search in the experiment space” (item 5). An exception is the Unit “Baking bread” which consists of four items, one item for the dimension “search in the hypothesis space”, one item for the dimension “data analysis” and two items for the dimension “search in the experiment space”. Therefore, the test booklet for version 1 – all items combined – contained 24 items.

In version 2, each unit has three items corresponding to three dimensions of experimentation, except for unit 2 “Chicken eggs” which has four items because there are two items in the dimension “search in the experiment space”. The test booklet for version 2 consists of 16 items.

The six students were randomly divided into two groups. One group did version 1 and the other group did version 2. The time provided was 45 minutes for each version. Students who finished the test before that time were asked to raise their hand so that the exact testing time could be recorded.

Answering format

In this test, we used simple multiple-choice items. Each item had four choices (A,B,C,D), with one correct answer, two partly correct options and an incorrect answer. The students had to choose one of the four options.

Coding the answer

We used a 0/1/2 scoring model to code the answers. This model describes two levels of competencies. If the students choose the correct option, they will gain value 2 (level 2), if they choose the incorrect answer they will gain value 0 (level 0) and if they choose a partly correct one they will achieve value 1 (level 1).

2. Findings

2.1. Testing time

The first student finished the test after 13 minutes and all six students finished completely after 25 minutes. Thus, the average time for each student for version 1 was 22 minutes (24 items), that means, the students need an average of 55 seconds to solve each item.

The average time for each student to solve version 2 was 18 minutes (16 items). This means that for each item in this version the students needed an average of 1.12 minutes.

2.2. Item profiles

Item profiles can be used to investigate whether the students consistently answered the items at a specific level. Item profiles give a general idea about the students' responses to the items. Trustworthy insights into item characteristics, however, can be gained only if a larger sample is considered and if the Cronbach's alpha is calculated. This is typically done after the cognitive laboratory in a qualitative pre-study of the test (cf. Chapter 4-6).

Based on students' answers for each item (see appendix, tables 1-3) in all three dimensions in experimentation, we had the following findings:

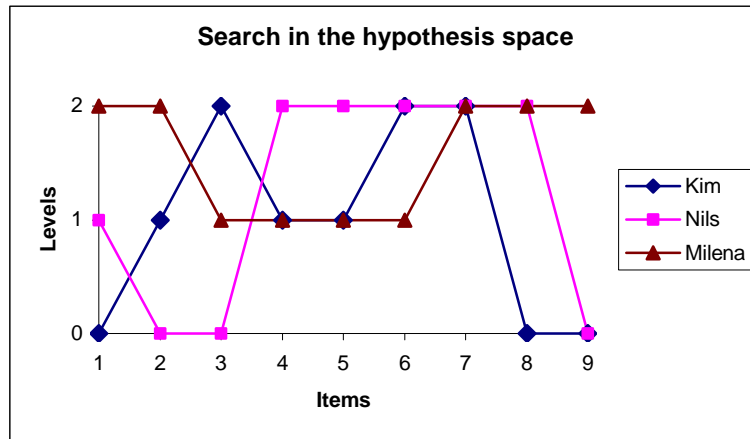
Version 1**Search in the hypothesis space**

Figure 2.4: Item profiles for items in the dimension “search in the hypothesis space” (Version 1) (Item 1 and 2 belonged to unit 1; item 3 and 4 to unit 2; item 5 and 6 to unit 3; item 7 to unit 4; item 8 and 9 to unit 5)

Figure 2.4 shows that in the dimension “search in the hypothesis space” Kim solved three items out of nine, while Nils and Milena fully solved five items each. The item profiles give further insights into the differences between Nils and Milena. Milena reached level 1 for the items that were not completely solved, whereas Nils reached level 1 for item 1 and level 0 for the items 2, 3 and 9. In this dimension Milena performed best, as can be seen in the items not fully solved that she consistently answered at an intermediate level.

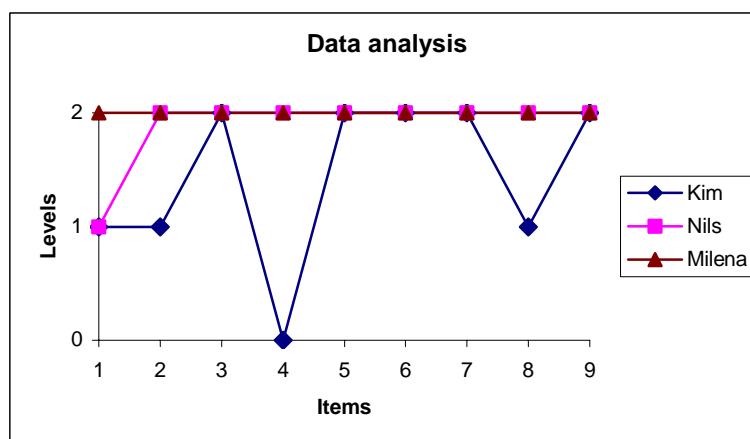
Data analysis

Figure 2.5: Item profiles for items in the dimension “data analysis” (Version 1) (Item 1 and 2 belonged to unit 1; item 3 and 4 to unit 2; item 5 and 6 to unit 3; item 7 to unit 4; item 8 and 9 to unit 5)

In “data analysis”, Milena answered all items correctly and Nils solved eight items out of nine. In contrast, Kim solved 5 items and gained level 0 for item 4 and level 1 for items 1, 2 and 8.

Thus, in this dimension, Milena was also the best student and Kim did all the tasks more poorly than two others.

Search in the experiment space

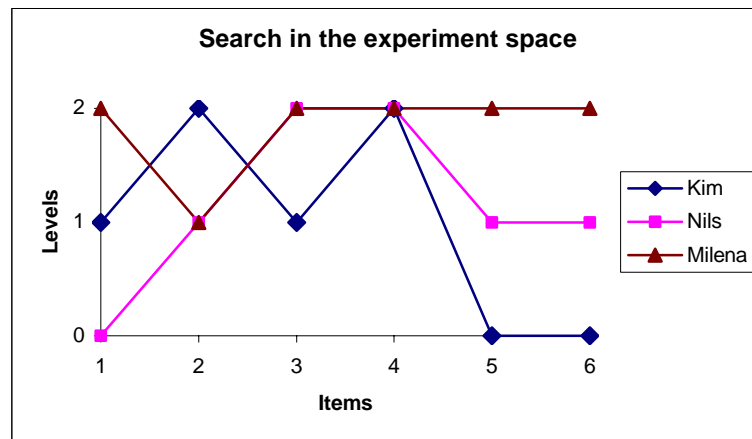


Figure 2.6: Item profiles for items in the dimension “search in the experiment space” (Version 1) (Item 1 belonged to unit 1, items 2 and 3 to unit 2, item 4 to unit 3, item 5 to unit 4 and item 6 to unit 5)

In this dimension, Milena reached level 2 for four items out of five, while Kim and Nils only answered two items at level 2. Moreover, Nils did not solve item 1 and Kim did not solve items 5 and 6. Their item profiles oscillated frequently, especially Kim’s item profile, as Kim received full credit for two items, partial credit for two items and no credit for two items.

Discussion

Generalizing comments cannot be made since only three students responded to the items in competency test version 1. However, the item profiles for the three item types reveal that the items were clearly within the students’ reach. The students were able to solve most items, even though sometimes only the intermediary level was reached (partial credit). The item type that presented the least challenge to the students, apparently, was “analysing data”, where most items were solved by the three students and where the three students’ item profiles also resemble each other the most with little fluctuation between the levels. This can be interpreted – with a lot of caution because this is only qualitative data – as an indication that there may be problems with the

items' power of discrimination. Concerning the other two item types, there are differences both between students who were able – respectively unable – to solve the items as well as differences between the items that proved to be sometimes easy and sometimes difficult for the same student, although the items belong to the same dimension. Items with oscillating item profiles need to be carefully analysed in order to ensure that it is not the text or the pictures, but the science contents of the item that presents the real challenge.

Version 2

Search in the hypothesis space

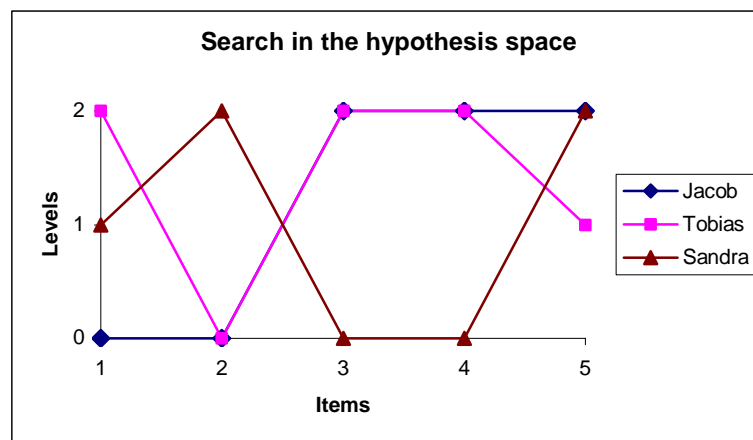


Figure 2.7: Item profile in the dimension “search in the hypothesis space” (Version 2)
(Each item belonged to one unit)

Figure 2.7 shows that in the dimension “search in the hypothesis space” Jacob and Tobias solved three items out of five, while Sandra fully solved two items. The item profiles give further insights into differences between Jacob and Tobias. Tobias gained level 1 for item 5 and level 0 for the item 2, whereas Jacob gained level 0 for both items (2 and 5). In this dimension, Tobias did the best and Sandra did worse than two others.

Data analysis

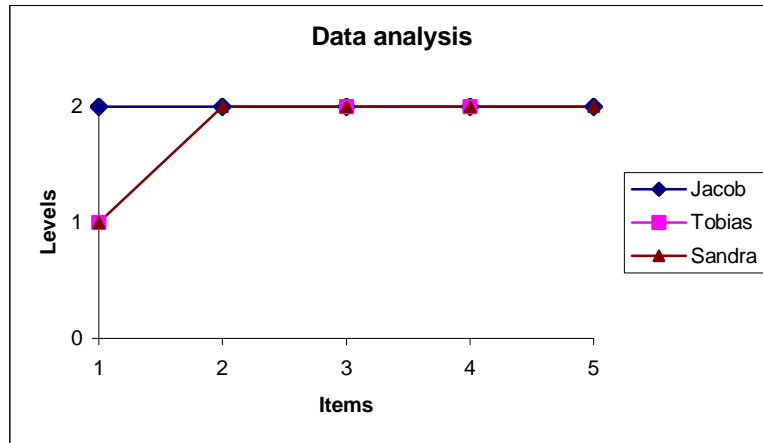


Figure 2.8: Item profile in the dimension “data analysis” (Version 2)
(Each item belonged to one unit)

In “data analysis”, Jacob answered all items correctly, Sandra solved four items out of five and item 1 was solved for partial credit. Tobias solved two items, he gained level 1 for item 1 and did not do items 2 and 5 (missing).

So, in this dimension, Jacob was the best student and Tobias did worse than Jacob and Sandra.

Search in the experiment space

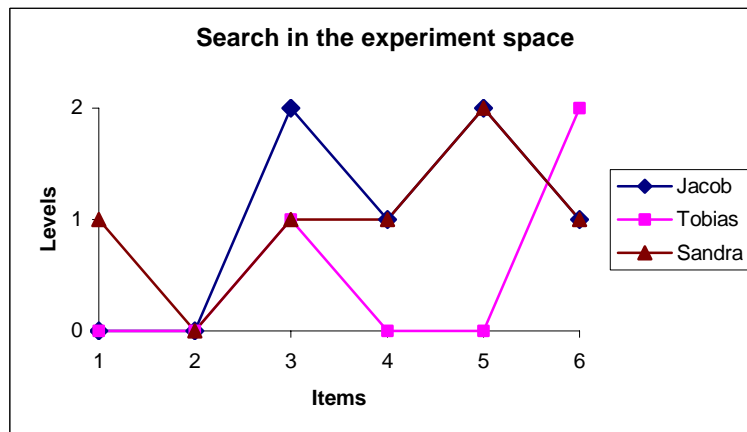


Figure 2.9: Item profile in the dimension “search in the experiment space” (Version 2)
(Item 1 belonged to unit 1; items 2 and 3 were from unit 2; item 4 from unit 3; item 5 from unit 4 and item 6 from unit 5)

In this dimension, Jacob gained level 2 for two items out of six, while Tobias and Sandra only answered one item at level 2. Moreover, Tobias did not solve 4 items and Sandra did not solve item 2. Their item profiles oscillated frequently, especially Jacob’s

item profile, as Jacob received full credit for two items, partial credit for two items and no credit for two items.

Discussion

The item profiles in version 2 reveal that the items in the two dimensions “search in the hypothesis space” and “data analysis” were clearly within the students’ reach. As well as in version 1, in the dimensions “data analysis” most items were solved by three students. However, in the dimension “search in the experiment space” most items were not fully solved; item 2 was incorrectly answered by all three students. In this version, it was not clear enough to differentiate two types of students, because one student did well in one dimension but in the others he/she did incorrectly. Item profiles also oscillated frequently, especially in “search in the experiment space”.

2.3. Mean level of competency of each student in each dimension of experimentation

Based on the students’ answers , we calculated the mean level of competency for each student in each dimension and drew the graph that allowed us to compare the levels of competency of students in three dimensions of experimentation.

Version 1

Name of student	Search in the hypothesis space	Data analysis	Search in the experiment space
Nils	1.22	1.78	1.6
Kim	1.0	1.44	1.0
Milena	1.56	2.0	1.83

Table 2.1: Mean level of competency of each student in the three dimensions of experimentation (Version 1)

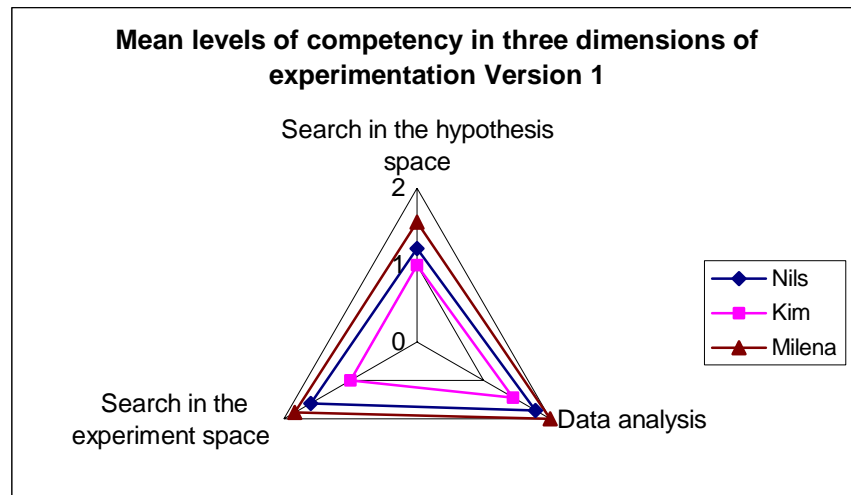


Figure 2.10: Mean levels of competency each student attained in the three dimensions of experimentation (Version 1)

Figure 2.10 shows that in version 1 Milena was the best student in solving all three dimensions in experimentation, she reached higher levels of competency in experimentation than Nils and Kim and Nils did better than Kim in all three dimensions. So, it might be said that if a student achieves a high level in a dimension of experimentation, she/he can also reach the high levels in the others.

Version 2

Name of student	Search in the hypothesis space	Data analysis	Search in the experiment space
Tobias	1.4	1.25	0.80
Jakob	1.2	2.0	1.0
Sandra	1.0	1.80	1.40

Table 2.2: Mean level of competency of each student in the three dimensions of experimentation (Version 2)

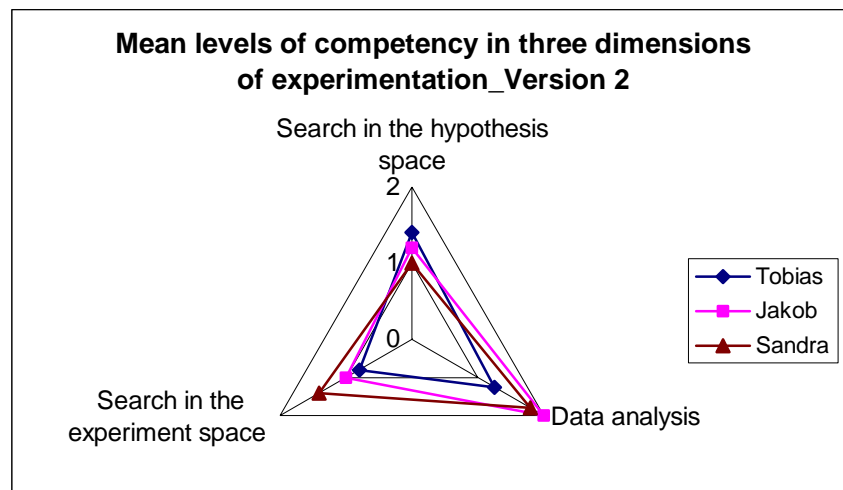


Figure 2.11: The mean levels of competency each student attained in the three dimensions of experimentation (Version 2)

It was different than in version 1. In this version, figure 2.11 shows that in “search in the hypothesis space” Tobias gained higher level of competency. However, in two other dimensions he attained a lower level than two others. On the other hand, in “data analysis” Jacob answered the best and in “search in the experiment space” Sandra reached the highest level.

2.4. Interview

After the paper-and-pencil test, we interviewed two volunteer students. We asked them about the level of difficulty of the items in the three dimensions and the two versions. The result was that students said that some units were easier than the others, for example, unit 4 “Baking bread” and unit 5 “The growth of plants” were familiar to students. They knew yeast is very important in baking bread, or they had learned that plants need water, light, warmth...etc. to grow. Thus, they did these units better than the others. It was also said that the items in “data analysis” were too easy and the items in “search in the experiment space” were difficult.

3. Conclusion

The students understood how to do the test and they did it quickly. About the content of the test, most items in “search in the hypothesis space” had the medium level of difficulty, since most items were correctly answered by one or two of three students. However, most of the tasks in the dimension “data analysis” in both versions were very easy, many items in this dimension were correctly answered by all three students. The reason for this may be the text was so clear, or the pictures were not good, or some units and some items were familiar to students, while the others were unfamiliar. So, these items must to be fixed to be more complex, we must change the text of answers or replace them in another way. On the other hand, most items in “search in the experiment space” were too difficult, especially in version 2, where in some units no students answered correctly. Therefore, for the items in this dimension we must make them easier.

On the other hand, we needed to develop four or five more units. Then we would have 9 or 10 units, each unit containing two versions and each unit in version 1 having six items and each unit in version 2 having three items that corresponded to the three dimensions of experimentation. We also designed a knowledge test to test the content knowledge of students.

Chapter 3: Pre-test 1

Introduction

In February 2005, the first pre-test was done. For this, we used two types of tests, a knowledge test and a competency test. We expected two question systems to be reliable enough to assess the levels of competencies of students in experimentation and the students' biological pre-knowledge to analyse the relationships between the students' pre-knowledge and the three dimensions in experimentation as well as the correlations within these three dimensions.

On the basis of the cognitive laboratory, we expected version 1 of the competency test to be more appropriate for students in fifth and sixth grades than version 2 in terms of item difficulty. Further, we investigated whether two different scoring models can be used to assess students' competencies in experimentation. One scoring model (full credit/no credit) is based on the assumption that there is no intermediate level; the other (partial credit) assumes the contrary.

1. Method

Sample

The participants of this study were 799 fifth and sixth grade students from 22 schools (four different types of school: Gymnasium, Hauptschule, Gesamtschule and Realschule) in Germany. In each school, one to four classes were randomly chosen and the participants took part in the test voluntarily. Their age ranged from 10 years and 1 month to 14 years and 3 months and the mean age for the sample was 12 years and 1 month.

School	Grade	Number of students	Booklet
Hans-Multscher-Gymnasium	5	14	111
Heinrich-Nordhoff-Gesamtschule	6	29	111
Von-Galen-Schule	6	30	111
Gesamtschule Schermbeck	6	29	121
Lise Meitner Gesamtschule -Köln	5, 6	55	121, 221
Ratsgymnasium	5, 6	78	121,131, 221
Elly-Heuss-Knapp-Realschule	5, 6	50	131,211
IGS Franzches Feld	6	13	131
Wilhelm-Bracke-Gesamtschule	6	26	131
Ernst-Barlach-Gymnasium	5, 6	53	211, 221
Ferdinand-Steinbeis-Realschule	5	51	211, 221
Geschwister-Scholl-Gesamtschule	6	105	211, 221, 231, 241
Haupt- und Realschule Meiendo	6	27	211
Oberwaldschule Hauptschule	5	20	221
Oswald-von-Nell-Schule	5	19	221
Gesamtschule			
Hermannsburg- Gesamtschule	5	27	221
Giordano-Bruno-Gesamtschule	5, 6	52	221, 241
Georg-Christoph-Lichtenberg-Gesamtschule	6	25	231
IGS Peine-Vöhrum	5	27	231
IGS Querum	5	18	231
IGS Franzches Feld	6	26	241
Max-Planck-Gymnasium	5	25	241

Table 3.1: The schools and the number of students took part in the test

Design

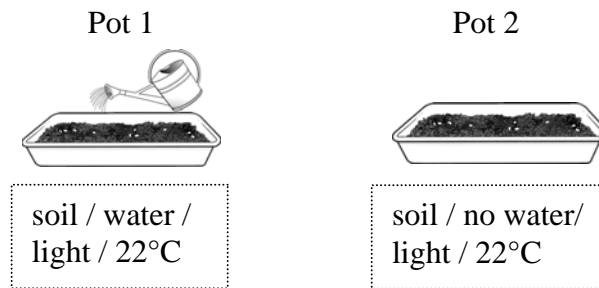
Each student took two types of test, the competency test and the knowledge test.

In the competency test, we used simple multiple choice questions as described in Chapter 1 (Cognitive laboratory). We also had two versions of this test. Version 2 was more difficult than version 1. In most items of version 1, only one hypothesis was tested, whereas in version 2, two hypotheses were tested.

The following two sample items illustrate the difference in complexity between the two versions, as version 1 requires comparing two tests and version 2 requires comparing three tests.

Sample Item: Search in the hypothesis space in version 1

A student named Jan did an experiment about seed germination. He used two pots with soil (pot 1 and pot 2). He sowed bean seeds in the two pots and put them in the sunlight at a temperature of 22°C. He watered pot 1, but he did not water pot 2.



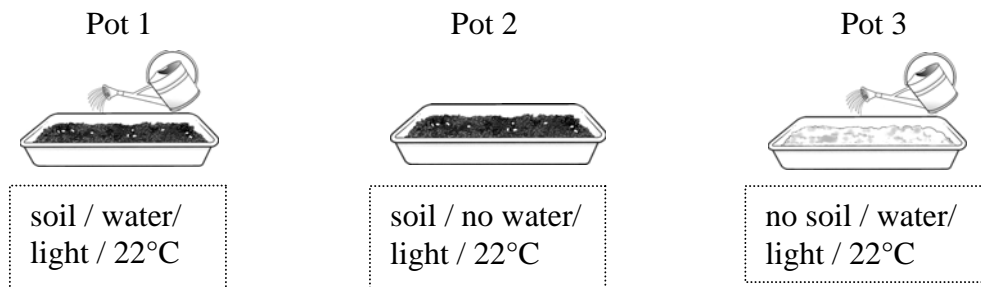
Why does Jan do this experiment?

- A Because he wants to make the seeds germinate faster.
- B Because he thinks that light and warmth are necessary for seed germination.
- C Because he thinks that water, soil, light and warmth are necessary for seed germination.
- D Because he thinks that water is necessary for seed germination.

Figure 2.1: Sample item: Search in the hypothesis space (Version 1)

Sample Item: Search in the hypothesis space in version 2

A student named Jan did an experiment about seed germination. He used two pots with soil (pot 1 and pot 2) and one pot with cotton wool instead of soil (pot 3). He sowed bean seeds in the three pots and put all three pots in the sunlight at a temperature of 22°C. He watered pot 1 and pot 3, but he did not water pot 2.



Why does Jan do this experiment?

- A Because he wants to make the seeds germinate faster.
- B Because he thinks that light and warmth are necessary for seed germination.
- C Because he thinks that water, soil, light and warmth are necessary for seed germination.
- D Because he thinks that water and soil are necessary for seed germination.

Figure 2.2: Sample item: Search in the hypothesis space (Version 2)

Besides, a knowledge test was also developed in order to assess the students' knowledge of the biological content of the experiment, for example, knowledge about seed germination. In the knowledge test, complex multiple-choice questions were used. Each unit of the knowledge test had five to seven items, each item consisted of four to six sub-questions.

What is contained in a seed?	Yes or no?
Nutrient	yes / no
Blooms	yes / no
Plant embryo	yes / no
Small seeds	yes / no
Water	yes / no
Parts of plant	yes / no

Figure 3.3: Knowledge test: Seed germination

Since the knowledge test does not exist in two versions, it was identical for students who took the competency test versions 1 and 2.

Nine units with different biological content knowledge were used. These units were: Unit 1: Seed germination; Unit 2: Chicken eggs; Unit 3: Apple wine; Unit 4: Baking bread; Unit 5: The growth of bean plants; Unit 6: Potatoes; Unit 7: Heart beat; Unit 8: Plant growth and Unit 9: Fish respiration.

The units were divided into seven booklets. Three booklets belonged to version 1 (booklets 111, 121 and 131) and four booklets belonged to version 2 (booklet 211, 221, 231 and 241). The unit "Seed germination" was considered to be the anchor unit and it was used in six of the seven booklets. The other eight units were distributed by chance among the seven booklets.

	Unit	Knowledge test	Competency test		
			Search in the hypothesis space	Data analysis	Search in the experiment space
Version 1	Seed germination	6 items	2 items	2 items	2 items
	Chicken eggs	7 items	2 items	2 items	2 items
	Apple wine	6 items	2 items	2 items	2 items
	Baking bread	5 items	2 items	2 items	2 items
	Bean plants	5 items	2 items	2 items	2 items
	Potatoes	6 items	2 items	2 items	2 items
	Heart beat	6 items	2 items	2 items	2 items
	Plant growth	5 items	2 items	2 items	2 items
	Fish respiration	5 items	2 items	2 items	2 items
Version 2	Seed germination	6 items	1 item	1 item	1 item
	Chicken eggs	7 items	1 item	1 item	1 item
	Apple wine	6 items	1 item	1 item	1 item
	Baking bread	5 items	1 item	1 item	1 item
	Bean plants	5 items	1 item	1 item	1 item
	Potatoes	6 items	1 item	1 item	1 item
	Heart beat	6 items	1 item	1 item	1 item
	Plant growth	5 items	1 item	1 item	1 item
	Fish respiration	5 items	1 item	1 item	1 item

Table 3.2: Design of the knowledge and competency test

In version 1, each booklet contained four units. Each unit had 6 items corresponding to the three dimensions of experimentation. This meant each booklet in version 1 had 24 items and each dimension had eight items.

Booklet 111 consisted of Unit 1: Seed germination, Unit 2: Chicken eggs, Unit 3: Apple wine and Unit 4: Baking bread.

Booklet 121 consisted of Unit 1: Seed germination, Unit 5: The growth of bean plants, Unit 6: Potatoes and Unit 7: Heart beat.

Booklet 131 involved Unit 1: Seed germination, Unit 2: Chicken eggs, Unit 8: Plant growth and Unit 9: Fish respiration.

In version 2, each booklet contained five units. Each unit contained three items, so each booklet contained 15 items in the competency test and each dimension had five items. In the knowledge test, five units consisted of 27 to 30 items.

Booklet 211 comprised Unit 1: Seed germination, Unit 2: Chicken eggs, Unit 3: Apple wine, Unit 4: Baking bread and Unit 5: The growth of bean plants.

Booklet 221 contained Unit 1: Seed germination, Unit 6: Potatoes, Unit 7: Heart beat, Unit 8: Plant growth and Unit 9: Fish respiration.

Booklet 231 consisted of Unit 2: Chicken eggs, Unit 3: Apple wine, Unit 5: The growth of bean plants, Unit 7: Heart beat and Unit 9: Fish respiration.

Booklet 241 comprised Unit 1: Seed germination, Unit 4: Baking bread, Unit 5: The growth of bean plants, Unit 6: Potatoes and Unit 8: The plant growth.

In version 1, each unit appeared only once, except for Unit 2: Chicken eggs which was present in two booklets (111 and 131) and Unit 1: Seed germination which was used in all three booklets.

In version 2, each unit appeared in two booklets, except for Unit 5: The growth of bean plants and Unit 1: Seed germination. Both of them were used in three booklets.

In the knowledge test, version 1 consisted of four units, so that each booklet contained 22 to 26 items. In version 2, each booklet comprised five units, so each booklet consisted of 27 to 30 items.

The testing time was one school period of 45 minutes. For the students who took test version 1, the testing time was 23 minutes for the knowledge test and 22 minutes for the competency test. For students who took test version 2, the testing time for the knowledge test was 27 minutes and for the competency test 18 minutes.

Answering format

In the knowledge test, complex multiple-choice questions were used. Each item had four to six sub-questions; the answering format was “yes” or “no”. In the competency test, we used simple multiple-choice items. Each item had four choices (A,B,C,D), in which only one choice was correct, two options were partly correct and the other was incorrect (cf. Chapter 2). The students were asked to choose only one of the four options.

Coding the answers

In the knowledge test, the answer “yes” was not always correct in all cases, thus the raw data were recoded if necessary. In the competency test, we used a partial-credit scoring model (cf. Chapter 1). This model consists of two levels of competencies. If the students chose the correct option, their response was coded 2 (level 2, full credit). If they chose the incorrect answer, their response was coded 0 (level 0, no credit) and if they chose the partly correct option, their response was coded 1 (level 1). Besides, we used another scoring model that does not distinguish an intermediate level of competency. In this scoring model, a correct response is coded 1 (full credit) and an incorrect response is coded 0 (no credit).

2. Findings

2.1. Item difficulty

2.1.1. Method

Nine units in the knowledge test – divided into seven booklets – were tested. For reporting on item difficulties, students from different test booklets who responded to the same unit were combined. However, in order to calculate the mean item difficulty for the booklet, this was done separately for each booklet. First, frequencies for each sub-question were calculated and then mean item difficulties for the items, the units and the booklet.

2.1.2. Findings

2.1.2.1. Item difficulty in the knowledge test

a) Item difficulty for the sub-questions

The item difficulty for most of sub-questions ranged from 20% to 80%. However, in some sub-questions, the number of students who answered correctly was higher than 80%, even higher than 90%. For example, in Unit 7 “Heart beat” or in Unit 4 “Baking bread” nearly a half of the sub-questions were correctly answered with a probability of over 80%. Item 5 of Unit 7 “Heart beat” and item 1 of Unit 9 “Fish respiration” were correctly answered by over 85% of the students (See appendix, tables 7-15).

Although the item difficulty of some sub-questions was high, some sub-questions were correctly answered by less than 20% of the students. For example, item 7 of Unit 2 “Chicken eggs”, item 2 of Unit 4 “Baking bread” and three sub-questions in Unit 6 “Potatoes” were correctly answered by less than 17 % of the students. (See appendix, tables 7-15).

b) Mean item difficulty for the items and for the units

Unit	Number of students	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Mean item difficulty for unit
Seed germination	705	60.4	80.0	61.6	75.6	70.9	54.4		67.2
Chicken eggs	395	52.0	71.4	52.5	66.2	67.0	72.9	44.9	61.0
Apple wine	298	72.6	58.3	71.8	70.3	55.8	67.6		66.1
Baking bread	310	78.3	49.5	60.7	69.4	72.7			66.1
Bean plants	413	62.7	72.3	74.6	82.9	66.6			71.8
Potatoes	404	85.7	57.2	40.1	36.6	53.4	63.3		56.1
Heart beat	392	71.9	75.3	71.5	70.7	95.1	86.0		78.4
Plant growth	419	80.4	65.9	38.3	48.0	81.6			62.8
Fish respiration	407	90.1	70.4	64.9	68.0	55.6			69.8

Table 3.3: Mean item difficulty for the items and for the units in the knowledge test

The mean item difficulty for the items in the knowledge test ranged from 35% to 80%. However, six items were correctly answered with a probability of over 80%, in particular, item 4 of unit 5, item 1 of unit 6, item 1 of unit 9, item 6 of unit 7 and items 1 and 5 of unit 8. The mean item difficulty for the unit ranged from 56.1% to 78.4%. The item difficulty for Unit 6 “Potatoes” was the highest and it was the lowest for Unit 7 “Heart beat”.

c) Mean item difficulty for the booklets in the knowledge test

Booklet	Mean item difficulty for booklet
111	66.1
121	70.3
131	65.0
211	65.6
221	67.4
231	69.6
241	63.9

Table 3.4: Mean item difficulty for the booklets in the knowledge test

The mean item difficulty for the booklets ranged from 63.9% to 70.3%. In terms of item difficulty, there was not much variation between the seven booklets.

Discussion

Item selection, typically, follows the rule that items that are too easy to solve (> 80%) and too difficult to solve (< 20%) must be revised. The mean item difficulty for most items in this test stayed within the range above 35% and below 80% which means that in terms of item difficulty, the test is appropriate for students who are in grades five and

six. However, the items which identified in the previous section as too easy and too difficult need to be revised and retrialed.

2.1.2.2. Item difficulty in the competency test

a) Item difficulty for the items

Version 1

Unit	Search in the hypothesis space		Data analysis		Search in the experiment space	
	H1	H2	D1	D2	E1	E2
Seed germination	39.0	50.6	50.0	45.8	68.3	52.8
Chicken eggs	52.5	56.3	32.5	50.0	40.0	38.5
Apple wine	64.1	55.2	67.6	69.7	57.1	56.3
Baking bread	58.0	80.3	56.1	46.2	59.4	54.0
Bean plant	56.6	60.8	51.4	70.0	45.8	57.1
Potatoes	35.5	38.2	54.2	58.3	59.7	31.4
Heart beat	64.9	72.6	78.1	72.6	69.3	37.5
Plant growth	38.4	55.6	36.4	52.8	30.7	62.1
Fish respiration	41.4	34.5	48.9	28.1	41.7	30.6

Table 3.5: Item difficulty for the items in the competency test - Version 1 (H1, H2 were item 1 and item 2 in the dimension “search in the hypothesis space”; D1, D2 were items in the dimension “data analysis”; E1, E2 were items in the dimension “search in the experiment space”)

In version 1, the item difficulty for most items in the competency test ranged from 30% to 80%. No item was correctly answered with a probability of below 20%. Only one item, namely item H 2 of unit 4, was correctly answered with a probability of 80.3%.

Version 2

Unit	Search in the hypothesis space	Data analysis	Search in the experiment space
1	24.1	42.2	54.4
2	40.4	49.5	33.3
3	45.4	44.7	41.5
4	40.3	39.4	47.6
5	51.8	48.7	25.7
6	42.9	42.5	49.3
7	61.5	57.1	44.5
8	42.3	57.9	49.7
9	42.5	41.7	39.5

Table 3.6: Item difficulty for the items in the competency test - Version 2

In version 2, the item difficulty for items in the competency test ranged from 24.1% to 61.5%. Most of the items were correctly answered with a probability of over 40%.

b) Mean item difficulty for the dimensions and for the units in both of versions

Version	Unit	Mean item difficulty for “Search in the hypothesis space”	Mean item difficulty for “Data analysis”	Mean item difficulty for “Search in the experiment space”	Mean item difficulty for unit
1	1	44.8	47.9	60.6	51.1
	2	54.4	41.3	39.3	45.0
	3	59.7	68.7	56.7	61.7
	4	69.2	51.2	56.7	59.0
	5	58.7	60.7	51.5	57.0
	6	36.9	56.3	45.6	46.2
	7	68.8	75.4	53.4	65.8
	8	47.0	44.6	46.4	46.0
	9	38.0	38.5	36.2	37.5
		All units	53.1	53.8	49.6
2	1	24.1	42.2	54.4	40.2
	2	40.4	49.5	33.3	41.1
	3	45.4	44.7	41.5	43.9
	4	40.3	39.4	47.6	42.4
	5	51.8	48.7	25.7	42.1
	6	42.9	42.5	49.3	44.9
	7	61.5	57.1	44.5	54.4
	8	42.3	57.9	49.7	50.0
	9	42.5	41.7	39.5	41.2
		All units	43.5	47.1	42.8

Table 3.7: Mean item difficulty for the dimensions and for the units in the competency test

The percentage of students who correctly answered the items in the competency test ranged from 36.6% to 75.4% in version 1 and from 24.1% to 61.5% in version 2. In version 1, the mean item difficulty for the dimension “search in the hypothesis space” was 53.1% and 53.8% for the dimension “data analysis”. For the dimension “search in the experiment space” the mean item difficulty was only slightly lower.

In contrast to version 1, the mean item difficulties in version 2 were slightly lower for all three dimensions and ranged from 42.8% to 47.1%.

c) Mean item difficulty for the booklets and for the versions in the competency test

Booklet	Mean item difficulty for booklet	Mean item difficulty for version
111	57.1	51.6
121	54.7	
131	42.9	
211	39.7	44.1
221	46.3	
231	46.2	
241	44.3	

Table 3.8: Mean item difficulty for the booklets and for the versions in the competency test

The mean item difficulty for the booklets ranged from 39.7% to 57.1%. Most booklets were correctly answered with a probability of over 39%. The mean item difficulty for version 1 was 51.6% and for version 2 it was 44.1%.

Discussion

In version 1, the item difficulty for most items ranged from 30% to 80%. In version 2, the item difficulty of most items ranged from 40% to 50%. No item was correctly answered with a probability of over 80% or below 20% which means that the item difficulties of all items in this version were acceptable. However, some items were quite difficult, for example, the items in the dimension “search in the hypothesis space” in unit 1 and the items in the dimension “search in the experiment space” in unit 5 in version 2 with item difficulties of 24% and 25%. Both of these two items must be considered carefully and the texts or pictures must be changed in order to facilitate understanding.

The mean item difficulties for all three dimensions in version 1 can be considered ideal because they range around 50%. For version 2, mean item difficulties for the three dimensions are also acceptable, but the test is a little more difficult, as anticipated during item development.

2.2. Reliability

2.2.1. Method

The Cronbach's alpha and the corrected item-total correlation were calculated in order to assess the statistical qualities of the knowledge test, competency test version 1 and competency test version 2. If the Cronbach's alpha exceeds 0.70, a test can be considered acceptable, though higher reliability coefficients are of course preferred.

In the competency test, version 1, the scales "forming hypotheses", "data analysis" and "planning experiments" consisted of eight items each so that it was possible to calculate the Cronbach's alpha for each scale as well as the Cronbach's alpha for the complete test. However, for the competency test, version 2, the three dimensions consisted of only five items each so that the reliability of the test was assessed at the test booklet level and not at the scale level because of insufficient scale length.

2.2.2. Findings

2.2.2.1. Reliability of the knowledge test

a) Reliability of the knowledge test at the unit level

Unit 1: Seed germination (n = 705)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.43	
2	0.39	
3	0.39	0.65
4	0.42	
5	0.33	
6	0.33	

Table 3.9: Reliability of Unit 1: Cronbach's alpha and corrected item-total correlation

The reliability coefficient (Cronbach's alpha) at the unit level for Unit 1 "Seed germination" was 0.65 and all items had a corrected item-total correlation higher than 0.3.

Unit 2: Chicken eggs (n = 395)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.27	0.43
2	0.18	
3	0.17	
4	0.19	
5	0.29	
6	0.16	
7	0.12	

Table 3.10: Reliability of Unit 2: Cronbach's alpha and corrected item-total correlation

The reliability coefficient at the unit level for unit 2 was 0.43. Only two items had a corrected item-total correlation higher than 0.2.

Unit 3: Apple wine (n = 298)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.33	0.56
2	0.28	
3	0.48	
4	0.24	
5	0.39	
6	0.18	

Table 3.11: Reliability of Unit 3: Cronbach's alpha and corrected item-total-correlation

The reliability coefficient for unit 3 was 0.56. Only one item had a corrected item-total-correlation lower than 0.2. After deleting this item, the Cronbach's alpha for the unit was 0.58.

Unit 4: Baking bread (n = 310)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.32	0.59
2	0.42	
3	0.42	
4	0.37	
5	0.23	

Table 3.12: Reliability of Unit 4: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for this unit was 0.59. All five items had a corrected item-total correlation higher than 0.2.

Unit 5: The growth of bean plants (n = 413)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.35	0.47
2	0.22	
3	0.23	
4	0.28	
5	0.20	

Table 3.13: Reliability of Unit 5: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 5 was 0.47. All items had a corrected item-total correlation higher than 0.2.

Unit 6: Potatoes (n = 404)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.22	0.36
2	0.25	
3	0.29	
4	-0.03	
5	0.13	
6	0.18	

Table 3.14: Reliability of Unit 6: Cronbach's alpha and corrected item-total correlation

The reliability coefficient at the unit level for unit 6 was 0.36. Only three items had a corrected item-total correlation higher than 0.2. After deleting the other three items, the Cronbach's alpha was 0.45.

Unit 7: Heart beat (n = 392)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.44	0.70
2	0.52	
3	0.47	
4	0.48	
5	0.45	
6	0.27	

Table 3.15: Reliability of Unit 7: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 7 was 0.70. Furthermore, all six items had a corrected item-total correlation higher than 0.2.

Unit 8: Plant growth (n = 419)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.34	0.48
2	0.36	
3	0.10	
4	0.18	
5	0.37	

Table 3.16: Reliability of Unit 8: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 8 was 0.48. Three out of the five items had a corrected item-total-correlation higher than 0.2. The highest Cronbach's alpha after deleting the non-discriminating items was 0.53. However one item with a corrected item-total correlation lower than 0.2 was still left.

Unit 9: Fish respiration (n = 407)

Item	Corrected item-total correlation	Cronbach's alpha
1	0.38	0.56
2	0.33	
3	0.31	
4	0.30	
5	0.30	

Table 3.17: Reliability of Unit 9: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 9 was 0.56. However, all five items had a corrected item-total correlation higher than 0.3.

Unit	Cronbach's alpha	Unit	Cronbach's alpha
1. Seed germination	0.65	6. Potatoes	0.45
2. Chicken eggs	0.43	7. Heart beat	0.70
3. Apple wine	0.58	8. Plant growth	0.53
4. Baking bread	0.59	9. Fish respiration	0.56
5. Bean plant growth	0.47		

Table 3.18: Reliability coefficients (Cronbach's alpha) at the unit level in the knowledge test

Discussion

Most units in the knowledge test have reliability coefficients that stay below the acceptable limit of 0.70. As a consequence, in further analyses it will not be possible to look at correlations between knowledge scores and competency scores at the unit level. Rather, analyses of reliabilities at the test booklet level must reveal if it is possible to investigate correlations between pre-knowledge across different biological topics and the three dimensions in experimentation. This will be done in the following section.

b) Reliability of the knowledge test at the booklet level

Booklet	Cronbach's alpha for all items	Cronbach's alpha after item selection
111	0.67 (24items)	0.73 (15 items)
121	0.66 (23 items)	0.75 (15 items)
131	0.55 (24 items)	0.69 (15 items)
211	0.73 (29 items)	0.77 (20 items)
221	0.77 (28 items)	0.79 (24 items)
231	0.75 (29 items)	0.76 (23 items)
241	0.53 (27 items)	0.71 (16 items)

Table 3.19: Reliability coefficient (Cronbach's alpha) at the booklet level in the knowledge test before and after item selection

The reliability coefficients at the booklet level ranged from 0.53 to 0.77 before item selection and from 0.71 to 0.79 after item selection..

Discussion

The reliability coefficients (Cronbach's alpha) at the booklet level were higher than 0.7 in six out of seven booklets and Cronbach's alpha for booklet 131 was 0.69. As a consequence, the knowledge test at the booklet level is reliable enough and can be used to investigate correlations between the students' biological pre-knowledge and their competencies concerning the three dimensions of experimentation. It should be kept in mind, though that all scales used for these analyses combine items from different units, so that within the scope of the following analyses, it is impossible to make statements about knowledge and competencies concerning specific biological knowledge domains, like seed germination. Rather, when statements are made about students' knowledge and students' competences, they refer to a range of different biological topics.

2.2.2.2. Reliability of the competency test

a) Reliability of the competency test at the unit level

Version 1

Unit 1: Seed germination

Dimension	Item	Corrected item- total correlation (all items)	Cronbach's alpha (all items)	Corrected item- total correlation (item selection)	Cronbach's alpha (item selection)
Search in the hypothesis space	1	0.52	0.61	0.60	0.69
	2	0.40		0.43	
Data analysis	3	0.41		0.48	
	4	0.35		0.40	
Search in the experiment space	5	0.14			
	6	0.25			

Table 3.20: Reliability of Unit 1: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient at the unit level for unit 1 was 0.61. Five out of the six items had a corrected item-total correlation higher than 0.2. After selecting these items, the Cronbach's alpha was 0.69.

Unit 2: Chicken eggs (n = 170)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.23	0.61
	2	0.40	
Data analysis	3	0.36	
	4	0.34	
Search in the experiment space	5	0.43	
	6	0.30	

Table 3.21: Reliability of Unit 2: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient at the unit level for unit 2 was 0.61. All six items had a corrected item-total correlation higher than 0.2.

Unit 3: Apple wine (n = 73)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.55	0.74
	2	0.53	
Data analysis	3	0.52	
	4	0.39	
Search in the experiment space	5	0.49	
	6	0.45	

Table 3.22: Reliability of Unit 3: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient for unit 3 was 0.74. Moreover, all six items had a corrected item-total correlation higher than 0.3.

Unit 4: Baking bread (n = 73)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.69	0.75
	2	0.45	
Data analysis	3	0.67	
	4	0.50	
Search in the experiment space	5	0.40	
	6	0.29	

Table 3.23: Reliability of Unit 4: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient for unit 4 was 0.75. All six items had a corrected item-total correlation higher than 0.2.

Unit 5: Bean plant growth (n = 82)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.39	0.67
	2	0.64	
Data analysis	3	0.43	
	4	0.37	
Search in the experiment space	5	0.24	
	6	0.38	

Table 3.24: Reliability of Unit 5: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient for unit 5 was 0.67. All items had a corrected item-total correlation higher than 0.2.

Unit 6: Potatoes (n = 82)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.36	0.46
	2	0.10	
Data analysis	3	0.23	
	4	0.22	
Search in the experiment space	5	0.31	
	6	0.18	

Table 3.25: Reliability of Unit 6: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient for unit 5 was 0.46, with two items of low discriminatory power. After selecting items 2 and 6, the Cronbach's alpha was 0.48.

Unit 7: Heart beat (n = 82)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.31	0.56
	2	0.23	
Data analysis	3	0.33	
	4	0.44	
Search in the experiment space	5	0.32	
	6	0.21	

Table 3.26: Reliability of Unit 7: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient for unit 7 was 0.56. However, all items had a corrected item-total correlation higher than 0.2.

Unit 8: Plant growth (Plant nutrients) (n = 97)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.35	0.69
	2	0.56	
Data analysis	3	0.36	
	4	0.60	
Search in the experiment space	5	0.30	
	6	0.39	

Table 3.27: Reliability of Unit 8: Corrected item-total correlation and Cronbach's alpha

The Cronbach's alpha for unit 8 was 0.69. All items had a corrected-item-total correlation higher than 0.2.

Unit 9: Fish respiration (n = 97)

Dimension	Item	Corrected item-total correlation	Cronbach's alpha
Search in the hypothesis space	1	0.45	0.67
	2	0.42	
Data analysis	3	0.41	
	4	0.49	
Search in the experiment space	5	0.26	
	6	0.37	

Table 3.28: Reliability of Unit 9: Corrected item-total correlation and Cronbach's alpha

The Cronbach's alpha for unit 9 was 0.67. All items had a corrected item-total correlation higher than 0.2.

Unit	Cronbach's alpha	Unit	Cronbach's alpha
1. Seed germination	0.69	6. Potatoes	0.49
2. Chicken eggs	0.61	7. Heart beat	0.56
3. Apple wine	0.75	8. Plant growth	0.69
4. Baking bread	0.75	9. Fish respiration	0.67
5. Bean plant growth	0.67		

Table 3.29: Reliability coefficients (Cronbach's alpha) at the unit level in Version 1 in the competency test**Discussion**

The reliability coefficients at the unit level in the competency test, version 1, ranged from 0.49 to 0.75. In some units, the reliability was very low, for example, in unit 6 or unit 7, where the Cronbach's alpha was lower than 0.6. However, in some other units, the reliability coefficient was quite high, such as in unit 3 and unit 4, where the Cronbach's alpha was 0.75. Low reliabilities at the unit level, however, can be expected, because in a unit three different item types that require different competencies are mixed. More coherent scales, thus, should be formed by combining the same competency – i.e., search in the hypothesis space – across different units, if the competency is not related to content-specific knowledge (cf. table 3.39).

Version 2

Unit	Number of students	Corrected item-total correlation			Cronbach's alpha for unit
		Item 1: Search in the hypothesis space	Item 2: Data analysis	Item 3: Search in the experiment space	
1. Seed germination	453	0.04	- 0.03	- 0.12	- 0.10
2. Chicken eggs	225	0.05	0.08	0.04	0.11
3. Apple wine	225	0.35	0.24	0.19	0.43
4. Baking bread	237	0.13	0.02	0.11	0.19
5. Bean plant growth	331	0.24	0.24	0.05	0.31
6. Potatoes	339	0.01	0.07	0.06	0.10
7. Heart beat	310	0.13	0.26	0.19	0.34
8. Plant growth	339	0.33	0.32	0.21	0.46
9. Fish respiration	310	0.24	0.27	0.21	0.40

Table 3.30: Reliability at the unit level in the competency test in Version 2: Corrected item-total correlation and Cronbach's alpha

The reliability coefficients at the unit level in version 2 ranged from 0.1 to 0.4. Only three of the nine units (unit 3, unit 8 and unit 9) had a Cronbach's alpha higher than 0.4. Also, the corrected item-total-correlation for most items was lower than 0.2. Only in unit 8 and in unit 9 did all items have a corrected item-total correlation higher than 0.2.

Discussion

The reliability at the unit level in version 2 was lower than in version 1, with Cronbach's alphas for all units lower than 0.5. This is related to the fact that many items had a corrected item-total correlation lower than 0.2.

Possible reasons for this are the short scale length in version 2 and the higher item difficulty in version 2 than in version 1.

b) Reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level

Booklet 111

Unit	Item	Scale “forming hypotheses”		Scale “data analysis”		Scale “planning experiments”	
		CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Seed germination	1	0.46		0.38		0.39	
	2	0.36		0.19		0.33	
Chicken eggs	1	0.15	0.75	0.59	0.73	0.47	0.68
	2	0.58		0.43		0.27	
Apple wine	1	0.47		0.43		0.36	
	2	0.69		0.50		0.44	
Baking bread	1	0.46		0.59		0.41	
	2	0.48		0.32		0.28	

Table 3.31: The reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and reliability coefficient (Cronbach’s alpha) Booklet 111

The reliability coefficients of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level for booklet 111 ranged from 0.68 to 0.75. The Cronbach’s alphas of the scales “forming hypotheses” and “data analysis” were higher than 0.7. The Cronbach’s alpha of the scale “planning experiments” was 0.68. Also, most items in all three scales had a corrected item-total correlation higher than 0.2. Only two items, one in the scale “forming hypotheses” (item 1 of unit “Chicken eggs”) and the other in the scale “data analysis” (item 2 of unit “Seed germination”) had a corrected item-total correlation lower than 0.2. In the scale “planning experiments” all eight items correlated with the corrected sum score.

Booklet 121

Unit	Item	Scale “forming hypotheses”		Scale “data analysis”		Scale “planning experiments”	
		CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Seed germination	1	0.44		0.49		0.05	
	2	0.45		0.48		0.18	
Bean plant growth	1	0.49	0.61	0.39	0.70	0.38	0.51
	2	0.58		0.42		0.26	
Potatoes	1	0.38		0.33		0.40	
	2	0.05		0.39		0.23	
Heart beat	1	0.31		0.28		0.15	
	2	0.23		0.42		0.24	

Table 3.32: The reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and reliability coefficient (Cronbach’s alpha) Booklet 121

In booklet 121, the reliability coefficient of the scale “data analysis” was 0.70. However, the Cronbach’s alpha for the scale “forming hypotheses” was 0.62 and the reliability coefficient for the scale “planning experiments” was 0.51. However, most items had a corrected item-total correlation higher than 0.2. Especially, in the scale “data analysis”, all items correlated with the corrected sum score. However, one item in the scale “forming hypotheses” (item 2 of unit “Potatoes”) and three items out of eight in the scale “planning experiments” (items 1, 2 of unit 1 and item 1 of unit 7) had a corrected item-total correlation lower than 0.2.

Booklet 131

Unit	Item	Scale “forming hypotheses”		Scale “data analysis”		Scale “planning experiments”	
		CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Seed germination	1	0.41		0.39		0.29	
	2	0.51		0.41		0.34	
Chicken eggs	1	0.30	0.70	0.38	0.74	0.47	0.67
	2	0.44		0.40		0.38	
Plant growth	1	0.40		0.53		0.43	
	2	0.43		0.48		0.38	
Fish respiration	1	0.30		0.46		0.22	
	2	0.32		0.40		0.43	

Table 3.33: The reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and reliability coefficient (Cronbach’s alpha) Booklet 131

In this booklet, the reliability coefficients of the scales “forming hypotheses”, “data analysis” and “planning experiments” ranged from 0.67 to 0.74. As in booklet 111, the reliabilities of the scales “forming hypotheses” and “data analysis” were higher than the reliability of the scale “planning experiments” with Cronbach’s alphas of 0.70 and 0.74. Furthermore, all items of each scale had a corrected item-total correlation higher than 0.2.

Booklet 211

Unit	Scale “forming hypotheses”		Scale “data analysis”		Scale “planning experiments”	
	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Seed germination	-0.08		0.14		0.02	
Chicken eggs	0.21		0.08		0.27	
Apple wine	0.29	0.36	0.19	0.29	0.20	0.29
Baking bread	0.23		0.09		0.14	
Bean plant growth	0.25		0.19		0.08	

Table 3.34: The reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and reliability coefficient (Cronbach’s alpha) Booklet 211

In booklet 211, the reliability coefficients of all three scales were lower than 0.40. Furthermore, many items had a corrected item-total correlation lower than 0.2. Especially in the scale “data analysis” all items lacked discriminatory power and in the scale “planning experiments” four out of five items had a corrected item-total correlation lower than 0.2.

Booklet 221

Unit	Scale “forming hypotheses”		Scale “data analysis”		Scale “planning experiments”	
	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Seed germination	0.11		0.34		0.24	
Potatoes	0.12		0.34		0.24	
Heart beat	0.23	0.36	0.28	0.54	0.33	0.52
Plant growth	0.12		0.28		0.34	
Fish respiration	0.29		0.28		0.30	

Table 3.35: The reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and reliability coefficient (Cronbach’s alpha) Booklet 221

The reliability coefficients of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level for booklet 221 ranged from 0.36 to 0.54. The lowest Cronbach’s alpha can be found in the scale “forming hypotheses”. Also, in this scale, only two items had a corrected item-total correlation higher than 0.2. In the scales “data analysis” and “planning experiments”, however, all items had a corrected item- total correlation higher than 0.2.

Booklet 231

Unit	Scale “forming hypotheses”		Scale “data analysis”		Scale “planning experiments”	
	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Chicken eggs	0.34		0.33		0.31	
Apple wine	0.35		0.36		0.28	
Bean plant growth	0.33	0.58	0.23	0.54	0.13	0.50
Heart beat	0.35		0.35		0.32	
Fish respiration	0.31		0.27		0.33	

Table 3.36: The reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and reliability coefficient (Cronbach’s alpha) Booklet 231

The reliability coefficients of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level in this booklet were similar in all three scales and ranged from 0.50 to 0.58. Moreover, in the scales “forming hypotheses” and “data analysis”, all items had a corrected item-total correlation higher than 0.2. In the unit “Bean plant growth”, only one item had to be deleted.

Booklet 241

Unit	Scale “forming hypotheses”		Scale “data analysis”		Scale “planning experiments”	
	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Seed germination	0.18		0.38		0.20	
Baking bread	0.30		0.35		0.27	
Bean plant growth	0.34	0.49	0.27	0.58	0.31	0.44
Potatoes	0.29		0.37		0.29	
Plant growth	0.21		0.32		0.09	

Table 3.37: The reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and reliability coefficient (Cronbach’s alpha) Booklet 241

In this booklet, the reliability coefficient of the scales “forming hypotheses”, “data analysis” and “planning experiments” ranged from 0.44 to 0.58. In the scale “data analysis”, all items had a corrected item-total correlation higher than 0.2. One item (Unit 1: Seed germination) in the scale “forming hypotheses” and one item (Plant growth) in the scale “planning experiments” possessed a corrected item-total correlation lower than 0.2.

After deleting the non-discriminating items in each scale (see appendix, table 24), the Cronbach’s alpha increased. The reliability coefficients of the scales after item selection are shown in the following table:

Booklet	Scale “forming hypotheses”	Scale “data analysis”	Scale “planning experiments”
111	0.78 (7 items)	0.75 (7 items)	0.68 (8 items)
121	0.71 (7 items)	0.71 (8 items)	0.56 (5 items)
131	0.70 (8 items)	0.74 (8 items)	0.67 (8 items)
211	0.49 (4 items)	0.34 (2 items)	0.32 (2 items)
221	0.34 (2 items)	0.54 (5 items)	0.52 (5 items)
231	0.58 (5 items)	0.54 (5 items)	0.53 (4 items)
241	0.49 (4 items)	0.58 (5 items)	0.48 (4 items)

Table 3.38: The reliability coefficients of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level – Item selection

Discussion

When the whole sample is considered, version 1 of the competency test has a much higher reliability for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level than version 2. Most Cronbach’s alphas in version 1 were higher than 0.7. The others were higher than 0.6. Only the Cronbach’s alpha of the scale “planning experiments” in booklet 121 was lower than 0.6. On the other hand, most items in the booklets had a corrected item-total correlation higher than 0.2. In particular, in booklet 131 all items for each scale correlated with the sum score. In booklet 111 and 121, some items had a corrected item-total correlation lower than 0.2.

In contrast, in version 2, the reliability for the scales “forming hypotheses”, “data analysis” and “planning experiments” was very low. In some scales, the Cronbach’s alpha ranged between 0.2 and 0.3 (Booklet 211). Most Cronbach’s alphas were lower than 0.5. Often, the corrected item-total correlation was lower than 0.2 (Booklet 211, 221).

After deleting the non-discriminating items, the Cronbach’s alphas increased in all booklets. In version 1, the Cronbach’s alphas ranged from 0.56 to 0.78. All Cronbach’s

alphas for the scales “forming hypotheses” and “data analysis” were higher than 0.7. For the scale “planning experiments” the Cronbach’s alphas for booklet 111 and 131 were higher than 0.67.

However, in version 2 the reliability was still low, with Cronbach’s alphas ranging from 0.32 to 0.58. Among them, half of the Cronbach’s alphas were lower than 0.5.

Table 3.38 shows the reliability for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level for seven booklets after deleting the non-discriminating items. In booklet 111, the two scales “forming hypotheses” and “data analysis” have a Cronbach’s alpha higher than 0.7. The other scale has a Cronbach’s alpha of 0.68. Besides, in booklet 131, the Cronbach’s alphas for the scale “data analysis” and “planning experiments” are higher than 0.7 and for the other scale the Cronbach’s alpha is 0.67. Thus, the Cronbach’s alphas for all three scales in these two booklets indicate that the scales are quite reliable. Accordingly, these two booklets can be used to investigate the correlations between the students’ knowledge and the three dimensions in experimentation.

Booklet 121, however, cannot be used. Although for the two scales “forming hypotheses” and “data analysis” the Cronbach’s alphas are higher than 0.7, the scale “planning experiments” is not reliable enough with a Cronbach’s alpha of 0.56. The scales in all four booklets in version 2 cannot be used for further analyses because their Cronbach’s alphas were lower than 0.6. Thus, booklet 121 in version 1 and all booklets in version 2 cannot be used for further analyses.

Since the reliability coefficients of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level were much lower in version 2 than in version 1, it is necessary to determine why. Three possible explanations can be discussed. Maybe weaker students took version 1 and stronger students took version 2. Since we do not have any indicator of student achievement – like the Biology grade – it is impossible to find a definitive answer to this. Second, the difference in item difficulty between the units and the versions may be responsible for differences in test reliability. This is a very likely reason because difficult items often coincide with greater numbers of students who guess which is detrimental to test reliability. Third, the length of the test scale may be a reason, because in version 1 test booklets have more items. We know from test theory that longer scales have a higher reliability. In fact, the Spearman-Brown

formula describes the relationship between scale length and test reliability and can be used to compare scales of different length. This will be done in the following section.

We adjusted reliabilities according to the Spearman-Brown formula (see appendix, p. 294) to be able to compare the shorter scales in version 2 and the longer scales in version 1.

Booklet	Scale “forming hypotheses”(8 items)	Scale “data analysis”(8 items)	Scale “planning experiments”(8 items)
111	0.80	0.77	0.68
121	0.74	0.74	0.67
131	0.70	0.74	0.67
211	0.66	0.67	0.65
221	0.67	0.65	0.63
231	0.69	0.66	0.69
241	0.66	0.69	0.65

Table 3.39: The reliability coefficients (Cronbach’s alpha) of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level - after using Spearman-Brown to adjust the scale length.

Table 3.39 shows what the reliability coefficients would look like if the scales in version 2 were longer. The table shows that the reliability coefficients for version 2 (i.e. booklets 211, 221, 231 and 241) are still lower than in version 1 for all three scales.

c) Reliability of the booklet (all items of the four units combined) in the competency test

Booklet	Cronbach’s alpha of the booklet	Mean alpha for version
111	0.89 (24 items)	0.86
121	0.84 (24 items)	
131	0.85 (24 items)	
211	0.59 (15 items)	0.67
221	0.65 (15 items)	
231	0.74 (15 items)	
241	0.70 (15 items)	

Table 3.40: The reliability of the booklet (all items of the four or five units combined)

The reliability for all items of the test combined is much higher in version 1 than in version 2. In all three booklets in version 1, the Cronbach’s alpha is higher than 0.8, especially in booklet 111, where the Cronbach’s alpha is 0.89. Furthermore, most items in all three booklets have a corrected item-total correlation higher than 0.2 (see appendix, tables 25-27).

In version 2, only the Cronbach's alphas of booklets 231 and 241 were higher than 0.7, but the reliability for booklet 211 was low with a Cronbach's alpha of 0.59. Also, the mean reliability for version 1 was 0.86, while it was only 0.67 in version 2. The reasons for this have been discussed in the paragraphs above.

The Spearman-Brown formula was used to adjust the test scales of version 2: Table 3.41 shows the results of this operation.

Booklet	Cronbach's alpha of the booklet	Cronbach's alpha of the booklet, if all scales had had 24 items (adjusted after Spearman-Brown)
111	0.89 (24 items)	0.89
121	0.84 (24 items)	0.84
131	0.85 (24 items)	0.85
211	0.59 (15 items)	0.70
221	0.65 (15 items)	0.75
231	0.74 (15 items)	0.82
241	0.70 (15 items)	0.80

Table 3.41: The reliability of the booklet (all items of the four or five units combined) - before and after using Spearman-Brown formula

After adjusting the scale of version 2 according to Spearman-Brown, the reliability for the booklets increased. For booklet 231, for example, it increased from 0.74 to 0.82. However, compared with all three booklets in version 1, the reliability coefficients of all four booklets in version 2 were still lower. Possible reasons for this have been discussed in the paragraphs above.

d) Comparison of the reliability coefficients for the tests when two different scoring models are used

Booklet	Cronbach's alpha of the booklet (scoring model 0/1)	Cronbach's alpha of the booklet (scoring model 0/1/2)
111	0.89	0.87
121	0.84	0.80
131	0.85	0.83
211	0.59	0.56
221	0.65	0.65
231	0.74	0.72
241	0.70	0.72

Table 3.42: The reliability of the booklet (all items of the four or five units combined) for two scoring models.

Comparing test reliabilities that result from a 0/1 scoring model (full credit/no credit) and a 0/1/2 scoring model (partial credit), table 3.42 shows that the 0/1/2 scoring model

which is a partial credit scoring model, brings down booklet reliabilities to some degree, but not considerably.

From this follows that a partial credit scoring model can be used because level 1 the students probably do not shift very often between level 0 and level 2.

Discussion

The reliability coefficients of the seven test booklets tested in this study ranged from 0.59 to 0.89. All three booklets in version 1 had Cronbach's alphas higher than 0.8. And two booklets in version 2 had Cronbach's alphas higher than 0.7. Moreover, most items in version 1 had a corrected item-total correlation higher than 0.2, while in version 2, many items had a corrected item-total correlation lower than 0.2 (see appendix, tables 25-31).

So, the reliability in version 1 was much higher than in version 2. One of the reasons for this was the shorter scale length in version 2. After using the Spearman-Brown formula to adjust the same scale length in both versions, the reliability coefficient of the booklet (all items of the four or five units combined) was still lower in version 2 than in version 1. The other possible explanation for lower reliabilities – higher item difficulties that may lead to higher percentages of students who guess – was in part substantiated by the finding that items in version 2 were more difficult to solve than items in version 1 (cf. table 3.8).

Furthermore, a comparison of the two scoring models showed that the reliability coefficients for the scoring model 0/1 were slightly higher than that for the scoring model 0/1/2. Thus, both models can be used to assess correlation coefficients and levels of competency of students in experimentation.

GENERAL DISCUSSION

In the knowledge test, the reliability coefficient at the unit level was very low in most units. The Cronbach's alphas ranged from 0.43 to 0.70. Among them, only unit 5 had a Cronbach's alpha higher than 0.7. Thus, the knowledge test at the unit level was not reliable. However, at the booklet level, the reliability for six booklets out of seven was higher than 0.7 and the Cronbach's alpha for booklet 131 was nearly 0.70. This indicates the knowledge test at the booklet level was reliable and we could use all seven booklets to calculate the correlations between the students' pre-knowledge and the three dimensions in experimentation.

In the competency test, at the unit level, in version 1, the reliability was not so high in some units (units 6 and 7) as the Cronbach's alphas for these units were lower than 0.6,

whereas in other units (units 3 and 4) the Cronbach's alpha was higher than 0.7. However, in version 2, the reliability at the unit level was very low in all units. Here, the Cronbach's alphas were lower than 0.5.

At the level of the scales "forming hypotheses", "data analysis" and "planning experiments", the reliability coefficient in version 1 was also higher than in version 2. In version 1, two booklets (111 and 131) were reliable and had Cronbach's alphas for all three scales higher than 0.7 or nearly 0.7. These two booklets can be used to calculate the correlations between the students' pre-knowledge and the three dimensions of experimentation and the correlations within the three dimensions. In contrast, all four booklets in version 2 were not reliable, because all Cronbach's alphas were lower than 0.6. This is also true for booklet 121 though two scales had acceptable Cronbach's alphas higher than 0.7. In the scale "planning experiments" the Cronbach's alpha was only 0.56.

On the other hand, the reliability for all items combined, i.e., the items that form the three scales "forming hypotheses", "data analysis" and "planning experiments", was very high in version 1. Here, all Cronbach's alphas were higher than 0.8, whereas it was not so high in version 2. Only two booklets (231, 241) had a Cronbach's alpha higher than 0.7 and booklet 211 had a Cronbach's alpha lower than 0.6. At the booklet level in the competency test, five booklets out of seven were reliable.

Thus, in the competency test, at the unit level or at the booklet level, the reliability in version 2 was much lower than in version 1. One of the reasons for this was that the scale length was not the same in two versions (version 1 had 4 units with 24 items; version 2 had 5 units with 15 items). So, we used the Spearman-Brown formula to adjust the scale length in both versions to be the same. However, the reliability at all levels in version 2 was still lower than in version 1. Therefore, the reliability at the unit level or at the booklet level was influenced by the item difficulty in which was lower in version 1 than in version 2.

2.3. Latent class analysis for the competency test

2.3.1. Method

In the competency test, the main goal is to assess levels of competency of students in experimentation. The problem can be stated in different terms: Even though two students may have the same total score, they may still represent different levels of competencies. For example, in the dimension “search in the hypothesis space”, there are 8 items in each booklet. If both students gained the same scores, but one student answered the tasks only partly correctly and the other was able to solve many tasks fully correctly but missed some of the tasks, the sum score would not reveal the differences between them. So the question arises: Do both students have the same patterns of solving items correctly? In order to answer this question, we used the latent class analysis (LCA). One of the purposes of LCA is to assign students into different latent classes and to find out the correlation between the classes of students and the three dimensions of experimentation. By means of latent class analysis it is possible to identify subgroups of students with different levels of competency in experimentation.

We used the Winmira program to run the latent class analysis.

However, Winmira can be only used with a relatively large number of persons. In all our booklets, the number of students was not large enough and many students had missing values. Thus, we combined all booklets in each version in one file, regardless of the fact if the units in the booklets were different. However, because the reliabilities at the unit level, booklet level and scale level in version 2 were not high enough, we used only version 1 for the latent class analysis. In version 1, the number of students that responded to booklet 111, 121 and 131 was 252.

We used the 0/1/2 scoring model for these analyses.

The Winmira program can estimate the parameters for 8 different latent class analysis models for ordinal variables. In order to compute the classifying test models, the number of classes had to be specified before. Thus, the number of classes is not represented by a model parameter estimated with the modelling.

In our study, we considered three main models with two to four latent classes. In the two-class model, students are assigned into two groups. In the one group, the students are experts with a high level of competency and in the other the students are less accomplished with low levels of competencies.

In the three-class model, students are assigned into three classes. People in class 1 do not have well-developed competencies. In class 2 there are people who possess only factual

knowledge, in the third class; there are persons with great factual knowledge and high levels of competencies. In the four-class model, there are the same classes as in the three-class model, but there is one more class with people who have factual knowledge about a domain, but do not always solve items that test competencies correctly.

2.3.2. Findings

Latent class analysis for all items in the competency test

In order to explain the data, we tested four probable test models, a quantification one (Rasch model) and three classifications (Latent class analysis). The following table shows the associated model value index for the tested model.

Model	BIC index	CAIC index
One dimension Rasch model (Rasch 1)	7498	7444
Classified model with two latent classes (LCA 2)	7078	7175
Classified model with three latent classes (LCA 3)	7162	7308
Classified model with four latent classes (LCA 4)	7317	7512

Table 3.43: Value of the probable tested models

Table 3.43 indicates that the Rasch model was not adequate, because its BIC index and CAIC index were the highest in all four models. Therefore, one of three quality models should be chosen. The two-class model was the best with both the lowest BIC index and CAIC index. However, the two-class model also provided the differentiation about the quantification and the item profile for this model was not as good as the three-class model. On the other hand, we assessed levels of competency here for all items of the test. Therefore, we could also choose the second solution, the three-latent-class model. This model delivered a stronger differentiation between item profiles than the two-class model. In order to ensure the reliability of the selected test model, especially when it was not the best solution, we looked at another criterion which is the mean of response probability of students, shown in the following table.

Statistics of expected class membership:

Class	Expected size	Mean probability	Assignment probability	Assignment probability	Assignment probability	Number of students
			Class 1	Class 2	Class 3	
1	0.474	0.989	0.989	0.011	0.000	75
2	0.353	0.958	0.040	0.958	0.001	54
3	0.173	0.993	0.000	0.007	0.993	27

Table 3.44: Mean of response probability for three latent classes model

If we assign students into three classes, the mean of maximum response probability of one class in three cases was at least 95%. The alternative second and third-highest assignment probabilities did not exceed a value of 4% for any class. Thus, the three-latent-class model was also reliable and could be used to assign students into three classes of competency in experimentation.

We assigned students into two and then three classes of competency and we calculated the mean score for each dimension in experimentation in each class. Figures 3.4 and 3.5 show that the resulting relationship between the groups of students and the three dimensions in experimentation.

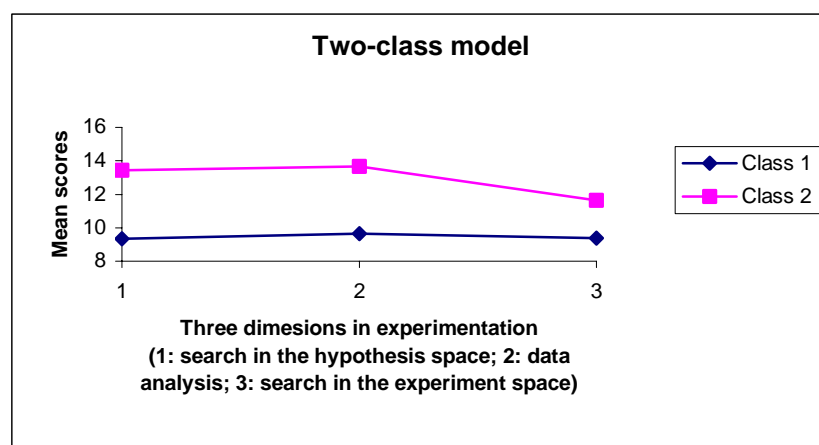


Figure 3.4: The relationship between groups of students and the three dimensions in experimentation (Two-class model)

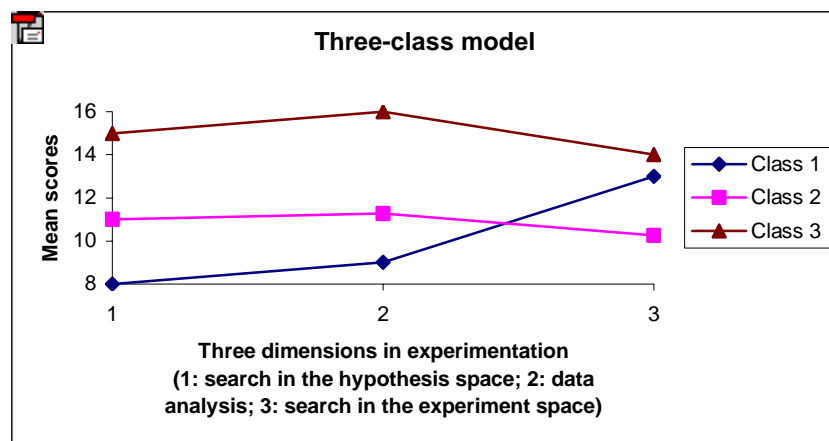


Figure 3.5: The relationship between groups of students and the three dimensions of experimentation (Three-class model)

According to the two-class model, students in class 2 solved the tasks in all three dimensions of experimentation better than those in class 1.

In the three-class model, students in class 3 always solved the tasks in all three dimensions better than the students in the other two classes. However, there was a difference between class 1 and class 2. Students in class 2 solved the tasks in the dimensions “search in the hypothesis space” and “data analysis” better than the students in class 1. In contrast, students in class 1 solved the tasks in the dimension “search in the experiment space” better than the students in class 2.

Discussion

Latent class analyses shows that there are three types of students. One type consists of high-achievers who outdo the other types in all three dimensions. The other two types have different profiles. They are different insofar as one type has lower competencies in the dimensions “search in the hypothesis space” and “data analysis” and higher competencies in the dimension “search in the experiment space” than the other. This can be interpreted as a student type that possesses a greater amount of content knowledge about the science content behind the experiments. On the basis of this knowledge, the students are better at forming hypotheses and analyzing data. On the other hand, these students have lower methodological knowledge about the method of experimentation as they are worse in this dimension than the other students.

However, it must be kept in mind that these interpretations are based on a rather small sample and that the data were compiled from three different test booklets. Thus, when interpreting these findings, some caution needs to be exerted and it remains to be seen in the main study if the results are stable.

In order to determine the relationship between students in different classes and three dimensions in experimentation with a greater degree of certainty, the test needs to be repeated with only one test booklet and a larger sample.

2.4. Correlations

2.4.1. Method

We calculated the correlation coefficients between the students' knowledge and the three dimensions in experimentation. We also investigated the correlations between the three dimensions of experimentation. In order to do this, we calculated sum scores for the knowledge test at the booklet level and sum scores for each dimension in experimentation.

In the knowledge test, the reliability at the booklet level for all booklets was high enough (all of the Cronbach's alphas were higher than 0.7) and we used all of them. We also used booklets 111 and 131 in version 1 in order to calculate the correlations between the three dimensions in experimentation

2.4.2. Findings

2.4.2.1. Correlations (Spearman) between the students' pre-knowledge and the three dimensions in experimentation

Booklet	Pre-knowledge * Search in the hypothesis space	Pre-knowledge * Data analysis	Pre-knowledge * Search in the experiment space
111 (n = 73)	0.020	0.123	0.223
131 (n = 97)	0.197	0.264**	0.110

Table 3.45: Correlation coefficients (Spearman) between the students' pre-knowledge and the three dimensions in experimentation

The correlation coefficients between the students' pre-knowledge and the three dimensions in experimentation in booklet 111 ranged from 0.020 to 0.223. In booklet 131, the correlation coefficients ranged from 0.110 to 0.264. The correlation coefficient between the pre-knowledge and the dimension "data analysis" in booklet 131 was the highest and significant.

Discussion

The correlation coefficients between the students' pre-knowledge and the three dimensions in experimentation were low and there was quite some variation between the two booklets. In booklet 111, the correlation coefficient between the students' pre-knowledge and the dimension "search in the experiment space" was higher than in the other two combinations. In contrast, in booklet 131, the correlation between the pre-knowledge and the dimension "data analysis" was the highest. These findings contradict our expectations. We expected high correlations, especially between the students' pre-knowledge and the dimensions "search in the hypothesis space" and "data analysis" because these two dimensions were hypothesized to be influenced by the biological pre-knowledge rather than by methodological knowledge.

Possible reasons for these findings are that the number of students in this test was not high enough and that many students had missing items or missing units. This is the case for 11 students out of 73, who had had two missing units. Ten students out of 73 had missing items in booklet 111 in the knowledge test and 15 students out of 73 had missing items in the competency test.

2.4.2.2. Correlations (Spearman) between the three dimensions in experimentation

Booklet	Search in the hypothesis space * Data analysis	Data analysis * Search in the experiment space	Search in the hypothesis space * Search in the experiment space
111	0.716**	0.623**	0.589**
131	0.666**	0.274**	0.323**

Table 3.46: Correlation coefficients (Spearman) between the three dimensions in experimentation

The correlation coefficients for the relationship between the three dimensions in experimentation ranged from 0.589 to 0.716 in booklet 111 and from 0.274 to 0.666 in booklet 131. The correlation coefficients for the relationship between "search in the hypothesis space" and "data analysis" were the highest in both booklets.

On the other hand, in booklet 111 the correlation coefficients for the relationships between "search in the hypothesis space" and "search in the experiment space" and between "data analysis" and "search in the experiment space" were medium. However, in booklet 131, they were low in these two combinations.

Discussion

We expected different interactions between the three dimensions in experimentation. In particular, we assumed that there are higher correlations between the dimensions “search in the hypothesis space” and “data analysis” than between the dimensions “search in the hypothesis space” and “search in the experiment space”. We also assumed that there are higher correlations between the dimensions “search in the hypothesis space” and “data analysis” than between the dimensions “data analysis” and “search in the experiment space”. These hypotheses were based on the assumption that the dimensions “search in the hypothesis space” and “data analysis” are driven by the students’ pre-knowledge about the science contents of the experiment while the dimension “search in the experiment space” should prove more dependent on the students’ methodological knowledge. However, correlation statistics did not allow us to substantiate any of our hypotheses, because in booklet 131 the correlation between “search in the hypothesis space” and “data analysis” was much higher than in two other combinations. On the contrary, in booklet 111, though the correlation between “search in the hypothesis space” and “data analysis” was also the highest, correlation statistics suggested that the correlations between the three dimensions are quite high. This is not an unusual finding, however, as high achieving students possess high content knowledge as well as high methodological knowledge and do well in all three dimensions. The opposite is true for low achieving students. This accounts for the fairly high level of the correlation coefficients. The difference between students in class 2 and 3 however, can be regarded as evidence for the fact that there were students in our sample who differed in exactly the dimensions we hypothesized, i.e. “search in the experiment space”.

3. Conclusion

Test development was successful for the competency test. Reliabilities for the three scales are higher, for example, than for the national PISA 2003 science test. Especially, the reliability at the booklet level for booklet 111 was very high, where the Cronbach's alpha was 0.89. This booklet should be chosen for the next test.

Comparing the two versions of the competency test from a psychometrical perspective, version 1 of the competency test needs to be favoured over version 2 of the competency test for students at grade 5-6.

Moreover, both scoring models proved to be useful, though no added value was found for the partial credit scoring model.

However, work needs to be done on item development for the knowledge test, where the reliability at the unit level was not reliable enough.

Chapter 4: Pre-test 2

In September 2005, pre-test 2 was done. In this pre-test, we only tested knowledge. Prior to the test, knowledge test items had been revised with the intention to develop a test reliable enough at the unit level so that it can be used in the main study to investigate correlations between knowledge about the topic of a unit and competencies in experimentation. This was done, mainly, by adding additional test items to the existing test, thus extending scale length. No changes were made to the answering format which was complex multiple choices.

1. Method

Sample

The participants in this study were 122 sixth grade students, who came from five classes from three schools, one Gymnasium, one Hauptschule and one Realschule in Germany. The students' age ranged from 9 years and 11 months to 11 years and 6 months. The average age for the sample was 12 years and 1 month.

Design

In this pre-test, we only tested knowledge. The test is similar to the knowledge test described in chapter 2 (pre-test 1), however, some more new items were tested. Again, complex multiple-choice questions were used, with items that consist of four questions each. Four units with different content knowledge were used: Unit 1: Seed germination, Unit 2: Chicken eggs, Unit 3: Apple wine, and Unit 4: Baking bread. These units were chosen because they belong to booklet 111 which had high reliabilities in pre-test 1. Each unit had 8 questions. Thus, each student had to answer 32 questions.

Each student was given 35 minutes to do the test.

Answering format

As in pre-test 1, complex multiple-choice questions were used. Each question (item) had 4 to 6 sub-questions. The answering format was “yes” or “no”.

2. Findings

2.1. Item difficulty

2.1.1. Method

Mean item difficulties for the items in each unit, as well as the mean item difficulty for the booklet were calculated after the items had been scored as correct or incorrect. In complex multiple choice items, this entails determining the cut-off before an item can be scored.

2.1.2. Findings

Mean item difficulty for the items and for the units

	Unit 1	Unit 2	Unit 3	Unit 4
Item 1	70.2	73.9	80.5	81.2
Item 2	88.4	74.3	68.7	83.3
Item 3	70.6	81.2	77.8	73.3
Item 4	70.0	69.8	66.6	69.9
Item 5	67.1	68.7	57.0	63.5
Item 6	57.9	79.5	68.3	56.0
Item 7	60.6	87.6	52.0	62.7
Item 8	73.9	81.4	72.6	78.7
Mean item difficulty for unit	69.8	77.1	67.9	71.1

Table 4.1: Mean item difficulty for the items and for the units

The mean item difficulty for all items was quite low. Most items were solved correctly by 50% to 80% of the students. However, some items were correctly answered with a probability of over 80%, for example, item 2 of unit 1 or item 7 of unit 2.

The mean item difficulty for all four units was also low. It ranged from 67% to 77%.

The mean item difficulty for the booklet was 71.5%.

Discussion

The mean item difficulty for the items in this pre-test was very low. In particular, some items were correctly answered by nearly 90% of the students (in unit 1, unit 2). However, the mean item difficulties for the units and the booklet were acceptable and indicate that the test is at an appropriate level for students in sixth grade.

Therefore, the item difficulty for the test in general was low; students in sixth grade could do the test very well.

2.2. Reliability

2.2.1. Method

To calculate the reliability at the unit and at the booklet level, we summed all sub-questions for each item in all four units, considered the cut of point and scored each item. Then, we calculated the reliability coefficient for each unit and booklet and calculated the corrected item-total correlation for each item. If an item had a corrected item-total correlation lower than 0.2, it was deleted.

2.2.2. Findings

2.2.2.1. Reliability at the unit level

Unit 1: Seed germination

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (Item selection)
1	0.23		0.22	
2	0.33		0.30	
3	0.06		0.19	
4	-0.03	0.38		0.48
5	0.31		0.36	
6	0.10		0.16	
7	0.05			
8	0.28		0.22	

Table 4.2: Reliability of Unit 1: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 1 was 0.38. Many items had a corrected item-total correlation lower than 0.2. After deleting the non-discriminating items, the highest Cronbach's alpha was 0.48.

Unit 2: Chicken eggs

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (Item selection)
1	0.16			
2	- 0.01			
3	0.36		0.35	
4	- 0.06	0.44		0.58
5	0.28		0.37	
6	0.19		0.23	
7	0.41		0.42	
8	0.27		0.32	

Table 4.3: Reliability of Unit 2: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 2 was 0.44. Like in unit 1, many items had a corrected item-total correlation lower than 0.2. After deleting the non-discriminating items, the Cronbach's alpha was 0.58.

Unit 3: Apple wine

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (Item selection)
1	-0.21		0.24	
2	-0.06		0.29	
3	0.20		0.29	
4	-0.09	-0.29		0.45
5	-0.28			
6	-0.06			
7	-0.17			
8	-0.01			

Table 4.4: Reliability of Unit 3: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 3 was very low. The Cronbach's alpha was -0.29. All items had a corrected item-total correlation lower than 0.2. After deleting some non-discriminating items, this unit had only three items left and the highest Cronbach's alpha was 0.45.

Unit 4: Baking bread

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (Item selection)
1	0.29		0.26	
2	0.04		0.22	
3	0.08			
4	0.15	0.07	0.25	0.41
5	-0.30			
6	-0.04			
7	0.11		0.19	
8	-0.02			

Table 4.5: Reliability of Unit 4: Cronbach's alpha and corrected item-total correlation

Like in the three above units, the reliability at the unit level for unit 4 was low. The Cronbach's alpha was 0.07 and only one item had a corrected item-total correlation higher than 0.2. After deleting the non-discriminating items, the highest Cronbach's alpha was 0.41.

Discussion

The reliability at the unit level was very low in all four units. All reliability coefficients were lower than 0.6; and many items had a corrected item-total correlation lower than 0.2. This indicates that the knowledge test at the unit level was not reliable.

2.2.2.2. Reliability at the booklet level

	Items	Corrected item-total correlation (all items)	Corrected item-total correlation (items selection)
Unit 1: Seed germination	1	0.26	0.22
	2	0.28	0.29
	3	0.15	0.32
	4	0.14	
	5	0.30	0.33
	6	0.20	0.29
	7	0.11	
	8	0.28	0.32
Unit 2: Chicken eggs	1	0.24	
	2	0.03	
	3	0.35	0.40
	4	0.14	
	5	0.43	0.52
	6	0.17	0.24
	7	0.36	0.45
	8	0.31	0.34
Unit 3: Apple wine	1	-0.06	
	2	0.03	
	3	0.25	
	4	0.24	0.30
	5	0.01	
	6	0.19	0.24
	7	-0.10	
	8	-0.06	
Unit 4: Baking bread	1	0.34	0.20
	2	0.05	
	3	0.06	
	4	0.11	
	5	0.04	
	6	-0.09	
	7	0.13	
	8	0.20	0.31

Cronbach's alpha 1 = 0.59

Cronbach's alpha 2 = 0.71

Table 4.6: Reliability at the booklet level: Cronbach's alpha and corrected item-total correlation

The reliability coefficient at the booklet level was 0.59. Furthermore, many items had a corrected item-total correlation lower than 0.2. After deleting the non-discriminating items, the highest Cronbach's alpha was 0.71 and fifteen strong items were left.

Discussion

Although the attempt was made to extend the scales of the knowledge test, the reliability of the test at the unit level was very low. All Cronbach's alphas for the units were lower than 0.6. On the other hand, many items were weak with a corrected item-total correlation lower than 0.2. However, at the booklet level the reliability after item selection was higher than 0.7. The remaining fifteen items can be used to measure knowledge with satisfactory reliability.

	Cronbach's alpha for units in pre-test 1	Cronbach's alpha for units in pre-test 2
Unit 1. Seed germination	0.65	0.48
Unit 2. Chicken eggs	0.43	0.58
Unit 3. Apple wine	0.57	0.45
Unit 4. Baking bread	0.59	0.41

Table 4.7: Reliability coefficient (Cronbach's alpha) at the unit level in pre-test 1 and pre-test 2

Comparing the reliability coefficient at the unit level between pre-test 1 and pre-test 2, table 4.7 shows that the reliability at the unit level in pre-test 2 was only slightly lower than in pre-test 1 in three of the four units. The reliability at the unit level in both pre-tests was also lower than 0.7. Thus, all Cronbach's alphas at the unit level were not reliable.

However, at the booklet level, the knowledge test was similarly reliable in both pre-tests as the Cronbach's alphas were 0.74 in pre-test one and 0.71 in pre-test two. However, in both pre-tests the number of discriminating items at the booklet level was only fifteen items.

In sum, although the knowledge test in pre-test 2 had more items than in pre-test 1, the reliability coefficient at the unit level and at the booklet level in this pre-test was still not higher than in pre-test 1.

3. Conclusion

In this pre-test, we developed some more items for all units so that each unit consisted of 8 items. However, the item difficulty for the test was also low. The number of students who answered correctly in most items ranged from 50% to 80%. Four units were correctly answered with a probability of over 67%, and the mean item difficulty for booklet was 71.5%. Thus, this test was appropriate for students in sixth grade.

However, the reliability at the unit level was very low. All Cronbach's alphas were lower than 0.6 and many items had a corrected item-total correlation lower than 0.2. Especially in unit 3, all items had low discriminating power. Compared with pre-test 1, the reliability at the unit level in this pre-test was slightly lower.

However, the reliability coefficient at the booklet level was higher than 0.7, but many items had a corrected item-total correlation lower than 0.2. Only 15 out of 32 items correlated well with the corrected sum score.

Because item development failed to improve test quality, another possibility of item development not used so far can be tried – change of the answering format to simple multiple choice.

Chapter 5: Pre-test 3

1. Method

Sample

This pre-test was carried out in July 2006.

The participants of this study were 77 students (30 girls and 47 boys) of sixth grade, who came from a high school (Gymnasium) in Kiel, Germany. Their age ranged from 10 years and 3 months to 13 years and 8 months and the mean age for the sample was 12 years and 5 month.

Design

Each student worked on two types of test, the competency test and the knowledge test. If the students finished early, they were asked to answer opinion questions (cf.2.1.2.1 b).

In this pre-test, booklet 111 in the competency test was used (cf. Chapter 2 for a description of the design of the test). This booklet contained 4 units: Unit 1: Seed germination; Unit 2: Chicken eggs; Unit 3: Apple wine and Unit 4: Baking bread.

In the knowledge test, simple multiple choice items were used. Each question consisted of 4 options, one correct answer and three distractors. Each unit had 7 to 10 questions. Thus, the knowledge test, consisted of 35 questions.

Sample Item

Which statement is correct?

- a) Seeds need soil to germinate.
- b) Seeds must take up nutrients to germinate.
- c) Seed germinate faster when they are fertilized.
- d) Seeds can germinate in the dark.

Figure 5.1: Sample item of the knowledge test

In the opinion test, we asked students how important they considered specific variables used in the items of the competency test. By means of this test we wanted to investigate students' beliefs about the importance of factors that may be responsible for the phenomena presented. One of these questions was a complex multiple choice question (see appendix, p. 263).

Each student was given 45 minutes. The time for the knowledge test was 23 minutes; for the competency test 22 minutes.

Answering format

In the knowledge test, simple multiple-choice questions were used. Each question had four choices, one of which was the correct answer.

The competency test had the same answering format as the test used in pre-test 1.

In the opinion test, we asked students to assess how important specific factors are on a Likert scale with the following categories: very important, important, not very important and unimportant. The students were asked to tick the category they most agreed with. In the other questions, a complex multiple choice was used. Each question had four or five sub-questions. The answering format was “yes” or “no”.

Coding the answers

In the knowledge test, we used a scoring model that distinguished between “full credit” and “no credit”. If the students chose the correct answer, their response was scored 1. If the students chose an incorrect answer, they received a 0 score. The Likert scale for the opinion questions was coded in the following way: very important: 4, important: 3, not very important: 2 and unimportant: 1. In the complex multiple-choice questions, each correct answer was coded 1. After coding the sub-questions, we summed the scores for each question, determined the cut-off point and scored the item again.

In the competency test, we also used two scoring models (0/1 and 0/1/2) as well as in pre-test 1.

2. Findings

2.1. Item difficulty

2.1.1. Method

In both the knowledge test and the competency test, we calculated the percentage of correctly solved items, mean item difficulties for the items of a unit and mean item difficulties for the items of a booklet.

Besides, in the competency test we also calculated the mean item difficulties for each dimension of experimentation.

2.1.2. Findings

2.1.2.1. Item difficulty in the knowledge test

a) Item difficulty for the items and the mean item difficulty for the units

Items	Seed germination	Chicken eggs	Apple wine	Baking bread
1	80.5	93.3	86.8	90.5
2	89.3	28.0	26.3	94.6
3	93.5	70.7	75.3	62.9
4	27.4	81.1	61.4	91.5
5	28.6	92.2	6.5	40.0
6	34.2	76.0	78.9	15.3
7	18.9	36.8	53.3	79.2
8		80.3	70.3	57.5
9		34.2	15.6	8.8
10		29.7		
Mean item difficulty for unit	53.2	62.2	52.7	60.0

Table 5.1: Item difficulties for the items and the mean item difficulties for the units in the knowledge test

The item difficulty for most items in the knowledge test ranged from 20% to 80%. However, some items were correctly answered with a probability of over 80%, some even over 90%. For example, item 3 (unit 1), item 1 and 5 (unit 2), and items 1, 2 and 4 (unit 4) were solved by more than 90% of the students.

On the other hand, some items had a rather high item difficulty, for example, item 5 (unit 3) with 6.5% and item 9 (unit 4) with 8.8%.

However, the mean item difficulty for the unit ranged from 52% to 62%. The mean item difficulty for the booklet was 57%.

Discussion

The item difficulty for items in the knowledge test varied considerably between the items of different units and even between items in the same unit.. The reasons for this difference may be that the contents of some items was familiar to the students while the contents of another item may have been unfamiliar. Also, possibly some items were more difficult to answer because these items were inconsistent with the students' prior beliefs and conceptions. The very easy or very difficult items must be looked at closely and revised for the main study.

The mean item difficulty for the units ranged from 52 % to 62%. In pre-test 1, the mean item difficulty for the unit ranged from 56% to 78% and the mean item difficulty for booklet 111 in pre-test 1 was 66.1%. Thus, the item difficulty for the units and for the

booklet was also higher in pre-test 1 than in this pre-test, but still acceptable for students in grade 6.

b) Assessment of students' beliefs about the importance of variables that are relevant / irrelevant for the phenomenon investigated in the experiments

Unit 1: Seed germination

Water

In the question “What happens when seeds germinate?” only 27.4% of the students (20 students out of 74) answered correctly that seeds absorb water, while 50.7% answered that seeds absorb nutrients. 20.5% said that seeds absorb light. However, 87% of the students indicated that water is very important for seed germination and only 2.8% said that water is not important. Furthermore, 95.7% of the students confirmed that one has to water dry seeds in order to make them germinate.

Soil

33.8% of the students (24 students out of 71) answered that soil is very important or important for seed germination. 32.4% of the students said that soil is not important. However, 87.3% of the students indicated that seeds can germinate in materials other than soil.

Light

Only 18.9% of the students (14 students out of 74) indicated that bean seeds can germinate in the dark. 87.3% of the students (62 students out of 71) confirmed that light is very important or important for seed germination. Only 4.2% of the students said that light is not important for seed germination.

Temperature

49.3% of the students (35 students out of 71) confirmed that one has to put seeds in a warm place to facilitate germination. Moreover, 73.2% of the students indicated that the temperature is very important or important for seed germination. Only 9.7% said that the temperature is not important.

Variable	Light	Warmth	Air	Soil	Water
Mean score	3.41	2.76	2.89	2.17	3.82
STD	0.82	0.81	0.94	1.05	0.56

Table 5.2: Means and standard deviations (STD) for the variables in Unit 1: seed germination (4: very important, 3: important, 2: not very important and 1: unimportant).

Table 5.2 shows the students' beliefs about the importance of the variables in bean seed germination. In particular, the students believe that water is the most important factor for seed germination (M 3.82, STD 0.56). For the students, the second most important factor is light (M 3.41, STD 0.82). The temperature and air are also considered important. Soil is the least important factor for the students (M 2.17, STD 1.05).

Unit 2: Chicken eggs

Temperature

79.8% of the students (56 students out of 71) indicated that the temperature is very important for hatching chicken eggs. Furthermore, 94% of the students confirmed that chicken eggs need high temperatures to hatch. Besides, 76% of the students knew that it is the best to set the temperature in the incubator to 38°C, if one wants to hatch eggs.

Humidity

70.4% of the students (50 students out of 71) said that the humidity was very important or important for hatching chicken eggs. 67% of the students indicated their disagreement with the statement: "Chicken eggs hatch in humid air as fast as in dry air." So, most of students believe that chicken eggs need humid air to hatch.

Size of eggs

Only 12.7% of the students (9 students out of 71) indicated that they believe that the size of the eggs is very important or important for hatching eggs, while 60.6 % of the students said that it is not important. In addition, 64% of the students indicated that small eggs hatch as fast as big eggs. These students knew that the size of the eggs does not influence the hatching of chicken eggs.

Variable	Light	Warmth	Humid air	Size of eggs
Mean score	2.06	3.77	2.97	1.54
STD	0.99	0.45	1.11	0.75

Table 5.3: Means and standard deviations (STD) for the variables in Unit 2: chicken eggs (4: very important, 3: important, 2: not very important and 1: unimportant).

Table 5.3 shows that the students believe that the temperature is the most important factor for hatching chicken eggs (M 3.77, STD 0.45). According to the students, the second most important factor that influences the hatching of eggs is the humidity (M 2.97, STD 1.10). Light was not considered very important (M 2.06, STD 0.99) and the size of the eggs was considered unimportant for hatching chicken eggs (M 1.54, STD 0.75).

Unit 3: Apple wine

Temperature

Only 12.9% of the students (9 students out of 71) said that high temperatures are very important and 38.6% of the students said that high temperatures are important for making wine. 21% of the students indicated that high temperatures are not important. However, 63% confirmed that the temperature influences the success of making wine. So, roughly two thirds of the students believe that the temperature has an influence on making wine.

Amount of sugar

84.5% of the students (60 students out of 71) indicated that they believe that the amount of sugar is very important for making wine. However, the question “When does wine have much alcohol?” was answered correctly by only 6.5% of the students. 79.2% of the students believe that it is the duration of storing wine – not the amount of sugar in the grape juice – that has an effect on the amount of alcohol in the wine.

Yeast

61.4% of the students (43 students out of 70) indicated that yeast was important for making wine. However, in the question “What is yeast?” only 15.3% of the students chose the answer “Yeast is a creature”. In contrast, 20% of the students gave the answer: “Yeast is an enzyme”.

Construction of the wine-making vessel

Only 15.5% of the students (11 students out of 71) believed that one should use a specific wine-making vessel that allows gases to escape, but prevents air from the outside to enter the vessel. In contrast, 72.7% of the students believed that one should use a completely closed container. Moreover, 26.3% of the students knew that oxygen damages the process of making wine.

Variable	Light	Warmth	Good pot without air coming in	Amount of sugar
Mean score	1.33	2.37	3.75	3.18
STD	0.63	1.02	0.63	0.80

Table 5.4: Means and standard deviations (STD) for the variables in Unit 3: making wine (4: very important, 3: important, 2: not very important and 1: unimportant).

For Unit 3 “Making wine”, the students believed that the specific vessel that prevents air from coming in was the most important factor (M 3.75, STD 0.62). The second important factor, according to the students, was the amount of sugar (M 3.18, STD 0.80). Most of the students think that high temperatures were important (M 2.37, STD 1.02) but not as important as the amount of sugar and the use of a specific wine-making vessel. The rather large standard deviation for the mean for this item indicates that there was quite some variation among the students concerning this question. Light was considered the least important factor for making wine (M 1.33, STD 0.63).

Unit 4: Baking bread

Temperature of water

85.9% of the students (61 students out of 71) indicated that the temperature of the water was either very important or important for making bread. Moreover, 62.9% of the students said that one should not use boiling hot water (100°C) to mix yeast dough, because it damages the yeast. Besides, 79.2% of the students confirmed that the water temperature affects the growth of the yeast. In addition, 40% of the students knew that a water temperature of below 40°C was appropriate for making bread.

Yeast and baking powder

91.5% of the students (65 students out of 71) indicated that they believe that yeast is either very important or important for making bread. Furthermore, 94.6% of the students indicated that one could use baking powder instead of yeast to make bread. In addition, 76% of the students answered that the bread will be hard if one forgets the yeast when mixing dough.

Flour

81.7% of the students (58 students out of 71) said that flour was either very important or important, so that yeast dough rises.

Most of students believed that flour was important for making bread soft and airy.

Butter

Over 70% of the students (50 students out of 71) believe that butter influences the rising of yeast dough and it is important for baking bread. However, butter does not influence the rising of the yeast dough.

Variable	Yeast	Flour	Sugar	Butter	Temperature of water
Mean score	3.70	3.11	2.45	2.86	3.27
STD	0.67	0.87	0.97	1.05	0.91

Table 5.5: Mean scores and standard deviations (STD) for variables in Unit 4: Baking bread (4: very important, 3: important, 2: not very important and 1: unimportant).

In the Unit “Baking bread”, the students believed that yeast was the most important factor (M 3.70, STD 0.67). Besides, the temperature of the water and the use of flour were also considered important for yeast dough rising (M 3.27, STD 0.91). The use of butter was considered important (M 2.86, STD 1.05). The rather large standard deviation for this item shows that there was quite some variation in student thinking about the importance of butter. Sugar was considered the least important factor for baking bread (M 2.45, STD 0.97).

2.1.2.2. Item difficulty in the competency test

a) Item difficulty for items

Unit	Search in the hypothesis space		Data analysis		Search in the experiment space	
	H1	H2	D1	D2	E1	E2
Seed germination	55.3	72.0	57.1	79.2	82.7	72.4
Chicken eggs	74.3	74.3	56.8	54.7	82.9	66.2
Apple wine	78.9	79.7	83.8	83.8	72.0	68.9
Baking bread	91.9	89.3	86.5	86.7	75.7	72.0

Table 5.6: Item difficulty for items in the competency test

The item difficulty for the items in the competency test ranged from 54.7% to 91.9%. Most of the items had an item difficulty that stayed within the range from 54% to 80%. However, a third of the items were correctly answered with a probability of over 80%, such as items E1 of unit 1, item E1 of unit 2, items D1 and D2 of unit 3 and four items of unit 4. In particular, in Unit 4: Baking bread item 1 in the dimension “search in the hypothesis space” had an item difficulty of 91%.

Discussion

At the item level, the highest item difficulty was 54.7% (item 4 of unit 2). 8 out of 24 items were correctly answered with a probability of over 80%. In contrast, the item difficulty for items in the competency test in pre-test 1 ranged from 32% to 70%. Only one item was correctly answered with a probability of 80%.

This means that in this pre-test some items were too easy for the students. A possible reason for this is that the students who took this pre-test were at the end of grade 6 and may have possessed more knowledge than those students who took pre-test 1 at the beginning of the second semester of grade 5 and grade 6.

As a consequence of these findings, item revision is necessary. Some specific recommendations for item revision can be derived from this study: In Unit 1 “Seed germination,” 82.7% of the students solved the task that required assessing the importance of light for seed germination. The item belongs to the competency test and assesses competencies in the dimension “search in the experiment space”. Most of the students (87.3%) believed that light was important for germination. The variable light was consistent with the students’ belief. Therefore, the percentage of students who solved this task was very high.

In Unit 2 “Chicken eggs,” 82.9% of the students solved the item that required assessing the effect of the size of eggs on the time necessary to hatch eggs. The item belongs to the competency test and assesses competencies in the dimension “search in the experiment space”.

This variable was also consistent with the students’ thinking. The number of students who solved this item was very high. Another reason may be that the design of the item made it easy for those students (71.4%), who knew that the humidity affects the hatching of eggs, to identify the correct incubator. There were only two incubators with humid air, whereas the other two had dry air, so that it may have been an easy task to control variables when making a choice between the two options.

In Unit 3 “Apple wine,” 78.9% of the students solved the item about the temperature in the dimension “search in the hypothesis space” and 83% of the students solved the task in “data analysis”. In addition, 79.7% of the students solved the item about the amount of sugar in the dimension “search in the hypothesis space”. 83.8% of the students solved the item in the dimension “data analysis”.

These percentages are too high to be acceptable by the standards of item development. The reason for these high percentages may be that the variables in these items were not

specific for the students. For example, the variable “apple juice” remained unchanged in all items and did not present any problems to the students who were asked to control variables or interpret the effect of a variable. On the other hand, the students did not know about the importance of the variable “air can go into the pot or not”, because in the knowledge test only 15.5% of the students answered that one should use the pot that allows air to go out but not to go in, while 72.7% of the students said that one should use a completely closed pot.

In Unit 4 “Baking bread,” 91% of the students solved the task about the temperature of water in the dimension “search in the hypothesis space” and 86.5% of the students solved the item in the dimension “data analysis”. Moreover, the number of students who solved the task about yeast in the dimension “search in the hypothesis space” was 89.3% and in the dimension “data analysis” it was 86.7%. These tasks were completely consistent with students’ thinking, so they could solve them easily.

b) Mean item difficulty for the units

	Unit 1	Unit 2	Unit 3	Unit 4
Mean item difficulty	69.8	68.2	77.9	83.7

Table 5.7: Mean item difficulties for the units in the competency test

The mean item difficulties for the units in the competency test ranged from 68% to 83%. Among them, unit 4 had the lowest item difficulty.

c) Mean item difficulty for dimensions and booklet

	Search in the hypothesis space	Data analysis	Search in the experiment space	Booklet
Mean item difficulty	77.0	73.6	74.1	74.9

Table 5.8: Mean item difficulties for the three dimensions of the competency test and for the booklet

The mean item difficulty for the three dimensions in experimentation ranged from 73% to 77%. The item difficulty for the booklet was 74.9%.

Discussion

The mean item difficulty for three units ranged from 68% to 77%. However, the mean item difficulty for unit 4 was 83.7%. In pre-test 1, the mean item difficulty was much higher than in this pre-test, where the same test booklet had been tested. Specifically, the mean item difficulty for units in booklet 111 in the pre-test 1 ranged from 45% to 61.7%.

Furthermore, the mean item difficulty for all three dimensions was higher than 73%. In pre-test 1, the item difficulty for the three dimensions ranged from 49% to 53%.

The mean item difficulty for the complete booklet in this pre-test was 74.9%. It was also much higher than in pre-test 1, where it was 57.1%.

So, the item difficulty at all levels in this pre-test was lower than that in pre-test 1. The reason for this might be that the students in this test had more content knowledge than those in pre-test 1, because the students in the pre-test 1 were at the beginning of fifth or sixth grade and those for this pre-test were at the end of sixth grade.

2.2. Reliability

2.2.1. Method

In both the knowledge test and the competency test, we calculated the reliability coefficient at the level of the unit and at the level of the booklet level based on the corrected item-total correlation for each item. If any item had a corrected item-total correlation lower than 0.2, it was taken out.

Besides, in the competency test, we also calculated the reliability coefficient for the scales “forming hypotheses”, “data analysis” and “planning experiments” as well as for the complete booklet.

2.2.2. Findings

2.2.2.1. Reliability of the knowledge test

a) Reliability of the knowledge test at the unit level

Unit 1: Seed germination (n = 252)

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (item selection)
1	0.03	0.25		0.37
2	0.03			
3	-0.10			
4	0.16		0.19	
5	0.12		0.19	
6	0.05			
7	0.40		0.25	

Table 5.9: Reliability of the knowledge test, Unit 1: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 1 was 0.25. Also, many items had a corrected item-total correlation coefficient lower than 0.2. After deleting the non-discriminating items, there were only three items left and the highest Cronbach's alpha was 0.37.

Unit 2: Chicken eggs

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (item selection)
1	-0.06	0.36		0.45
2	0.18		0.20	
3	0.23		0.26	
4	0.02			
5	0.20		0.19	
6	0.24		0.19	
7	-0.03			
8	0.33		0.28	
9	0.20		0.25	
10	0.07			

Table 5.10: Reliability of the knowledge test, Unit 2: Cronbach's alpha and corrected item-total correlation

As in unit 1, the reliability coefficient for unit 2 was also low. The Cronbach's alpha was 0.36. Also, many items had a corrected item-total correlation coefficient lower than 0.2. After deleting the non-discriminating items, there were five items left and the Cronbach's alpha for the item selection was 0.45.

Unit 3: Apple wine

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (item selection)
1	-0.01			
2	0.26		0.44	
3	0.34		0.40	
4	0.18	0.44	0.25	0.58
5	-0.20			
6	0.28		0.27	
7	0.47		0.33	
8	0.18			
9	0.00			

Table 5.11: Reliability of the knowledge test, Unit 3: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 3 was 0.44. Some items had a corrected item-total correlation coefficient lower than 0.2. After deleting the non-discriminating items, there were only five items left and the highest Cronbach's alpha was 0.58.

Unit 4: Baking bread

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach's alpha (item selection)
1	-0.11			
2	-0.02			
3	0.17		0.26	
4	-0.06	-0.07		0.41
5	-0.03			
6	-0.20			
7	0.01			
8	0.02		0.26	
9	-0.03			

Table 5.12: Reliability of the knowledge test, Unit 4: Cronbach's alpha and corrected item-total correlation

The reliability coefficient at the unit level for unit 4 was very low. The Cronbach's alpha was -0.07. All items had a corrected item-total correlation coefficient lower than 0.2. After deleting the nondiscriminating items, there were only two items left and the Cronbach's alpha was 0.41.

Cronbach's alpha (item selection)	
Unit 1: Seed germination	0.37
Unit 2: Chicken eggs	0.45
Unit 3: Apple wine	0.58
Unit 4: Baking bread	0.41

Table 5.13: Reliability coefficients (Cronbach's alpha) for all units of the knowledge test

Discussion

The reliability at the unit level for all four units in this pre-test was very low. The reliability coefficients ranged from 0.37 to 0.58. Furthermore, many items had a corrected item-total correlation coefficient lower than 0.2. After deleting the non-discriminating items, there were only a few strong items left in each unit. In particular, unit 4 had only two strong items left, while unit 1 had three items left. Therefore, the knowledge test at the unit level cannot be considered reliable.

b) Reliability at booklet level

	Item	Corrected item-total correlation (all items)	Corrected item-total correlation (item selection)
Unit 1: Seed germination	1	0.27	0.23
	2	0.12	0.20
	3	0.12	
	4	0.15	0.16
	5	0.20	0.18
	6	-0.08	
	7	0.30	0.27
Unit 2: Chicken eggs	1	-0.20	
	2	0.24	0.25
	3	0.23	0.28
	4	0.13	
	5	0.26	0.24
	6	0.28	0.36
	7	0.15	0.20
	8	0.45	0.44
	9	0.30	0.34
	10	0.23	
Unit 3: Apple wine	1	-0.11	
	2	0.29	0.39
	3	0.43	0.43
	4	0.29	0.43
	5	-0.17	
	6	0.38	0.31
	7	0.32	0.24
	8	0.16	0.15
	9	-0.11	

Unit 4: Baking bread	1	0.08	
	2	-0.01	
	3	0.27	0.35
	4	0.16	0.20
	5	-0.15	
	6	0.03	
	7	-0.02	
	8	0.45	0.53
	9	0.03	
Cronbach's alpha		0.62	0.74

Table 5.14: Reliability of the knowledge test at the booklet level before and after item selection: Cronbach's alpha and corrected item-total correlation

The reliability coefficient at the booklet level for all items (35 items) was 0.62 before item selection. Many items in all four units had a corrected item-total correlation coefficient lower than 0.2. After taking the low items out, the Cronbach's alpha for the booklet was 0.74. There were 21 strong items left in the booklet. Unit 1 had 5 items left, unit 2 had 7 items, unit 3 had 6 items and unit 4 had only 3 items left.

Discussion

The reliability coefficients at the unit level in the knowledge test were low. The highest Cronbach's alpha was 0.58 and many items had a corrected item-total correlation coefficient lower than 0.2.

A comparison of the reliability coefficient at the unit level in the knowledge test for all three pre-tests, table 5.15 shows that the Cronbach's alphas for the units that were tested in pre-test 1 were slightly higher than those tested in the other two pre-tests. Besides, all Cronbach's alphas at the unit level in the knowledge test were lower than 0.7. This means that the knowledge test at the unit level in all three pre-tests was not reliable despite the efforts that were made to increase the reliability of the test. The final change in the answering format (from complex multiple choice to simple multiple choice) brought the reliability down a little bit which can be traced back either to the fact that either there were differences between the samples or to the fact that in complex multiple choice items it is possible to adjust the difficulty level of the items to the specific sample, whereas this is not possible when simple multiple choice items are used.

Unit	Cronbach's alpha at unit level in pre-test 1	Cronbach's alpha at unit level in pre-test 2	Cronbach's alpha at unit level in pre-test 3
1. Seed germination	0.65	0.48	0.37
2. Chicken eggs	0.43	0.58	0.45
3. Apple wine	0.57	0.45	0.58
4. Baking bread	0.59	0.41	0.41

Table 5.15: Reliability Cronbach's alpha at the unit level in the knowledge test in pre-test 1, pre-test 2 and pre-test 3

However, the reliability coefficients for the three knowledge tests (pre-tests 1-3) were higher than 0.7 at the level of the booklet. In particular, the Cronbach's alpha of pre-test 1 was 0.74. For pre-test 2, it was 0.71 and for pre-test 3 it was 0.74.

Therefore, the knowledge test at the booklet level in all three tests is reliable and it is possible to use the knowledge test in order to calculate the relationships between the pre-knowledge and the three dimensions in experimentation. For the main study, described in the following chapter, it is possible to use any of the three knowledge tests. However, the complex multiple choice questions were more difficult to handle for the students than the simple multiple choice questions because each sub-question had to be assessed. Accordingly, pre-test 3 was chosen for further use in the main study.

2.2.2.2. Reliability of the competency test

a) Reliability at the unit level

	Unit Seed germination	Unit Chicken eggs	Unit Apple wine	Unit Baking bread
Item H1	0.58	0.50	0.78	0.53
Item H2	0.29	0.43	0.75	0.48
Item D1	0.53	0.52	0.32	0.57
Item D2	0.21	0.24	0.72	0.47
Item E1	0.42	0.26	0.65	0.61
Item E2	0.49	0.44	0.58	0.38
Cronbach's alpha for unit	0.69	0.66	0.84	0.75

Table 5.16: Reliability at the unit level: Corrected item-total correlation coefficients and Cronbach's alpha

The reliability coefficients at the unit level in the competency test ranged from 0.66 to 0.84. All items had a corrected item-total correlation coefficient higher than 0.2. Many items had a corrected item-total correlation coefficient higher than 0.5. Unit 3, for

example, had five items out of six with a corrected item-total correlation coefficient higher than 0.5

Discussion

The reliability coefficient at the unit level for all four units was quite high. In particular, the Cronbach's alpha for Unit 3 "Apple wine" was 0.84; and for Unit 4 "Baking bread" it was 0.75. Furthermore, all items had a corrected item-total correlation coefficient higher than 0.2.

A comparison of the reliability coefficients for the units of the competency test in pre-test 1 and pre-test 3 shows that there are higher reliability coefficients for the units in pre-test 3 than for the units in pre-test 1 (cf. table 5.17). An exception is the reliability coefficient for unit 4.

Unit	Cronbach's alpha at unit level in pre-test 1	Cronbach's alpha at unit level in pre-test 3
1. Seed germination	0.61	0.69
2. Chicken eggs	0.61	0.66
3. Apple wine	0.75	0.84
4. Baking bread	0.75	0.75

Table 5.17: Reliability coefficient (Cronbach's alpha) at the unit level in pre-test 1 and in pre-test 3

b) The reliability of the scales "forming hypotheses", "data analysis" and "planning experiments" at the booklet level

Unit	Item	Scale "forming hypotheses"		Scale "data analysis"		Scale "planning experiments"	
		CITC	Cronbach's alpha	CITC	Cronbach's alpha	CITC	Cronbach's alpha
Seed germination	1	0.48		0.44		0.51	
	2	0.44		0.29		0.49	
Chicken eggs	1	0.40	0.77	0.47	0.67	0.29	0.79
	2	0.45		0.20		0.53	
Apple wine	1	0.57		0.27		0.61	
	2	0.66		0.59		0.63	
Baking bread	1	0.45		0.38		0.45	
	2	0.40		0.35		0.47	

Table 5.18: Reliability of the scales "forming hypotheses", "data analysis" and "planning experiments" at the booklet level: Corrected item-total correlation (CITC) and Cronbach's alpha

The reliability coefficient of the scales "forming hypotheses", "data analysis" and "planning experiments" at the booklet level ranged from 0.67 to 0.79. Among them, the reliability coefficient for the scale "data analysis" was lower than the reliability

coefficient for the other two scales. Furthermore, all items had a corrected item-total correlation coefficient higher than 0.2.

	Scale “forming hypotheses”	Scale “data analysis”	Scale “planning experiments”
Pre-test 1	0.78 (7 items)	0.75 (7 items)	0.68 (8 items)
Pre-test 3	0.77 (8 items)	0.67 (8 items)	0.79 (8 items)

Table 5.19: Reliability coefficients (Cronbach’s alpha) for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level in pre-test 1 and pre-test 3

Discussion

The reliability coefficients of the scales “forming hypotheses” and “planning experiments” were quite high. In particular, the scale “planning experiments” possessed a Cronbach’s alpha of 0.79. However, the reliability for the scale “data analysis” was not so high. The Cronbach’s alpha for this scale was lower than 0.7. On the other hand, most items had a corrected item-total correlation coefficient higher than 0.2.

Comparing the reliability of the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level in pre-test 1 and pre-test 3, table 5.19 shows that the reliability coefficients in pre-test 1 and pre-test 3 were similar for the scale “forming hypotheses.” However, for the two other scales the reliability coefficients were different. The Cronbach’s alpha for the scale “data analysis” was higher in pre-test 1 than in pre-test 3, whereas the Cronbach’s alpha for the scale “planning experiments” was lower in pre-test 1 than in pre-test 3.

c) Reliability of the booklet (all items of the four units combined)

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H1	0.55	Apple wine	H1	0.67
	H2	0.41		H2	0.69
	D1	0.47		D1	0.21
	D2	0.14		D2	0.66
	E1	0.53		E1	0.65
	E2	0.56		E2	0.68
Chicken eggs	H1	0.42	Baking bread	H1	0.54
	H2	0.39		H2	0.46
	D1	0.61		D1	0.55
	D2	0.26		D2	0.33
	E1	0.24		E1	0.61
	E2	0.50		E2	0.52

Cronbach’s alpha = 0.89

Table 5.20: Reliability of the booklet (all items of the four units combined): Corrected item-total correlation and Cronbach’s alpha

The reliability for the three scales in the complete test booklet was very high in this pre-test. The Cronbach's alpha was 0.89. Furthermore, only one item out of twenty-four items (item D2 of unit 1) had a corrected item-total correlation coefficient lower than 0.2.

Discussion

A comparison of the reliability coefficients for the complete test booklet (cf. table 5.20) and the individual scales “forming hypotheses”, “data analysis” and “planning experiments” (cf. table 5.19) shows that the Cronbach's alpha increased when the three scales were combined. This is an interesting finding, insofar as longer scales are typically more reliable, but in this test, the different scales were hypothesized to be motivated by different kinds of knowledge, in particular methodological knowledge about the method of experimentation and knowledge about the biological contents of the experiments. An increased test reliability for the complete test can be interpreted as an indication that the three dimensions in experimentation may be highly correlated. The correlation coefficients between the three dimensions and between the students' biological content knowledge and their competencies in experimentation were analysed in the following section of this chapter.

d) Comparison of the Cronbach's alphas for the test when two different scoring models are used (0/1 and 0/1/2)

		Scoring Model 0/1	Scoring Model 0/1/2
Unit	Seed germination	0.69	0.66
	Chicken eggs	0.66	0.64
	Apple wine	0.84	0.80
	Baking bread	0.75	0.75
Scale	Forming hypothesis	0.77	0.78
	Data analysis	0.67	0.69
	Planning experiment	0.79	0.78
All three scales “forming hypothesis”, “data analysis” and “planning experiment” at combined booklet level		0.89	0.88

Table 5.21: Cronbach's alphas for the units, for the scales “forming hypothesis”, “data analysis” and “planning experiment” at the booklet level and for all three scales combined – for two scoring models

A comparison of the reliability of the test scored with two different models (0/1 and 0/1/2) shows that the reliability was similar at the level of the individual unit, at the

level of the scale and at the level of the complete test booklet. Generally, the Cronbach's alpha was slightly higher for the 0/1 model than for the 0/1/2 model.

Discussion

Although the item difficulty for the competency test was low, the reliability coefficients for the individual scales “forming hypotheses”, “planning experiments” and “analysing data” were quite high. In particular, the reliability at the level of the scales “forming hypotheses”, “data analysis” and “planning experiments” reaches or surpasses the acceptable reliability coefficient of 0.7. In addition, most items had a corrected item-total correlation higher than 0.2. Also, the high Cronbach's alpha for the test at the booklet level indicate that the competency test can be used to calculate the correlations between the students' pre-knowledge and their competencies in experimentation.

The comparison of the two scoring models reveals that there is no added value in terms of test reliability if an intermediate level of competency is assumed and tested with multiple choice distractors derived from the empirical investigation of student competencies in experimentation. This does not question, of course, the existence of such an intermediary level that can be revealed with other methods of investigation, for example observations of student behaviour during experimentation (Schauble et al. 1991).

2.3. Correlations

2.3.1. Method

In the knowledge test, we summed the scores for the strong items at the booklet level in order to calculate the sum score of the students' content knowledge .

In the competency test, we calculated the total scores for each dimension of experimentation (“search in the hypothesis space”, “search in the experiment space” and “data analysis”) using the 0/1 scoring model.

This was done in order to the correlation coefficients (Spearman) between the sum score of the knowledge test and the sum scores for each dimension of experimentation.

2.3.2. Findings

2.3.2.1. Correlations between the students' pre-knowledge and the three dimensions in experimentation

	Pre-knowledge * Search in the hypotheses space	Pre-knowledge * Data analysis	Pre-knowledge * Search in the experiment space
Correlation coefficient	0.231*	0.144	0.172

Table 5.22: Correlations (Spearman) between the students' pre-knowledge and the three dimensions of experimentation

The correlation coefficients between the students' pre-knowledge and the three dimensions in experimentation ranged from 0.144 to 0.231. Among these coefficients, the correlation between the content knowledge and the dimension "search in the hypothesis space" was slightly higher than the correlation coefficients for the two other combinations.

Discussion

Similar to the findings in pre-test 1, the correlation coefficients between the students' pre-knowledge and the three dimensions in experimentation were low. Most correlation coefficients were lower than 0.2. In contrast, we expected that there are high correlations between the students' pre-knowledge and specific dimensions, especially, between the students' pre-knowledge and the dimensions "search in the hypothesis space" and "data analysis", because, these dimensions were hypothesized to be influenced by the students' pre-knowledge about the biological contents of the experiment whereas the dimension "search in the experiment space" was hypothesized to be affected by the students' methodological knowledge.

Thus, our hypotheses about high correlations between the students' biological pre-knowledge and the dimensions "search in the hypothesis space" and "data analysis" were not confirmed in this test.

There are two possible reasons for this. On the one hand, there may be a general factor that underlies all three dimensions. This factor was not assessed in this study, for example the students' intelligence, i.e. their ability to think logically and – on the basis of this – make inferences. This explanation is corroborated by the fact that all three item types used in this study – items that require identifying the hypothesis that can be tested in a given experiment, items that require planning an experiment for a given hypothesis and items that require the interpretation of data of a given experiment – require logical

thinking. As a second explanation, it is possible that solving the items does require different kinds of knowledge – i.e., methodological knowledge about the method of experimentation and biological knowledge about the biological contents of the experiments – but that our hypotheses did not capture the specific degree to which both components are necessary for solving all three item types. For example, for solving the item type “analysing data”, it is necessary to have biological content knowledge, as studies have shown that students’ alternative conceptions severely hamper the students’ ability to interpret data objectively. On the other hand, the students’ ability to analyse data may also be supported by methodological knowledge, e.g. knowledge about the differences between test variables and variables that need to be controlled and knowledge about the methodological convention to vary only the test variable. If this kind of knowledge is not available, it is difficult for the students to identify unconfounded experiments and make correct inferences about the cause of an experimental result. The same is true for the ability to solve items in this study that test the ability to form hypotheses.

According to David Klahr (2000), people form hypotheses either from their pre-knowledge or from experimental findings. The way in which this competence is operationalised in this study was to present students with the findings from an experiment and asks them to identify the hypothesis that can be tested with the experiment. Essentially, this item type may also require two types of knowledge. Biological content knowledge may very well influence the ability to identify the correct hypothesis in ways that a similar to the correct interpretation of experimental findings. Also, this item type requires an understanding of the basic principles of experimental design in order to identify the correct hypothesis.

As a general concluding remark, it may be argued that it may be very difficult to design items that test competencies motivated by exclusively methodological knowledge – if these competencies are tested in a biological domain. A different approach – not chosen in this study – consists of designing items that do not require any specialized knowledge about the contents of the experiment, for example items about events from daily life (e.g. Tschirgi 1980). These types of items, however, were not used in this study because the aim of this study was to investigate the students’ biological knowledge and their competencies in biological experiments.

2.3.2.2. Correlations between the students' knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation

	Search in the hypothesis space	Data analysis	Search in the experiment space
Knowledge about which factors are relevant and irrelevant for the biological phenomenon examined	0.315**	0.295*	0.473**

Table 5.23: Correlations (Spearman) between the students' knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation

For calculating these correlations, the means for the students' knowledge about relevant and irrelevant factors (cf. table 5.2 – 5.5) for the biological phenomenon under consideration (e.g., seed germination) were correlated with the test scores for the three dimensions of the competence test. The irrelevant factors assessed in this study, for example, concerning the germination of bean seeds were soil and light. Relevant factors for the same experiment, for example, were air, warmth and water. The correlations between the students' knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation ranged from 0.295 to 0.473. Among them, the correlations between the dimension “planning experiments” and knowledge about which factors are relevant and irrelevant for the biological phenomenon examined are the highest.

Discussion

The items that test knowledge about factors which are relevant and irrelevant for the biological phenomenon represent a selection of the knowledge test (cf. tables 5.2 – 5.4). They can be considered a more specific indicator of the students' biological pre-knowledge than the sum score of the knowledge test because the latter also contains questions that are more peripheral to the experiment presented in the competence test (cf. Appendix, p. 263, 264). This is the reason for the differences between the correlations in table 5.22 and 5.23. The findings indicate that if students possess knowledge about which factors are relevant and which are irrelevant, they are more likely to achieve higher scores in the competence test than students who do not possess this kind of knowledge. This finding further substantiates the points made about the role of the student's pre-knowledge in the previous section (cf. 2.3.2.1).

2.3.2.3. Correlations between the three dimensions in experimentation

	Search in the hypothesis space * Data analysis	Data analysis * Search in the experiment space	Search in the hypothesis space * Search in the experiment space
Correlation coefficient	0.666**	0.600**	0.646**

Table 5.24: Correlations (Spearman) between the three dimensions in experimentation

The correlation coefficients between the three dimensions of experimentation ranged from 0.600 to 0.666, all of which were significant. Among them, the correlations between the dimensions “search in the hypothesis space” and “data analysis” were slightly higher than in the two other combinations.

Discussion

Two findings are crucial and need to be discussed.

On the one hand, the correlation coefficients for the relationships between the three dimensions in experimentation were much higher than the correlation coefficients for the relationships between the students’ pre-knowledge and the three dimensions in experimentation. That is, the competence, for example, analyzing experimental data, seems a stronger predictor for other competences in experimentation (e.g., forming hypotheses, planning experiments) than knowledge about the biological content of the experiment, for example knowledge about which factors are relevant for explaining a specific phenomenon under consideration. Given the rather small sample size of 77 students in this study, this finding needs to be considered with some caution, but it seems to indicate that there is a difference between the constructs “knowledge” and “competences”. This difference, for which we find empirical proof here, was the basis for constructing independent tests to measure knowledge and competences.

On the other hand, the high correlation coefficients for the relationships between the three dimensions in experimentation need to be discussed. Interestingly, this finding is consistent with the findings in pre-test 1, where the same correlations were also analysed (cf. table 3.46). In pre-test 1, the correlation coefficient for the dimension “search in the hypothesis space” and “data analysis” in booklet 111 was 0.716. For the dimension “data analysis” and “search in the experiment space” the correlation coefficient was 0.623 and for the dimension “search in the hypothesis space” and “search in the experiment space” it was 0.589. Again, the small sample sizes in both studies require to exercising some caution in interpreting these results. But there seem

to be less differences between the three dimensions in experimentation than hypothesized. The reasons for this have been discussed before (see Chapter 3, section 2.4.2.2).

3. Conclusion

This pre-test was successful insofar as the knowledge test and the competency test proved reliable.

Although the reliability at unit level was not high in the knowledge test, the reliability coefficient at the booklet level was 0.74. Thus, the knowledge test at the booklet level was reliable and could be used to assess the relationship between the students' pre-knowledge and their competences in experimentation.

In the competency test, the reliability coefficients at the level of the unit, the scales and the test booklet were high. For example, the Cronbach's alpha for the test booklet was 0.89, so that it is fair to say that it seems easier to construct a reliable competence test than a reliable knowledge test. The reasons for this are uncertain, but the biological contents of the units are probably not very familiar to the students which results in lower reliabilities in the knowledge test.

The opinion items successfully assessed the students' beliefs about factors that were relevant and irrelevant for the biological phenomena that were presented in the experiments. The students in grade six were found to possess correct content knowledge in some areas and they could answer some questions correctly. For example, concerning seed germination, 87% of the students believe that water is very important. Over 70% of the students indicate that warmth is very important or important. 87% of the students said that seed can germinate in other materials, not only in the soil. However, for some variables, students had alternative conceptions. For example, 86% of the students indicate that light is very important or important for seed germination.

The correlation coefficients for the relationship between the students' pre-knowledge and the three dimensions of experimentation were low. Higher correlation coefficients were found for the relationships between the three dimensions in experimentation. This is an interesting finding because – according to this study with a limited number of students – biological content knowledge is a less strong predictor for the competencies of, for example, planning experiments and analyzing data than, for example, the competence of forming hypotheses. It remains to be seen in the main study if this finding is robust enough to reoccur.

Part III

Main Study and Conclusion

Chapter 6: Main study

1. Method

Sample

In this study, 1006 students (511 girls and 495 boys) in grades five and six from 24 secondary schools in Germany participated. 753 students were in grade six and 253 students were in grade five. Their age ranged from 9 years and 11 months to 14 years and 2 months. The mean age for the sample was 11 years and 8 months.

Design

Each student worked on two types of test, the competency test and the knowledge test. The same test as described in the previous Chapter “Pre-test 3” was used. This test had four biological contents: Unit 1: Seed germination, Unit 2: Chicken eggs, Unit 3: Apple wine and Unit 4: Baking bread.

In the knowledge test, each unit had seven to eight questions. In the competency test, there were six questions for each unit.

Answering format

In the knowledge test, simple multiple-choice questions were used. Each question had four choices, only one of which was correct. One opinion question was used to assess the students’ beliefs about the importance of variables in experiments. For this item, a Likert scale with four categories was used: very important, important, not very important and not important.

The competency test consisted of simple multiple choice questions with one correct option and three distractors. Two of the distractors were partly correct.

Coding the answers

As in pre-test 3, a scoring model that distinguishes between “full credit” and “no credit” was used for the knowledge test. For a correct answer, the code 1 was used, for an incorrect answer the code 0.

For the competency test, two scoring models were used: a scoring model that distinguishes between “full credit” (code 2) and “no credit” (code 0) and a partial credit scoring model that distinguishes an additional intermediary level.

2. Results

2.1. Item difficulty

2.1.1. Method

As in pre-test 3, item difficulties were calculated for individual items as well as mean item difficulties for the units and for the booklets for both the knowledge test and the competency test.

In addition, mean item difficulties were calculated for each dimension of experimentation.

2.1.2. Findings

2.1.2.1. Item difficulty in the knowledge test

a) Item difficulty for the items in the knowledge test

Item	Unit 1: Seed germination	Unit 2: Chicken eggs	Unit 3: Apple wine	Unit 4: Baking bread
2	60.5	81.0	47.7	70.2
3	78.4	46.4	48.5	67.1
4	37.3	84.3	52.2	55.8
5	33.7	64.7	10.3	84.6
6	39.7	42.1	63.3	41.2
7	11.9	59.9	44.5	55.0
8		19.4	56.1	47.9

Table 6.1: Item difficulty for the items in the knowledge test (n = 1006)

The item difficulty for the items in the knowledge test ranged from 10.3% to 84.6%. Most of the items had an item difficulty larger than 20% and smaller than 80%. Only three items (item 7 of unit 1, item 8 of unit 2 and item 5 of unit 3) were correctly answered with a probability of lower than 20% and three items (item 2, 4 of unit 2, item 5 of unit 4) had a low item difficulty with more than 80% of the students who solved these items correctly.

In unit 1, the item difficulty ranged from 11.9% to 78.4%. Five items out of six were correctly answered by more than 30% of the students. In unit 2, the item difficulty ranged from 19.4% to 84.3%. Two items (items 2 and 4) had a low item difficulty and were solved by over 80% of the students.

In unit 3, the item difficulty ranged from 10.3% to 63.3%. Item 5 had a very high item difficulty and was answered by only 10% of the students.

In unit 4, all items were correctly answered with a probability of over 40% and only item 5 had a very low item difficulty with a probability of 84% of the students who answered this item correctly.

b) Mean item difficulty for the units in the knowledge test

	Unit 1: Seed germination	Unit 2: Chicken eggs	Unit 3: Apple wine	Unit 4: Baking bread
Both grades five and six combined (n = 1006)	43.6	56.8	46.1	60.3
Grade five (n = 253)	40.3	53.8	40.1	55.8
Grade six (n = 753)	45.5	57.9	48.2	61.8

Table 6.2: Mean item difficulties for the units in the knowledge test

The mean item difficulties for all four units ranged from 40% to 62%. The mean item difficulty for unit 4 was the lowest and unit 1 had the highest item difficulty.

Comparing the mean item difficulties for units in fifth grade and sixth grade, table 6.2 shows that the item difficulty for the units in fifth grade is higher than in sixth grade in all four units. The item difficulty for the units in fifth grade ranged from 40% to 55%, while it ranged from 45% to 61% in sixth grade.

The mean item difficulty for the booklet was 51.7%. For grade five, it was 47.5% and for grade six, it was 53.4%.

Discussion

The item difficulties for most items in the knowledge test ranged from 20% to 80%. Only a few items had an unacceptably high or low item difficulty.

The reasons for this can be discussed. Item 7 of unit 1 had an item difficulty of 11.9%. In this item, the students were asked to interpret data from an experiment investigating the factors that influence seed germination. Only 11.9% of the students answered the item correctly and took into consideration that bean seeds can germinate in the dark, while 48.9% of the students thought that seeds need to absorb nutrients to germinate, although this is not consistent with the experimental results presented to the students. Another item in the test, the opinion item, can be used to interpret this finding. Many students did not solve this item because most of them (87%) believe that seeds need light to germinate. Thus the students' conceptions (alternative conceptions) interfered with their interpretation of so-called anomalous data, i.e. data that do not meet the students' expectations. This is a well described phenomenon, as students seem to possess many different ways to deal with non-confirming evidence (Chinn & Brewer 1998).

The item difficulty for item 5 of unit 3 was 10.3%. This item presents an experiment investigating the factors that affect the production of alcohol in apple wine. The correct interpretation of the experiment was: “Wine will have a lot of alcohol when one puts a lot of sugar in apple juice.” However, many students (64%) believed that “apple wine will have much alcohol when it is stored a very long time” which is another student conception.

Thus, these two items had a high item difficulty because the beliefs of the students about the variables were in conflict with the scientific conceptions.

However, the mean item difficulty for all four units ranged from 40% to 60%. Among them, Unit 1: Seed germination and Unit 3: Apple wine had similar item difficulties and they were more difficult than Unit 2: Chicken eggs and Unit 4: Baking bread.

A comparison of the item difficulty between grade five and grade six, table 6.2 shows that the item difficulty for grade six was lower than that for grade five. However, there was also some variation (4% to 8%) among the units.

Thus, the knowledge test had an acceptable difficulty for all students in both grades.

2.1.2.2. Assessment students’ beliefs about the importance of variables that are relevant / irrelevant for the phenomenon investigated in the experiments

Unit 1: Seed germination

	Very important	Important	Not very important	Not important	Mean for each variable	STD
Light	69.8	17.5	7.3	5.4	3.52	0.85
Warmth	35.1	40.6	18.2	6.1	3.05	0.88
Air	35.4	31.5	22.3	10.8	2.91	1.00
Soil	61.3	18.3	13.9	6.4	3.35	0.94
Water	83.9	12.2	2.1	1.7	3.78	0.56

Table 6.3: Frequency of students (%) and mean for each variable (4: very important, 3: important, 2: not very important and 1: unimportant) in unit 1.

The question for unit 1 was “How important are the above variables for seed germination?”

87.3% of the students indicated that light is important or very important for seed germination. However, bean seeds can germinate in the dark (dark germination plants). Thus, the experimental results presented to the students did not confirm their beliefs. Similarly, 79.6% of the students thought that soil is important for seed germination. However, seeds can also germinate in other materials, for example, in cotton wool.

On the other hand, 75.7% of the students believed that warmth was important. 66.9% of the students thought that air was important and, significantly, 96.1% of the students indicated that water was important for seed germination. For these variables, the students' beliefs were confirmed by the experimental data presented in the items.

Unit 2: Chicken eggs

	Very important	Important	Not very important	Not important	Mean for each variable	STD
Light	24.6	24.9	28.5	22.0	2.52	1.08
Warmth	93.1	5.4	0.9	0.6	3.91	0.37
Humid air	7.1	24.9	31.6	36.4	2.03	0.92
Size of eggs	6.7	12.0	26.4	54.9	1.70	0.92
Color of eggs	3.4	7.6	14.6	74.4	1.40	0.77

Table 6.4: Frequency of students (%) and mean score for each variable (4: very important, 3: important, 2: not very important and 1: unimportant) in unit 2.

In unit 2, the question was “How important are the above variables, so that chicken eggs hatch as quickly as possible?”

93.1% of the students believed that the temperature is a very important factor. This belief was confirmed in the following items, in which the students are presented an experiment which shows that hatching chicken eggs is affected by the temperature. In contrast, not all students recognized that light does not influence the hatching of chickens, because 49.5% of the students thought that light was important, whereas, in fact, light does not affect hatching eggs. In contrast, only 32% of the students believed that humid air was important for hatching eggs. However, in fact, the humidity also affects hatching eggs. On the other hand, for two other factors, the size of eggs and the colour of eggs, the students' beliefs were confirmed in the following experiments. Most of them (81.3% and 89%) believed these two factors were not very important for hatching the eggs.

Unit 3: Apple wine

	Very important	Important	Not very important	Not important	Mean for each variable	STD
Light	17.9	14.7	27.4	40.0	2.11	1.12
Warmth	20.9	30.7	27.5	20.9	2.52	1.04
A pot without air coming in	71.1	18.3	5.4	5.2	3.55	0.82
Amount of sugar	15.4	33.1	31.7	19.8	2.44	0.98
Yeast	26.0	20.8	14.8	38.4	2.34	1.23

Table 6.5: Frequency of students (%) and mean score for each variable (4: very important, 3: important, 2: not very important and 1: unimportant) in unit 3.

In unit 3, the question was “How important are the above variables, so that apple wine is made successfully?”

Most of the students (67.4%) believed that light was not important for making wine, but 89.4% thought a wine-making vessel, constructed to prevent air from coming in, was important. Thus, the beliefs of the students concerning these two variables were confirmed. However, only 51.6% of the students indicated that the temperature was important. For the amount of sugar, only 48.5% of the students and for yeast 46.8% of the students thought that these factors were important. In fact, the amount of sugar has an influence on the amount of alcohol in the wine and yeast is a prerequisite for making wine. So, for these variables, the students’ beliefs were not consistent with the biological facts.

Unit 4: Baking bread

	Very important	Important	Not very important	Not important	Mean for each variable	STD
Yeast	88.7	9.5	1.5	0.3	3.87	0.41
Flour	59.2	33.1	5.9	1.7	3.50	0.69
Sugar	11.9	22.5	38.3	27.3	2.19	0.97
Butter	24.9	38.6	24.3	12.1	2.76	0.96
Water temperature	18.8	21.9	25.9	33.4	2.26	1.11

Table 6.6: Frequency of students (%) and mean score for each variable (4: very important, 3: important, 2: not very important and 1: unimportant) in unit 4.

In unit 4, the question was “How important are the above variables, so that yeast dough rises?”

88.7% of the students believed that yeast was very important, but only 40.7% of the students thought that the temperature of the water has an influence on making bread. In fact, both yeast and the water temperature are important so that yeast dough rises. In

contrast, 92.3% of the students indicated that flour was important and 63.5% of the students believed that butter was important. So the students' beliefs about the water temperature, flour and butter were not congruent with the biological facts. On the other hand, only 34.4% of the students thought that sugar was important and this variable was confirmed.

Discussion

We assessed students' beliefs about the importance of 20 variables in all four units. The students' beliefs about seven variables were not congruent with the biological facts and it can be expected that student conceptions affect the interpretation of data in biological experiments. For example, in unit 1 the students considered light and soil important for seed germination. In fact, seeds can germinate in the dark and in materials other than soil, as long as the materials are permeable to water and oxygen. In unit 2, students did not consider humid air important for the hatching of eggs. In unit 3, the students believed that yeast and the amount of sugar were not important for making wine. In unit 4, the students indicated that flour and butter affect the rising of yeast dough.

2.1.2.3. Item difficulty in the competency test

a) Item difficulty for the items in the competency test

Unit	Search in the hypothesis space		Data analysis		Search in the experiment space	
	H1	H2	D1	D2	E1	E2
Seed germination	54.3	56.8	56.0	74.4	69.6	55.8
Chicken eggs	61.4	65.2	46.2	46.1	62.1	46.6
Apple wine	60.0	67.0	77.3	72.7	54.8	50.2
Baking bread	74.4	77.6	73.2	69.1	66.2	57.4

Table 6.7: Item difficulty for the items in the competency test (n = 1006)

The item difficulty for the items in the competency test ranged from 46% to 77%.

In Unit 1: Seed germination, the item difficulty ranged from 54% to 74%. Item 2 in the dimension "data analysis" had the lowest item difficulty. In this unit, items in the dimension "search in the hypothesis space" were more difficult than items in the other two dimensions.

In Unit 2: Chicken eggs, the item difficulty ranged from 46% to 65%. Among them, two items in the dimension "data analysis" and one item in the dimension "search in the experiment space" were more difficult than the three remaining items. In this unit, the items in the dimension "search in the hypothesis space" had the lowest item difficulty.

In Unit 3: Apple wine, the item difficulty ranged from 50% to 77%. Contrary to unit 2, in this unit, the items in “data analysis” had the lowest item difficulties and two items in “search in the experiment space” were more difficult than the remaining items.

In Unit 4: Baking bread, the item difficulty ranged from 57% to 77%. Like in unit 2, the items in the dimension “search in the hypothesis space” had a lower item difficulty than the items in the two other dimensions. However, most items in this unit were correctly answered with a probability of over 60%.

b) Mean item difficulty for the units in the competency test

	Unit 1: Seed germination	Unit 2: Chicken eggs	Unit 3: Apple wine	Unit 4: Baking bread
Both grades five and six combined	61.2	54.6	63.7	69.7
Grade five	52.0	46.5	53.8	61.6
Grade six	64.2	57.3	67.0	72.3

Table 6.8: Mean item difficulty for the units in the competency test

The mean item difficulty for the units in the competency test ranged from 52% to 72%. Unit 2 had the highest mean item difficulty of all four units, unit 4 the lowest.

Comparing grade five and grade six, table 6.8 shows that in grade five the mean item difficulty of the units ranged from 46% to 61% and that it ranged from 57% to 72% in grade six. In all four units, the mean item difficulty in grade six was 11% to 14% lower than the mean item difficulty for grade five.

c) Mean item difficulty for the three dimensions in experimentation and for the booklet in the competency test

	Search in the hypothesis space	Data analysis	Search in the experiment space	Mean item difficulty for booklet
Both grades five and six combined	64.6	64.4	57.8	62.3
Grade five	54.2	57.5	48.7	53.5
Grade six	68.1	66.7	60.9	65.2

Table 6.9: Mean item difficulty for the three dimensions and for the booklet in the competency test

The mean item difficulty for the three dimensions of experimentation ranged from 49% to 68%. Among them, the mean item difficulty for both dimensions “search in the

hypothesis space” and “data analysis” was 64% and it was 57% for the dimension “search in the experiment space”.

Besides, the mean item difficulty for the three dimensions in grade six was from 11% to 14% lower than that in grade five.

Discussion

The item difficulty for items in the competency test was ideal from a test theory perspective. It ranged from 46% to 77%. Unit 4 contained the items with the lowest item difficulty and unit 2 contained the items with the highest item difficulty.

For each unit, there was some variation concerning the item difficulty of items in one dimension. This was not to be expected because the items of the same unit that belong to one dimension were designed to test the same competence. For example, in unit 1, in the dimension “data analysis”, the item difficulty of item 1 was 56%, but that of item 2 was 74%. Also, in the dimension “search in the experiment space”, item 1 had an item difficulty of 69%, whereas item 2 had an item difficulty of 55%. The reasons for this may be that competencies in experimentation were influenced by the beliefs of the students about the variables as well as by the students’ pre-knowledge. Also, the design of the item may have made a difference, because in “search in the experiment space” the two items in one unit were differently designed. Slight variations between the items of the same dimension across different units can be explained because of the effect of the domain-specific knowledge on the competence. This interpretation is corroborated by the fact that, interestingly, item difficulties are not similar between the different dimensions of the same unit. For example, in unit 1 the dimension “search in the hypothesis space” had a higher item difficulty than the two other dimensions. However, in unit 2, the item difficulty was the highest in the dimension “data analysis” and in units 3 and 4 the dimension “search in the experiment space” was more difficult than the other two dimensions. These differences may go back to interpretations are warranted because the items within one dimension are characterized by a great degree of homogeneity across the different units.

The mean item difficulty for the units ranged from 54% to 69%. It was somewhat lower in unit 4 and higher in unit 2 than in the two other units. The mean item difficulty for the units in grade five was much higher than in grade six. Depending on the unit, the mean item difficulty for units in grade five and 6 ranged from 11% to 14%. However, the mean item difficulty for grade five was also lower than 50% in three units out of four. This means that the competency test was suitable for both of grades 5 and 6.

The mean item difficulty for the three dimensions in experimentation was similar in the dimensions “search in the hypothesis space” and “data analysis”. The mean item difficulty for in the dimension “search in the experiment space” was higher than in the other two dimensions. This is an interesting new finding that has not been described in the literature before. The reasons for this finding are not apparent, but may be related to the item format used in this study. For further clarification, the different items need to be solved by students using the “thinking out loud” method so that insights into the specific difficulties posed by the different item types can be gained.

2.2. Factor analysis

Aims of the factor analysis

In the competency test, each unit includes 6 items, two each for the dimensions “search in the hypothesis space”, “data analysis” and “search in the experiment space”. These 24 items were entered into a confirmatory factor analysis in order to investigate whether the 24 items load on three factors corresponding to the three dimensions in experimentation.

2.2.1. Method

For the factor analysis, the number of factors was set to three. The 24 items of the competency test were analysed as 24 variables. The items in “search in the hypothesis space” were signed “H” (H1: item 1, H2: item 2), those in “data analysis” were marked “D” and in “search in the experiment space” were “E”.

2.2.2. Findings

When no specific number of factors was pre-determined, twenty-four items in the competency test were extracted into five factors explaining 47.588% variance.

However, as explained above, we wanted to see whether twenty-four items load on three factors corresponding to the three dimensions in experimentation. So, a three factor solution was pre-determined.

Twenty-four items in the competency test were extracted into three factors explaining 37.06% variance.

The factor loadings are depicted in table 6.10.

Rotated Components matrix (a)

	Components		
	1	2	3
H1_unit 3	0.643	0.137	0.207
H2_unit 4	0.632	0.090	0.152
D4_unit 3	0.620	0.185	0.175
H1_unit 4	0.608	0.079	0.153
D3_unit 3	0.597	0.059	0.153
H2_unit 3	0.554	0.115	0.260
D4_unit 4	0.538	0.134	0.132
D3_unit 4	0.495	0.144	0.131
H2_unit 2	0.474	0.211	0.304
E5_unit 3	0.431*	0.389*	0.118
D4_unit 2	0.409	0.188	0.017
D3_unit 2	0.357	0.238	0.118
H1_unit 2	0.337	0.184	0.126
E6_unit 1	-0.014	0.716	0.141
E5_unit 2	0.228	0.572	0.011
E6_unit 2	0.324	0.544	-0.082
E6_unit 4	0.170	0.486	0.226
E5_unit 4	0.166	0.475	0.191
E5_unit 1	-0.063	0.461	0.451
E6_unit 3	0.355*	0.428*	0.047
H2_unit 1	0.219	0.085	0.660
D3_unit 1	0.190	0.174	0.660
H1_unit 1	0.345	0.145	0.586
D4_unit 1	0.243	0.024	0.538

* a tendency for a double loading of this item

Table 6.10: Results of the factor analysis of the competency test (VariMax rotation, 3 factors solution pre-determined) (the main loading is underlined in grey)

Factor 1 (variance 17.458%) consisted of all the items in the dimension “search in the hypothesis space” of unit 2, unit 3 and unit 4 and all the items in the dimension “data analysis” of unit 2, unit 3 and unit 4 as well as item 5, unit 3, of the dimension “search in the experiment space”.

Factor 2 (variance 10.412%) contained all items of all four units in the dimension “search in the experiment space” except for item 5 of unit 3.

Factor 3 (variance 9.190%) comprised four items of unit 1 in the dimensions “search in the hypothesis space” and “data analysis”.

Thus, the findings support the three factor solution only to some degree because factor one combined two dimensions, while factor 2 contained items that loaded on a particular unit.

Therefore, the factor analysis was repeated and the number of factors set to two in order to constrain the 24 items to two factors. The two factors explain 32.11% of the variance. Table 6.11 shows the results of this factor analysis.

Rotated Components matrix (a)

	Components	
	1	2
H1_unit 3	0.666	0.160
H2_unit 4	0.639	0.101
D4_unit 3	0.630	0.198
H1_unit 4	0.617	0.091
D2_unit 3	0.608	0.072
H2_unit 3	0.603	0.158
D4_unit 4	0.541	0.141
H2_unit 2	0.538	0.265
H1_unit 1	0.522	0.284
D4_unit 4	0.500	0.153
H2_unit 1	0.434	0.252
E5_unit 3	0.421*	0.387*
D4_unit 1	0.417	0.159
D2_unit 1	0.402	0.339
D4_unit 2	0.376	0.167
D2_unit 2	0.361	0.245
H1_unit 2	0.349	0.197
E5_unit 1	-0.004	0.727
E6_unit 1	0.075	0.569
E5_unit 2	0.184	0.543
E6_unit 4	0.211	0.522
E5_unit 4	0.195	0.502
E6_unit 2	0.241	0.486
E6_unit 3	0.322*	0.409*

* a tendency for a double loading of this item

Table 6. 11: Results of the factor analysis of the competency test (VariMax rotation, 2 factors solution pre-determined) (the main loading is underlaid in grey)

Factor 1 (variance 20.182%) contained all items (units 1-4) in the dimension “search in the hypothesis space” and all items (units 1-4) in the dimension “data analysis”, as well as item 5 of unit 3 which belongs to the dimension “search in the experiment space”.

Factor 2 (variance 11.930%) consisted of all items in “search in the experiment space” for four units, except for item 5 of unit 3, although the latter has a double loading on both factors.

In the competency test, the items are organized in four thematic units. In order to rule out the possibility that thematic context of the items influences the results of the factor analysis, we randomly sorted the eight items of each dimension to two variables for each dimension.

For example, four items H1 from the dimension “search in the hypothesis space” were combined and a mean score was calculated.

Unit 1	H1	H2	D3	D4	E5	E6
Unit 2	H1	H2	D3	D4	E5	E6
Unit 3	H1	H2	D3	D4	E5	E6
Unit 4	H1	H2	D3	D4	E5	E6
	Mean H1 (Hypothesis 1)	Mean H2 (Hypothesis 2)	Mean D3 (Data analysis 1)	Mean D4 (Data analysis 2)	Mean E5 Experiment 1	Mean E6 Experiment 2

Table 6.12: Means for four randomly assigned items that belong to the dimensions “search in the hypothesis space”, “search in the experiment space” and “data analysis”

The first factor analysis over these six variables was done with the number of factors set to three. The factor loadings showed that factor 1 contained four variables in “search in the hypothesis space” and “data analysis”, factor 2 was one variable in “search in the experiment space” and factor 3 was the other variable in “search in the experiment space”. These results did not confirm the findings of the second factor analysis reported above.

Accordingly, the factor analysis for the six variables listed in table 6.13 was repeated with the number of factors set at two. These two factors explained 70.30% of the variance.

The factor loading are depicted in table 6.13:

Rotated Components matrix (a)

	Components	
	1	2
Hypothesis 1	0.821	0.216
Hypothesis 2	0.816	0.243
Data analysis 3	0.763	0.262
Data analysis 4	0.739	0.277
Experiment 5	0.196	0.875
Experiment 6	0.347	0.759

Table 6.13: Results of the factor analysis of the competency test (VariMax rotation, 2 factors solution pre-determined) (the main loading is underlined in grey)

Factor 1 (variance 43.759%) consisted of items 1 and 2 of the dimension “search in the hypothesis space” and items 3 and of the dimension “data analysis.”

Factor 2 (variance 26.546%) contained items 5 and 6 of “search in the experiment space”.

2.2.3. Discussion of the findings of the confirmatory factor analysis

The factor analyses confirmed to some extent the assumptions that were made for the construction of the competency test. Regardless of the fact whether the number of factors was set at three or at two, the items that belong to the dimensions “search in the hypothesis space” and “data analysis” appeared in conjunction, whereas the items that form the dimension “search in the experiment space” formed a single factor. This suggests that the two dimensions “search in the hypothesis space” and “data analysis” are formed by a common factor, possibly the students’ pre-knowledge. This was to be expected as we hypothesized these two dimensions to be motivated by the students’ pre-knowledge about the science content, whereas the dimension “search in the experiment space” was hypothesized as formed by the student’s methodological knowledge about the aims and purposes of experimentation. The effects of context-based assessment became visible when a third factor was formed with items that belong to a specific unit in the first factor analysis.

2.3. Reliability

Reliability analyses were performed to ascertain that correlations between the students’ pre-knowledge and the three dimensions in experimentation can be calculated.

2.3.1. Method

In the knowledge test, we calculated the reliability at the unit level and at the booklet level based on the corrected item-total correlation for each item and the reliability coefficient Cronbach’s alpha.

In the competency test, we also calculated the reliability at the unit level. Besides, we calculated the reliability for the three scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level and as well as the reliability for all three scales at the booklet level combined.

2.3.2. Findings

2.3.2.1. Reliability of the knowledge test

a) Reliability at the unit level

Unit 1: Seed germination

Item	Corrected item-total correlation	Cronbach's alpha
2	0.13	0.32
3	0.11	
4	0.13	
5	0.10	
6	0.25	
7	0.15	

Table 6.14: Reliability of Unit 1: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 1 was 0.32. Most of the items had a corrected item-total correlation lower than 0.2. However, after deleting the non-discriminating items, the Cronbach's alpha was even lower.

Unit 2: Chicken eggs

Item	Corrected item-total correlation	Cronbach's alpha
2	-0.04	0.20
3	0.10	
4	0.17	
5	0.09	
6	0.06	
7	0.11	
8	0.07	

Table 6.15: Reliability of Unit 2: Cronbach's alpha and corrected item-total correlation

As in unit 1, the reliability coefficient for unit 2 was very low. The Cronbach's alpha was 0.20. Moreover, all items had a corrected item-total correlation lower than 0.2. After deleting the non-discriminating items, the highest Cronbach's alpha was 0.27.

Unit 3: Apple wine

Item	Corrected item-total correlation	Cronbach's alpha
2	0.23	0.42
3	0.25	
4	0.29	
5	0.06	
6	0.28	
7	0.09	
8	0.13	

Table 6.15: Reliability of Unit 3: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 3 was 0.42. Four items out of six had a corrected item-total correlation lower than 0.2. After deleting the non-discriminating items, the Cronbach's alpha was 0.48.

Unit 4: Baking bread

Item	Corrected item-total correlation	Cronbach's alpha
2	0.18	0.35
3	0.17	
4	0.19	
5	0.17	
6	-0.01	
7	0.15	
8	0.21	

Table 6.17: Reliability of Unit 4: Cronbach's alpha and corrected item-total correlation

The reliability coefficient for unit 4 was 0.35. Most of the items had a corrected item-total correlation lower than 0.2. After deleting the non-discriminating items, the highest Cronbach's alpha was 0.41.

Unit	Cronbach's alphas in both grades five and six combined (n = 1006)	Cronbach's alphas in grade five (n = 253)	Cronbach's alphas in grade six (n = 753)
Seed germination	0.32	0.33	0.38
Chicken eggs	0.27	0.36	0.30
Apple wine	0.48	0.46	0.49
Baking bread	0.41	0.36	0.41

Table 6.18: The reliability coefficient (Cronbach's alpha) at the unit level in the knowledge test

At the level of the unit, the reliability coefficient was generally very low. The Cronbach's alphas ranged from 0.27 to 0.49.

The Cronbach's alpha ranged from 0.33 to 0.46 in grade five. In grade six, it ranged from 0.30 to 0.49.

b) Reliability at the booklet level

	Both grades 5 and 6 combined (n = 1006)	Grade five (n = 253)	Grade six (n = 753)
Cronbach's alpha	0.63	0.61	0.65

Table 6.19: Reliability coefficient (Cronbach's alpha) at the booklet level in the knowledge test

The reliability coefficient at the booklet level for grades 5 and 6 combined was 0.63. The Cronbach's alpha was 0.61 in grade five and it was 0.65 in grade six. Many items had a corrected item-total correlation lower than 0.2 (see appendix, table 49).

Discussion

The reliability coefficients at the unit level in the knowledge test were very low in all four units. All of the Cronbach's alphas were lower than 0.5 which means that the test is not reliable at the level of the unit.

The comparison of the reliability coefficient for grade five and grade six revealed that the Cronbach's alpha was a slightly higher in grade six than in grade five, but not considerably. This rules out the possibility of looking at a sub-sample of grade six students for the sake of increasing the reliability of the knowledge test to an acceptable limit.

Also, the reliability coefficient at the booklet level stayed below the acceptable level of 0.7 for grade five, grade six and both grades combined. This is an unfortunate – and unforeseen – finding. In fact, in the previous trial of the knowledge test, the Cronbach's alpha for the same test was 0.74 (cf. table 5.14). The difference between the two reliability coefficients can be traced back to slight differences between the two samples. The low reliability of the knowledge test limits the trustworthiness of the following calculation of correlations between pre-knowledge and the competencies.

2.3.2.2. Reliability of the competency test

a) Reliability at the unit level

Item	Seed germination	Chicken eggs	Apple wine	Baking bread
H1	0.53	0.41	0.59	0.50
H2	0.48	0.42	0.56	0.52
D1	0.50	0.38	0.51	0.46
D2	0.36	0.30	0.62	0.50
E1	0.36	0.32	0.47	0.36
E2	0.30	0.35	0.37	0.33
Cronbach's alpha for unit	0.69	0.63	0.77	0.71

Table 6.20: Reliability at the unit level: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient at the unit level in the competency test ranged from 0.63 to 0.77. Among them, two units (unit 3 and unit 4) had a Cronbach's alpha higher than 0.7. In particular, unit 3 had a high reliability coefficient with a Cronbach's alpha of 0.77. And two other units had Cronbach's alphas lower than 0.7. Unit 2 had the lowest reliability coefficient.

However, all items in the four units had a corrected item-total correlation higher than 0.2. Many items had a corrected item-total correlation even higher than 0.5. In particular, unit 3 had four items out of six with corrected item-total correlations higher than 0.5.

Unit	Cronbach's alpha (Both grades combined)	Cronbach's alpha (Grade five)	Cronbach's alpha (Grade six)
1. Seed germination	0.69	0.56	0.71
2. Chicken eggs	0.63	0.50	0.65
3. Apple wine	0.77	0.71	0.78
4. Baking bread	0.71	0.61	0.73

Table 6.21: Reliability coefficient (Cronbach's alpha) at the unit level in the competency test

A comparison of the reliability coefficient at the unit level in grade five and grade six, table 6.21 shows that, the reliability coefficient was much higher in grade six than in grade five. The Cronbach's alpha at the unit level in grade five ranged from 0.50 to 0.71, while it ranged from 0.65 to 0.78 in grade six. In grade five, only unit 3 had a Cronbach's alpha higher than 0.7 and it was 0.5 in unit 2. In contrast, in grade six, three out of four units had a Cronbach's alpha higher than 0.7. Only unit 2 had a Cronbach's alpha of 0.65.

Discussion

The reliability coefficient at the unit level for all four units was higher than 0.6. However, only for two units (unit 3 and unit 4) was the Cronbach's alpha higher than 0.7, although all items had a corrected item-total correlation higher than 0.2.

A comparison of the reliability coefficients at the unit level for the two groups of students showed that the Cronbach's alpha at the unit level in 6 grade was much higher than in 5 grade. In grade five, three out of four units had a Cronbach's alpha lower than 0.7. In contrast, in grade six only unit 2 had a Cronbach's alpha lower than 0.7.

Therefore, the competency test is more appropriate for students in grade six than for students in grade five.

b) The reliability for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level

	Item	Scale “forming hypotheses”	Scale “data analysis”	Scale “planning experiments”
Unit 1:	1	0.46	0.34	0.34
Seed germination	2	0.44	0.33	0.41
Unit 2:	1	0.37	0.32	0.41
Chicken eggs	2	0.50	0.25	0.41
Unit 3:	1	0.57	0.48	0.44
Apple wine	2	0.51	0.49	0.41
Unit 4:	1	0.52	0.43	0.36
Baking bread	2	0.54	0.44	0.44
Cronbach's alpha		0.78	0.69	0.71

Table 6.22: Reliability for the scales “forming hypotheses”, “data analysis”; and “planning Experiments” at the booklet level: Corrected item-total correlation and Cronbach's alpha

The reliability coefficient for the scale “forming hypotheses” was quite high. The Cronbach's alpha was 0.78. Moreover, the corrected item-total correlations for all items were higher than 0.3. Among them, five items out of eight had a corrected item-total correlation higher than 0.5.

The reliability coefficient for the scale “data analysis” was 0.69. The corrected item-total correlation coefficients for all items in this scale were higher than 0.2.

The reliability coefficient for the scale “planning experiments” was also quite high. The Cronbach's alpha was 0.71. The corrected item-total correlations for all items were higher than 0.3.

	Scale “forming hypotheses”	Scale “data analysis”	Scale “planning experiments”
Both grades five and six combined (n = 1006)	0.78	0.69	0.71
Grade five (n = 253)	0.76	0.63	0.56
Grade six (n = 753)	0.78	0.70	0.74

Table 6.23: Cronbach’s alpha for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level _ Three groups of students

The reliability coefficient for the three scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level ranged from 0.69 to 0.78. Among them the reliability coefficient for the scale “forming hypotheses” was the highest and only the scale “data analysis” had a Cronbach’s alpha of below 0.7 in grade five and grade five and 6 combined.

Moreover, the corrected item-total correlation coefficients for all items in all the three scales were higher than 0.2. In particular, in the scale “forming hypotheses”, seven items out of eight had a corrected item-total correlation higher than 0.4.

A comparison of the reliability coefficients for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level in grade five and 6, shows that the reliability coefficient was lower in grade five than in grade six in all the three scales (cf. table 6.23). In grade five, the Cronbach’s alphas ranged from 0.56 to 0.76. Among them, only the scale “forming hypotheses” had a Cronbach’s alpha higher than 0.7. In contrast, in grade six the Cronbach’s alphas ranged from 0.70 to 0.78. All three scales had a Cronbach’s alpha higher than 0.7.

Discussion

The reliability coefficients for the scale “forming hypotheses” at the booklet level were quite high. The Cronbach’s alpha was 0.78 and all items had corrected item-total correlation coefficients higher than 0.3. Moreover, the Cronbach’s alpha for the scale “planning experiments” was also higher than 0.7. Only the Cronbach’s alpha for the scale “data analysis” was 0.69, although all items had a corrected item-total correlation higher than 0.2.

However, all the three scales were also reliable enough and they could be used to calculate the correlation between pre-knowledge and the three dimensions in experimentation and within the three dimensions.

Besides, the reliability coefficient for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level in grade five and in grade six was calculated separately. In grade five, the reliability for scale “forming hypotheses” was

also quite high. The Cronbach's alpha was higher than 0.7. However, the reliability coefficient for the scale "data analysis" was not high enough. The Cronbach's alpha was 0.63. In particular, the reliability coefficient for the scale "planning experiments" was low with a Cronbach's alpha of 0.56.

However, in grade six, the reliability coefficient for the two scales "forming hypotheses" and "planning experiments" was higher than 0.7. In particular, the Cronbach's alpha for the scale "forming hypotheses" was 0.78. Therefore, we could calculate the correlations between the three dimensions in experimentation and between pre-knowledge and the three dimensions only for grade six, or for both grades five and six combined, but not only for grade five.

Furthermore, we can also do latent class analysis for each dimension of experimentation based on both grades combined or based on grade six.

c) Reliability of the booklet (all items of the four units combined)

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H1	0.51	Apple wine	H1	0.58
	H2	0.48		H2	0.55
	D1	0.43		D1	0.52
	D2	0.41		D2	0.59
	E1	0.37		E1	0.52
	E2	0.39		E2	0.46
Chicken eggs	H1	0.42	Baking bread	H1	0.52
	H2	0.56		H2	0.54
	D1	0.37		D1	0.47
	D2	0.29		D2	0.52
	E1	0.42		E1	0.38
	E2	0.41		E2	0.42

Cronbach's alpha = 0.88

Table 6.24: Reliability of the booklet (all items of the four units combined): corrected item-total correlation and Cronbach's alpha

The reliability coefficient for the three scales of the complete test booklet was very high in the main test. The Cronbach's alpha was 0.88. Moreover, all items had a corrected item-total correlation higher than 0.2. In particular, all items of unit 3 and five out of six items in unit 4 had a corrected item-total correlation higher than 0.4.

	Both grades 5 and 6 combined	Grade five	Grade six
Cronbach's alpha	0.88	0.84	0.89

Table 6.25: Reliability coefficients of the booklet (all items of the four units combined) _ In the three groups of students

The reliability coefficients for the three scales of the complete test booklet were very high in grade five as well as in grade six. The Cronbach's alphas were higher than 0.8. In particular, in grade six the Cronbach's alpha was 0.89.

Discussion

At the booklet level, the reliability coefficient for all three scales “forming hypotheses”, “data analysis” and “planning experiments” combined was very high in all three groups of students, in grade five, in grade six and in both grades combined). All Cronbach's alphas were higher than 0.8. Furthermore, in all three groups of students the corrected item-total correlation for most items was higher than 0.2 (see appendix, table 55).

Therefore, the competency test at the booklet level was highly reliable and could be used for the latent class analysis.

DISCUSSION

The reliability coefficient of the knowledge test was not so high. In particular, at the unit level, all Cronbach's alphas were lower than 0.5 and most items had a corrected item-total correlation lower than 0.2. However, the reliability coefficient at the booklet level was not so low. The Cronbach's alpha was higher than 0.6 in all three groups of students (grade five, grade six and both grades combined). Thus, the knowledge test could be used at the booklet level to calculate the correlations between pre-knowledge and the three dimensions in experimentation.

On the other hand, the reliability coefficient in the competency test was quite high, in particular, at the booklet level. All three Cronbach's alphas in the three groups of students were higher than 0.8. At the unit level, the reliability coefficient was higher than 0.7 in two units, whereas it was lower than 0.7 in the other two units. However, at the unit level, the scales for each dimension in experimentation were not very long, as there were only two items in each dimension. Accordingly, since the reliability coefficient at the unit level in the knowledge test was very low, we did not calculate the correlation between pre-knowledge and the three dimensions at the unit level.

At the booklet level, the reliability coefficient for the three scales “forming hypotheses”, “data analysis” and “planning experiments” was quite high. The Cronbach’s alphas for the two scales were higher than 0.7. Only in the scale “data analysis” the Cronbach’s alpha slightly below 0.7. Besides, the reliability coefficients for the three scales at the booklet level in grade six were also higher than 0.7.

Therefore, at the booklet level, the competency test and the knowledge test can be considered reliable and can be used to calculate the correlations between the students’ pre-knowledge and the three dimensions as well as the correlations within the three dimensions in experimentation. Moreover, for the competency test, it was also possible to calculate a latent class analysis for all items in the booklet and for items in each dimension in experimentation.

2.4. Latent class analysis for the competency test

As in pre-test 1, the Winmira program was used to do the latent class analysis and to assign students into different classes of competencies in experimentation.

2.4.1. Latent class analysis for all items in the three dimensions of experimentation

The competency test consists of 24 items which are organized in the three dimensions. We tested three probable test models (two-, three- and four-class models). Based on these models the students were assigned into two, three or four classes of competencies in experimentation. Table 6.26 shows the associated model value index for the tested models.

Model	BIC-Index	CAIC-Index
Classified model with two latent classes (LCA 2)	26317	26416
Classified model with three latent classes (LCA 3)	26437	26586
Classified model with four latent classes (LCA 4)	26649	26848

Table 6.26: Value of the probable tested models

Like in pre-test 1, the table of value index of the tested models (table 6.26) shows that the two-class model was the best with both the lowest BIC index and CAIC index and the three-class model was the second option. However, as pointed out in the discussion of table 3.43 in pre-test 1, the two-class model also provided differentiation of the quantification and the item profile for this model was not as good as the three-class model. The three-class model delivered a stronger differentiation of the item profiles than the two-class model. Therefore, we looked at the three-class model more closely.

As pointed out in chapter 3, in order to ensure the reliability of the selected test model, especially when it was not the best solution, we looked at the second criterion which was the mean of response probability of the students (cf. table 6.27).

Class	Expected size	Mean probability	Assignment probability	Assignment probability	Assignment probability
			Class 1	Class 2	Class 3
1	0.433	0.927	0.927	0.019	0.054
2	0.335	0.947	0.015	0.947	0.039
3	0.223	0.890	0.064	0.046	0.890

Table 6.27: Mean of response probability for the three latent class model

After assigning the students to three classes of competency in experimentation, the mean of maximum response probability of one class in three cases was at least 89%. The alternative second highest and third highest assignment probabilities did not exceed a value of 6.4% for any class. Thus, the three latent classes model was also reliable.

Therefore, the three-class model was used to assign students to three classes of competencies in experimentation.

2.4.2. The relationship between groups of students and the three dimensions of experimentation

We assigned students to two and then three classes of competencies and calculated the mean score for each dimension in experimentation for each class. Figures 6.1 and 6.2 show the resulting relationship between the groups of students and the three dimensions in experimentation.

Two-latent-class model

We used the two latent class analysis model and assigned students into two classes of competencies in experimentation. The mean scores for each dimension in experimentation are shown in table 6.28.

	Search in the hypothesis space	Data analysis	Search in the experiment space
Class 1	9.09	9.63	9.27
Class 2	14.38	14.17	12.80

Table 6.28: Mean score for each dimension in experimentation corresponding to two classes of students

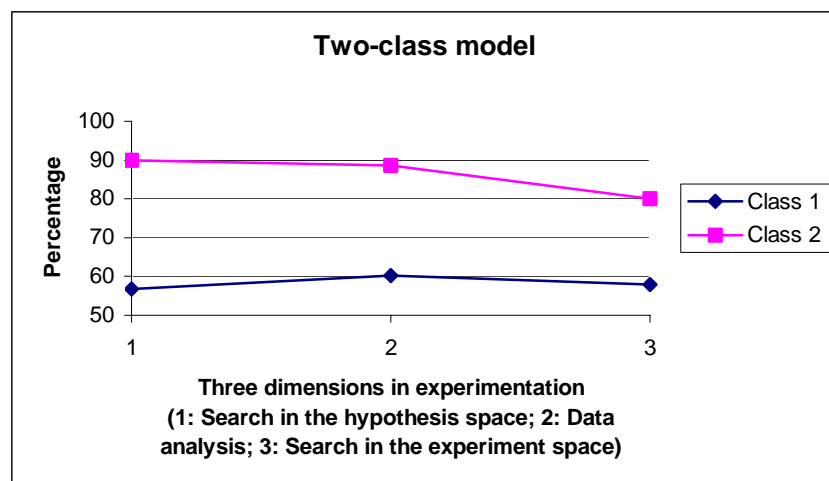


Figure 6.1: The relationship between the groups of students and the three dimensions in experimentation (The two-class model)

Figure 6.1 shows that based on a two latent class model, the students in class 2 always solved all the three dimensions of experimentation better than those in class 1.

In class 1 (33.5% of the students), all the three dimensions were correctly answered with a probability of 56% to 60%.

In class 2 (66.5% of the students), the students solved the items in the three dimensions with a probability of over 80%.

Three-latent-class model

Also, the three latent class analysis model was used to assign students to three classes of competencies in experimentation. The mean scores for each dimension in experimentation are shown in table 6.28.

	Search in the hypothesis space	Data analysis	Search in the experiment space
Class 1	9.09	9.69	9.27
Class 2	13.24	13.43	11.94
Class 3	14.96	14.50	13.23

Table 6.29: Mean score for each dimension in experimentation corresponding to three classes of students

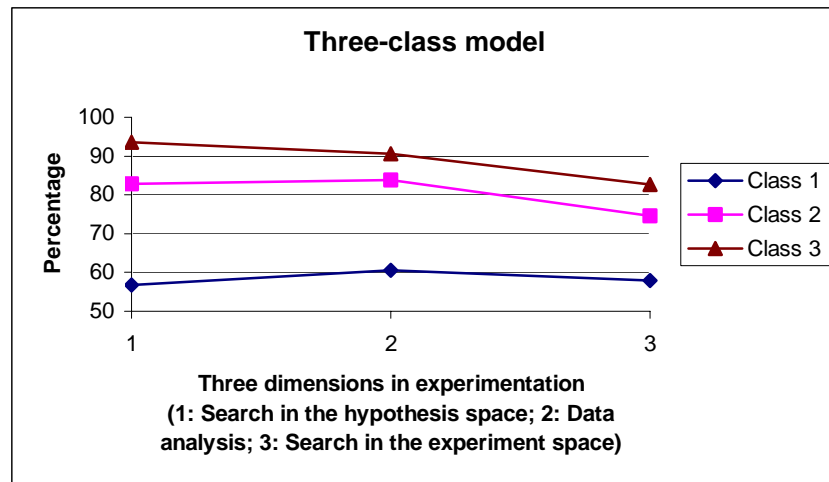


Figure 6.2: The relationship between the groups of students and the three dimensions in experimentation (The three-class model)

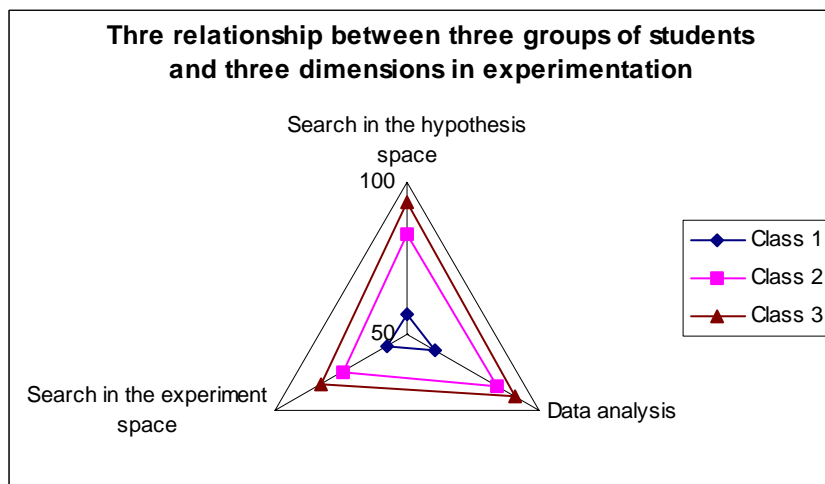


Figure 6.3: The relationship between the groups of students and the three dimensions in experimentation (The three-class model)

Based on the three-class model, students in class 3 always solved the items in all the three dimensions better than the students in class 1 and class 2. The students in class 2 were also better than the students in class 1 in all the three dimensions.

As in the two-class model, in class 1 (33.5% of the students), all the items in the three dimensions were correctly answered with a probability of 56% to 60%. The mean probability to answer the items correctly in class 1 was 58%.

In class 2 (22.3% of the students), the two dimensions “search in the hypothesis space” and “data analysis” were correctly answered with a probability of over 80%. Only the items in the dimension “search in the experiment space” were correctly answered with a

probability of 74%. The mean probability for solving items correctly in this class was 80%.

In class 3 (44.2% of the students), only items in the dimension “search in the experiment space” were correctly answered with a probability of lower than 90%. Items in the two other dimensions were correctly answered with a probability of 90% and 93%. The mean probability to answer correctly in this class was nearly 90%.

Discussion

If students are assigned to two classes, the students in class 2 always outdo the students in class 1 in all the three dimensions of experimentation. From these analyses follows that there are two types of students who resemble each other in their patterns of responding to the items. One group of students possesses a high level of competency in experimentation, the other a lower.

Furthermore, students are assigned to three classes, students in class 3 always outdo the two other classes and students in class 2 always outdo the students in class 1 in all the three dimensions of experimentation.

2.4.3. Latent class analysis for each dimension in experimentation

In the competency test, there were eight variables for each dimension in experimentation.

For each dimension of experimentation, students were assigned to different classes of competency. Three latent-class analysis models were also looked at, namely a two-class model, a three-class model and a four-class model. Similar to the analysis for all items in experimentation, the two-class model was the best with the lowest BIC index and CAIC index and the three-class model was the second option for all the three dimensions of experimentation (see appendix, tables 57-59).

Item profiles for the dimension “search in the hypothesis space”

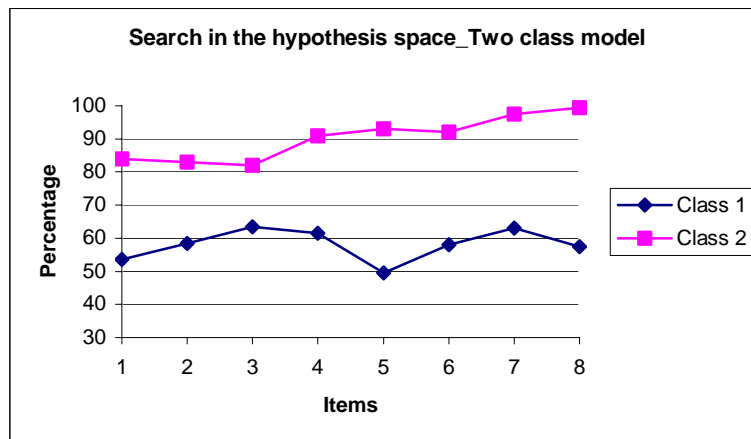


Figure 6.4: Item profile for the two-class model in “search in the hypothesis space”

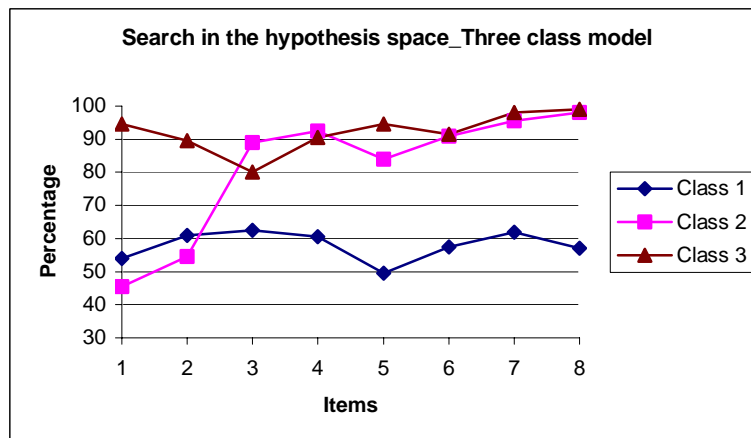


Figure 6.5: Item profile for the three-class model in “search in the hypothesis space”

In the dimension “search in the hypothesis space”, based on the two-class model, students were assigned to two classes. The students in class 2 (62.4% of the students) always solved all items better than those in class 1 (37.6% of the students). In class 1, most of items were correctly answered with a probability of 50% to 60%. In class 2, all items were correctly answered with a probability of over 80%.

However, if students were assigned to three classes, the students in class 3 solved six items out of eight better than the students in classes 1 and 2. Further, the students in class 1 did most items worse than the students in the two other classes, but students in class 2 did some items better than the students in class 3 and some items worse than the students in class 1.

Class 1 consisted of 36.1% of the students. In this class as well as in class 1 in the two-class model, most items were correctly answered with a probability of 50% to 60%.

Class 2 contained 15.3% of the students. In this class, items 1 and 2 were correctly answered with probabilities of 45% and 55%, while items 3 and 4 were correctly answered with high probabilities of 89% and 92%. Items 8 and 9 were correctly answered with probabilities of 95% and 98%. The mean probability for answering all of the items in this group correctly was 81.2%. So, in class 2, unit 1 was more difficult than the three other units. Class 3 consisted of 48.7% of the students. In this class, all items were correctly answered with a probability of over 80%. Among them only item 3 was correctly answered with a likelihood of below 90%. The mean probability for answering all of the items in this group correctly was 92%.

Item Profiles for the dimension “data analysis”

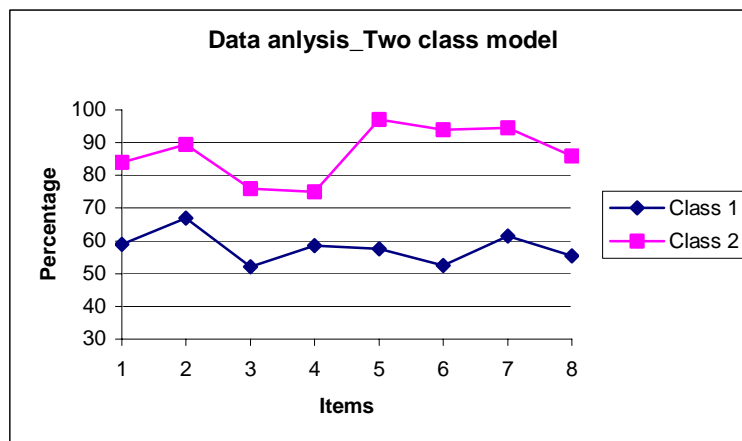


Figure 6.6: Item profile for the two-class model in “data analysis”

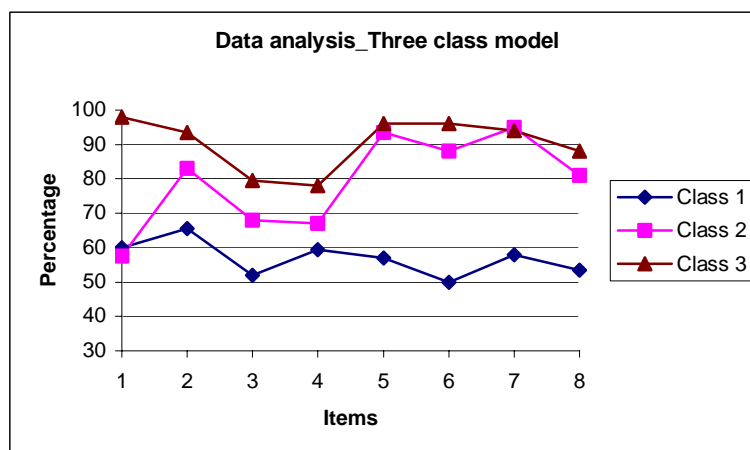


Figure 6.7: Item profile for the three-class model in “data analysis”

As in the dimension “search in the hypothesis space”, students were assigned to two classes in the dimension “data analysis”. Students in class 2 (70% of the students)

always solved all items better than the students in class 1 (30% of the students). In class 1, most items were correctly answered with a probability of 52% to 60%. Only item 2 was correctly answered with a probability of 65%. However, in class 2 all items were correctly answered with a probability of 75% to 97%. Among them, only items 3 and 4 were correctly answered with a probability of below 80%.

However, if students were assigned to three classes, students in class 3 solved all items but item 7 (Unit: baking bread) better than the students in class 2 and class 1. Students in class 2 did all items (except for item 1) better than the students in class 1. In contrast, students in class 1 solved item 1 better than the students in class 2. Students in class 1 and class 3 were also more consistent than those in class 2.

Class 1 consisted of 26.4% of the students. In this class most of the items were correctly answered with a probability of 50% to 60%. Only item 2 was answered with a probability of over 60%. The students in this class had a medium level of competency in evaluating evidence.

Class 2 contained 28.4% of the students. In this class, all items were correctly answered with a probability of 57% to 95%. Among them items 1, 3 and 4 were correctly answered with a probability of below 70% and items 5 and 7 were correctly answered with a probability of over 90%. The mean probability to answer correctly in this class was 79%. Thus, students in this class had quite a high level of competency, but they were inconsistent in evaluating evidence.

Class 3 consisted of 45.2% of the students. In this class most items were correctly answered with a probability of over 80%, except for item 4 which is characterized by a probability of 78%. The mean probability to answer correctly in this class was 90%.

Item profiles for the dimension “search in the experiment space”

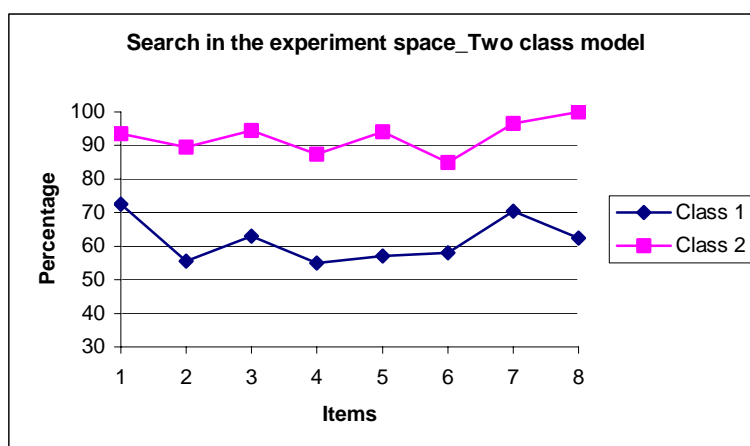


Figure 6.8: Item profile for the two-class model in “search in the experiment space”

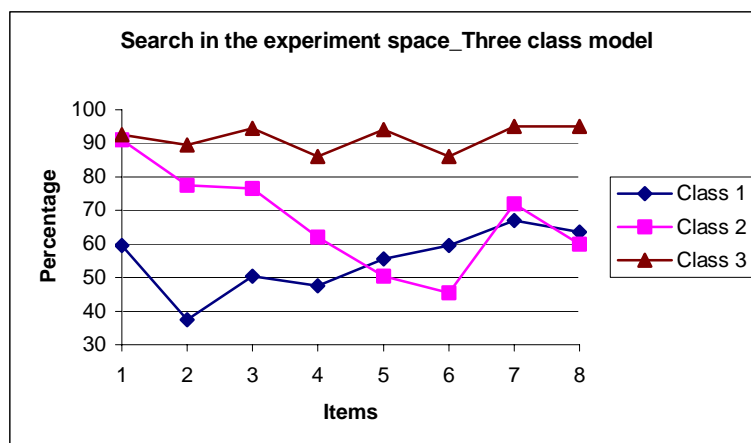


Figure 6.9: Item profile for the three-class model in “search in the experiment space”

Similarly, in the dimension “search in the experiment space” students were assigned to two classes. Students in class 2 always did all items better than the students in class 1.

In class 1 (65.8% of the students), all items were correctly answered with a probability of 55% to 72% and the mean probability to answer correctly in this class was 61%. This means, the students in this class had a medium level of competency in planning experiments.

In class 2 (34.2% of the students), all items were correctly answered with a probability of 85% to 96%. Students in this class had a high level of competency in planning experiments.

If students were assigned into three classes, students in class 3 solved all eight items better than the students in class 2 and class 1 and students in class 2 did five items out of eight better than the students in class 1. (Items 5 and 6 of unit 3: Apple wine and item 8 of unit 4: Baking bread is an exception to this.

Class 1 consisted of 37.8% of the students. In this class, most items were correctly answered with a probability of over 50%. Only items 2 and 4 were answered correctly with a probability below 50%, while items 7 and 8 were answered correctly with a probability of over 60%. The mean probability to answer correctly in this class was 55%.

Class 2 consisted of 22.2% of the students. In this class, most items were correctly answered with a probability of over 60%. However, items 5 and 6 were correctly answered with a probability of below 60% and item 1 was correctly answered with a probability of over 90%. The mean probability to answer correctly in this class was 67%. Students in this class were more competent in planning experiments than class 1, but they were inconsistent in all items.

Class 3 contained 40.1% of the students. In this class, all items were correctly answered with a probability of over 85%.

Discussion

For each dimension of experimentation, when students are assigned to two classes, students in class 2 consistently solve all items better than those in class 1.

However, if we assign students to three classes of competencies, in class 3 in all the three dimensions of experimentation students were correctly answered most items with a probability of over 80%. In class 1, most items were correctly answered from 50% to 60%; while in class 2 some items were correctly answered with a probability of over 80% and some items were correctly answered below 50%.

Students in class 1 had a medium level of competency, the students in class 3 solved all items, but the ones in class 2 were inconsistent in all cases, sometimes they do well and sometimes they do badly.

2.4.4. Cross tables

2.4.4.1. Cross tables between two of the three dimensions in experimentation

a) The two-class model

We assigned students to two classes of competency in each dimension of experimentation based on the two latent classes model, the relationships between each two of the three dimensions in experimentation were shown in the following tables:

Cross table between “search in the hypothesis space” and “data analysis”

			Data analysis		Total
			1	2	
Search in the hypothesis space	1	Number	213	102	315
		% of Hypothesis	67.6%	32.4%	100.0%
		% of Data analysis	85.9%	17.1%	37.3%
		% of Total	25.2%	12.1%	37.3%
	2	Number	35	495	530
		% of Hypothesis	6.6%	93.4%	100.0%
		% of Data analysis	14.1%	82.9%	62.7%
		% of Total	4.1%	58.6%	62.7%
Total	Number	248	597	845	
	% of Hypothesis	29.3%	70.7%	100.0%	
	% of Data analysis	100.0%	100.0%	100.0%	
	% of Total	29.3%	70.7%	100.0%	

Table 6.30: Cross table between “Data analysis” and “Search in the hypothesis space”

The number of students who gained class 1 in both dimensions “search in the hypothesis space” and “data analysis” was 25.2%, the students who achieved class 2 in both these two dimensions was 58.6%. That means, 83.8% of the students was consistent in both “search in the hypothesis space” and “data analysis” and only 16.2% of the students was inconsistent.

On the other hand, 93.4% of the students who gained class 2 in “search in the hypothesis space” also gained class 2 in “data analysis”. 82.9% of the students who achieved class 2 in “data analysis” also achieved class 2 in “search in the hypothesis space”.

Besides, 67.6% of the students who did not solve the tasks in “search in the hypothesis space” did not solve the ones in “data analysis” either and in contrast, 85.9% of the students who gained level 1 in “data analysis” also achieved level 1 in “search in the hypothesis space”.

Cross table between “data analysis” and “search in the experiment space”

			Search in the experiment space		Total
			1	2	
Data analysis	1	Number	203	13	216
		% of Data analysis	94.0%	6.0%	100.0%
		% of Experiment	41.5%	5.0%	28.9%
		% of Total	27.1%	1.7%	28.9%
	2	Number	286	246	532
		% of Data analysis	53.8%	46.2%	100.0%
		% of Experiment	58.5%	95.0%	71.1%
		% of Total	38.2%	32.9%	71.1%
Total	Number	489	259	748	
	% of Data analysis	65.4%	34.6%	100.0%	
	% of Experiment	100.0%	100.0%	100.0%	
	% of Total	65.4%	34.6%	100.0%	

Table 6.31: Cross table between “Data analysis” and “Search in the experiment space”

The number of students who was consistent in both dimensions “data analysis” and “search in the experiment space” was 60% and 40% of the students were inconsistent.

95% of the students who gained class 2 in “search in the experiment space” also achieved class 2 in “data analysis”; however, only 46.2% of the students who obtained class 2 in “data analysis” also achieved class 2 in “search in the experiment space”.

Moreover, 94% of the students who gained level 1 in “data analysis” also gained level 1 in “search in the experiment space”, while only 41.5% of the students who obtained level 1 in “search in the experiment space” also gained level 1 in “data analysis”.

Cross table between “search in the hypothesis space” and “search in the experiment space”

		Search in the experiment space		Total	
		1	2		
Search in the hypothesis space	1	Number	254	22	276
		% of Hypothesis	92.0%	8.0%	100.0%
		% of Experiment	51.2%	8.4%	36.4%
		% of Total	33.5%	2.9%	36.4%
	2	Number	242	240	482
		% of Hypothesis	50.2%	49.8%	100.0%
		% of Experiment	48.8%	91.6%	63.6%
		% of Total	31.9%	31.7%	63.6%
Total	Number	496	262	758	
	% of Hypothesis	65.4%	34.6%	100.0%	
	% of Experiment	100.0%	100.0%	100.0%	
	% of Total	65.4%	34.6%	100.0%	

Table 6.32: Cross table between “Search in the hypothesis space” and “Search in the experiment space”

The number of students who was consistent in both dimensions “search in the hypothesis space” and “search in the experiment space” was 65.2% and 34.8% of the students were inconsistent.

Only 49.8% of the students who gained class 2 in “search in the hypothesis space” also achieved class 2 in “search in the experiment space”. In contrast, 91.6% of the students who obtained class 2 in “search in the experiment space” also achieved class 2 in “search in the hypothesis space”.

Furthermore, 92% of the students who reached level 1 in “search in the hypothesis space” also gained level 1 in “search in the experiment space”. However, only 51.2% of the students who did not solve the tasks in “search in the experiment space” did not solve the ones in “search in the hypothesis space” either.

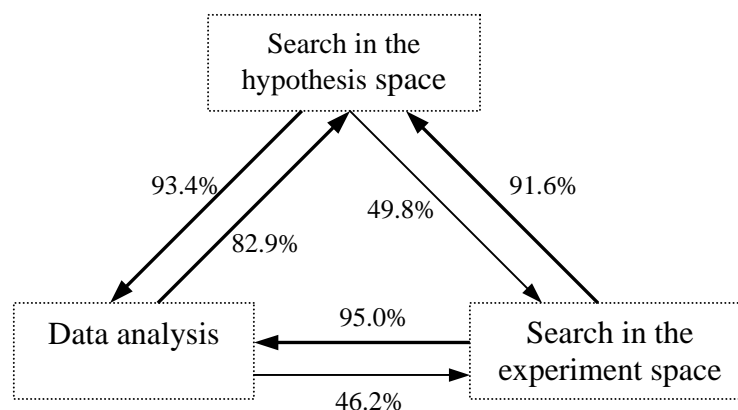


Figure 6.10: Percentage of students who gained class 2 in one dimension also gained level 2 in other dimension (The two-class model)

Discussion

If we assign students into two classes of competencies in experimentation, the percentage of students who was consistent in both dimensions “search in the hypothesis space” and “data analysis” was 83.8%, while the number of students who was consistent in both dimensions “data analysis” and “search in the experiment space” was only 60.0% and 65.2% of the students were consistent in both dimensions “search in the hypothesis space” and “search in the experiment space”.

Furthermore, the relationships between “data analysis” and “search in the experiment space” and between “search in the hypothesis space” and “search in the experiment space” were similar and not as close as between “search in the hypothesis space” and “data analysis”.

Concretely, figure 6.10 shows that 93% of the students who gained class 2 in “search in the hypothesis space” also gained class 2 in “data analysis”, and 82% of the students who gained class 2 in “data analysis” also gained class 2 in “search in the hypothesis space”. However, only 49% of the students who got class 2 in “search in the hypothesis space” and 46% of the students who achieved class 2 in “data analysis” also gained class 2 in “search in the experiment space”. In contrast, more than 90% of the students who do well in “search in the experiment space” can also do well in two other dimensions.

This indicated that the dimension “search in the experiment space” was influenced by methodological knowledge and if students show ability in doing experiments, they can achieve high level in all the three dimensions of experimentation. Besides, “search in the hypothesis space” and “data analysis” influenced each other and may be influenced by the content knowledge. If students solve the tasks in “search in the hypothesis space”, they can also solve the ones in “data analysis” and vice versa.

b) The three-class model

We also assigned students to three classes of competency in each dimension of experimentation based on the three latent class model. The relationships between each two of the three dimensions in experimentation were shown as the following tables:

Cross table between “search in the hypothesis space” and “data analysis”

		Data analysis			Total	
		1	2	3		
Search in the hypothesis space	1	Number	174	86	40	300
		% of Hypothesis	58.0%	28.7%	13.3%	100.0%
		% of Data analysis	79.5%	36.3%	10.3%	35.5%
		% of Total	20.6%	10.2%	4.7%	35.5%
	2	Number	25	67	36	128
		% of Hypothesis	19.5%	52.3%	28.1%	100.0%
		% of Data analysis	11.4%	28.3%	9.3%	15.1%
		% of Total	3.0%	7.9%	4.3%	15.1%
	3	Number	20	84	313	417
		% of Hypothesis	4.8%	20.1%	75.1%	100.0%
		% of Data analysis	9.1%	35.4%	80.5%	49.3%
		% of Total	2.4%	9.9%	37.0%	49.3%
Total	Number	219	237	389	845	
	% of Hypothesis	25.9%	28.0%	46.0%	100.0%	
	% of Data analysis	100.0%	100.0%	100.0%	100.0%	
	% of Total	25.9%	28.0%	46.0%	100.0%	

Table 6.33: Cross table between “Data analysis” and “Search in the hypothesis space”

The number of students who attained class 1 in both dimensions “search in the hypothesis space” and “data analysis” was 20.6%, the students who achieved class 2 in both these two dimensions was 7.9% and class 3 in both of dimensions was 37.0%. That means, 65.5% of the students were consistent in both “search in the hypothesis space” and “data analysis” and 34.5% of the students were inconsistent.

On the other hand, 75.1% of the students who attained class 3 in “search in the hypothesis space” also had class 3 in “data analysis”. 80.5% of the students who achieved class 3 in “data analysis” also achieved class 3 in “search in the hypothesis space”.

Besides, 58.0% of the students who did not solve the tasks in “search in the hypothesis space” did not solve the tasks in “data analysis” either and in contrast, 79.5% of the students who achieved only class 1 in “data analysis” also achieved class 1 in “search in the hypothesis space”.

Cross table between “data analysis” and “search in the experiment space”

			Search in the experiment space			Total
			1	2	3	
Data analysis	1	Number	121	57	14	192
		% of Data analysis	63.0%	29.7%	7.3%	100.0%
		% of Experiment	42.9%	34.8%	4.6%	25.7%
		% of Total	16.2%	7.6%	1.9%	25.7%
	2	Number	85	50	70	205
		% of Data analysis	41.5%	24.4%	34.1%	100.0%
		% of Experiment	30.1%	30.5%	23.2%	27.4%
		% of Total	11.4%	6.7%	9.4%	27.4%
	3	Number	76	57	218	351
		% of Data analysis	21.7%	16.2%	62.1%	100.0%
		% of Experiment	27.0%	34.8%	72.2%	46.9%
		% of Total	10.2%	7.6%	29.1%	46.9%
Total	Number	282	164	302	748	
	% of Data analysis	37.7%	21.9%	40.4%	100.0%	
	% of Experiment	100.0%	100.0%	100.0%	100.0%	
	% of Total	37.7%	21.9%	40.4%	100.0%	

Table 6.34: Cross table between “Data analysis” and “Search in the experiment space”

The number of students consistent in both of dimensions “data analysis” and “search in the experiment space” was 52% and 48% of the students were inconsistent.

72.2% of the students who were class 3 in “search in the experiment space” achieved class 3 in “data analysis”. In contrast, 62.1% of the students who obtained class 3 in “data analysis” also had class 3 in “search in the experiment space”.

On the other hand, 63% of the students who were class 1 in “data analysis” did not solve dimension “search in the experiment space” either. However, only 42.9% of the students who obtained class 1 in “search in the experiment space” also reached class 1 in “data analysis”.

Cross table between “search in the hypothesis space” and “search in the experiment space”

			Search in the experiment space			Total
			1	2	3	
Search in the hypothesis space	1	Number	154	83	28	265
		% of Hypothesis	58.1%	31.3%	10.6%	100.0%
		% of Experiment	53.3%	50.6%	9.2%	35.0%
		% of Total	20.3%	10.9%	3.7%	35.0%
	2	Number	46	18	50	114
		% of Hypothesis	40.4%	15.8%	43.9%	100.0%
		% of Experiment	15.9%	11.0%	16.4%	15.0%
		% of Total	6.1%	2.4%	6.6%	15.0%
	3	Number	89	63	227	379
		% of Hypothesis	23.5%	16.6%	59.9%	100.0%
		% of Experiment	30.8%	38.4%	74.4%	50.0%
		% of Total	11.7%	8.3%	29.9%	50.0%
Total	Number	289	164	305	758	
	% of Hypothesis	38.1%	21.6%	40.2%	100.0%	
	% of Experiment	100.0%	100.0%	100.0%	100.0%	
	% of Total	38.1%	21.6%	40.2%	100.0%	

Table 6.35: Cross table between “Search in the hypothesis space” and “Search in the experiment space”

The number of students consistent in both of dimensions “search in the hypothesis space” and “search in the experiment space” was 52.6% and 47.4% of the students were inconsistent.

59.9% of the students who attained class 3 in “search in the hypothesis space” also achieved class 3 in “search in the experiment space” and in contrast, 74.4% of the students who obtained class 3 in “search in the experiment space” also achieved class 3 in “search in the hypothesis space”.

On the other hand, 58.1% of the students who achieved class 1 in “search in the hypothesis space” also attained class 1 in “search in the experiment space”. 53.3% of the students who did not solve the tasks in “search in the experiment space” did not solve the ones in “search in the hypothesis space” either.

2.4.4.2. Cross tables between the students’ pre-knowledge and the three dimensions of experimentation

We also assigned students into three classes based on the sum scores in the knowledge test. Students in group 1 had total scores from 0 to 5 points. The ones in group 2 achieved six to ten points and the others in group 3 eleven to fifteen points.

We did the cross table between pre-knowledge and the three dimensions of experimentation based on the three class model.

Combining the cross tables between the three dimensions in experimentation, we had the following figure:

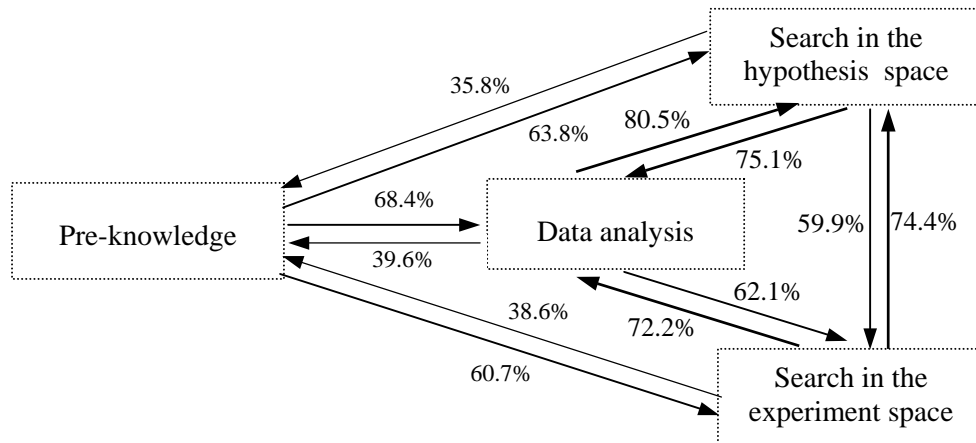


Figure 6.11: Relationship between pre-knowledge and the three dimensions in experimentation and within the three dimensions ($n = 1006$): Percentage of students who were in group 3 in the knowledge test also attained class 3 in each dimension of experimentation and the percentage of students who achieved class 3 in one dimension also achieved class 3 in another dimension

Figure 6.11 shows that, the number of students who had group 3 in pre-knowledge – where students have good content knowledge - also had class 3 in the three dimensions with a probability of over 60%. In contrast, only 35% to 39% of the students who attained class 3 in each of the three dimensions of experimentation – where students have good ability in doing experiments - also had group 3 in pre-knowledge.

On the other hand, over 70% of the students achieved class 3 in “search in the hypothesis space” also got class 3 in “data analysis” and vice versa. Over 70% of the students who achieved class 3 in “search in the experiment space” also attained class 3 in two other dimensions. However, only 59% of the students who achieved level 3 in “search in the hypothesis space” and 62% who had level 3 in “data analysis” also achieved level 3 in “search in the experiment space”.

Discussion

If we assign students to three classes of competency in each dimension of experimentation and to three classes in pre-knowledge, 63% of the students who achieved group 3 in pre-knowledge also achieved level 3 in “search in the hypothesis space”. 68% of them also achieved level 3 in “data analysis” and 60% obtained level 3 in “search in the experiment space”. That means, if students have good content

knowledge, a probability of over 60% of them can also achieve the highest level in each of the three dimensions of experimentation.

However, only 35% of the students who attained level 3 in “search in the hypothesis space”, 39% of the students who achieved level 3 in “data analysis” and 38% who had level 3 in “search in the experiment space” also reached group 3 in pre-knowledge. Thus, if students have methodological knowledge, they can attain a high level in experimentation.

On the other hand, the cross tables 6.33 to 6.35 show that the number of students who were consistent in both dimensions “search in the hypothesis space” and “data analysis” was 65.5%, while the percentage of students who were consistent in two other combinations was only 52%. Furthermore, 75.1% of the students who had class 3 in “search in the hypothesis space” also achieved class 3 in “data analysis”. In contrast 80.5% of the students who attained class 3 in “data analysis” also attained class 3 in “search in the hypothesis space”. While only 59.9% who achieved class 3 in “search in the hypothesis space” and 62.1% of the students who achieved class 3 in “data analysis” also attained class 3 in “search in the experiment space”. This means, the dimensions “search in the hypothesis space” and “data analysis” were more closely related than that between “search in the hypothesis space” and “search in the experiment space” and between “data analysis” and “search in the experiment space”.

However, if students achieve class 3 in “search in the experiment space”, over 70% of them also achieve class 3 in “data analysis” and in “search in the hypothesis space”. This indicated that if students have good ability in doing experiments, they can do well in “search in the experiment space” and in two other dimensions.

DISCUSSION

Cross tables can be used to identify the degree of similarity between dimensions in experimentation. The results of the cross tables support our hypotheses, in particular for the sub-sample of students who achieved class 2 in the two-class model and class 3 in the three-class model. The findings supported our assumption that the dimensions “search in the hypothesis space” and “data analysis” related more closely than between the dimensions “search in the hypothesis space” and “search in the experiment space” and between “data analysis” and “search in the experiment space”. These findings also support the assumption that “search in the experiment space” requires methodological knowledge about the conventions of scientific experimentation whereas the other two dimensions are more dependent on the student’s pre-knowledge.

2.5. Correlations

2.5.1. Method

As in chapter 5, we summed the scores of the strong items at the booklet level in the knowledge test to calculate the sum score of the students' pre-knowledge. In the competency test, we calculated the total scores for each dimension of experimentation (Search in the hypothesis space, search in the experiment space and data analysis).

Afterwards, we calculated the correlation coefficients (Spearman) between the sum scores of the knowledge test and the total scores for each dimension of experimentation. On the other hand, based on the total scores in the knowledge test (from 1 to 15 points), we assigned the students into three different classes with the same scale length: Students in group 1 (16.9% of the students) had total scores from 1 to 5 points in the knowledge test, the ones in group 2 (58.5% of the students) had total scores from 6 to 10 and the students in group 3 (24.6% of the students) had 11 to 15 points in the knowledge test. Then we looked at the correlation between the students' pre-knowledge and the three dimensions and within the three dimensions in experimentation based on these three classes of students.

Besides, based on the three latent-class model, we assigned students to three classes in experimentation based on latent class analysis and looked at the correlations between the students' pre-knowledge and the three dimensions and within the three dimensions.

On the other hand, we also calculated the correlation between the students' knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation.

2.5.2. Findings

2.5.2.1. Correlations between the students' pre-knowledge and the three dimensions of experimentation

	Pre-knowledge * Search in the hypotheses space (K*H)	Pre-knowledge * Data analysis (K*D)	Pre-knowledge * Search in the experiment space (K*E)
Both grades five and six combined (n = 1006)	0.400**	0.385**	0.353**
Grade six (n = 753)	0.380**	0.364**	0.366**

Table 6.36: Correlations (Spearman) between the students' pre-knowledge and the three dimensions of experimentation

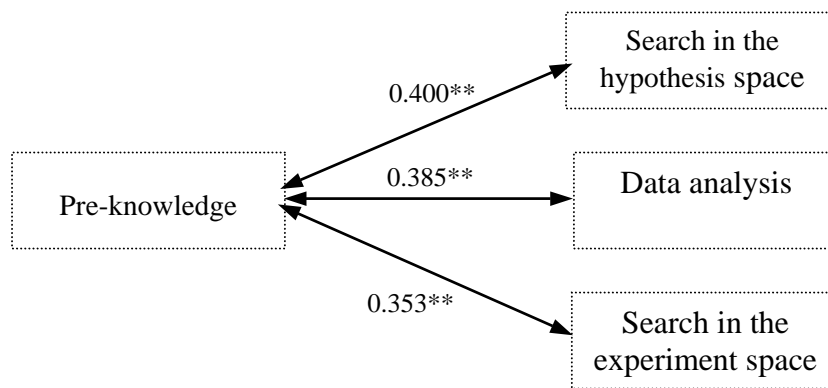


Figure 6.12: Correlations (Spearman) between the students' pre-knowledge and the three dimensions in experimentation (n = 1006)

	Z-score	1-tail p	2-tail p
(K*H) * (K*D)	0.397	0.3456	0.6912
(K*H) * (K*E)	1.227	0.1099	0.2198
(K*D) * (K*E)	0.083	0.2033	0.4067

Table 6.37: Difference between two independent correlation coefficients (Correlation coefficients between the students' pre-knowledge and the three dimensions in experimentation)

Table 6.36 shows that the correlation coefficients between the students' pre-knowledge and the three dimensions of experimentation ranged from 0.353 to 0.400. Among them, the correlation between the content knowledge and the "search in the experiment space" (K*E) was slightly lower than the correlations between pre-knowledge and the "search in the hypothesis space" (K*H) and between pre-knowledge and the "data analysis" (K*D). However, the three correlation coefficients between pre-knowledge and the three dimensions in experimentation were significant.

In grade six, the correlation coefficients between pre-knowledge and the three dimensions were also low; they ranged from 0.364 to 0.380. The correlation in the relationship pre-knowledge and the "search in the hypothesis space" was the highest.

Discussion

With the large number of students in the test, in the main study the correlations between students' pre-knowledge and the three dimensions were also low in all three relationships. Among them, the correlation between pre-knowledge and "search in the experiment space" was the lowest. However, when we look at the difference between two independent correlation coefficients, table 6.37 shows that, the correlation coefficients of each two of three combinations were not significant. That means, the

difference between three correlation coefficients of the students' pre-knowledge and the three dimensions was not considerable.

We expected that there would be high correlations between the students' pre-knowledge and the three dimensions in experimentation, especially, between the students' pre-knowledge and the dimensions "search in the hypothesis space" and "data analysis", because these dimensions were hypothesized to be influenced by biological pre-knowledge rather than by methodological knowledge. The findings did not completely support our expectations.

2.5.2.2. Relationships between the students' pre-knowledge and the three dimensions in experimentation in three classes of students based on the sum scores in content knowledge

We assigned students into three different classes based on total scores in the knowledge test and calculated the mean score for each dimension of experimentation in each group, resulting in the following table:

Classes of students (based on the knowledge test)	Search in the hypothesis space		Data analysis		Search in the experiment space	
	Mean scores	STD	Mean scores	STD	Mean scores	STD
1 (n = 169)	3.62	2.315	3.72	2.003	3.25	1.921
2 (n = 589)	4.94	2.353	4.92	2.089	4.20	2.174
3 (n = 247)	6.15	2.027	5.98	1.760	5.41	2.203

Table 6.38: Mean score and Standard deviation (STD) in each dimension of experimentation for the three group of students

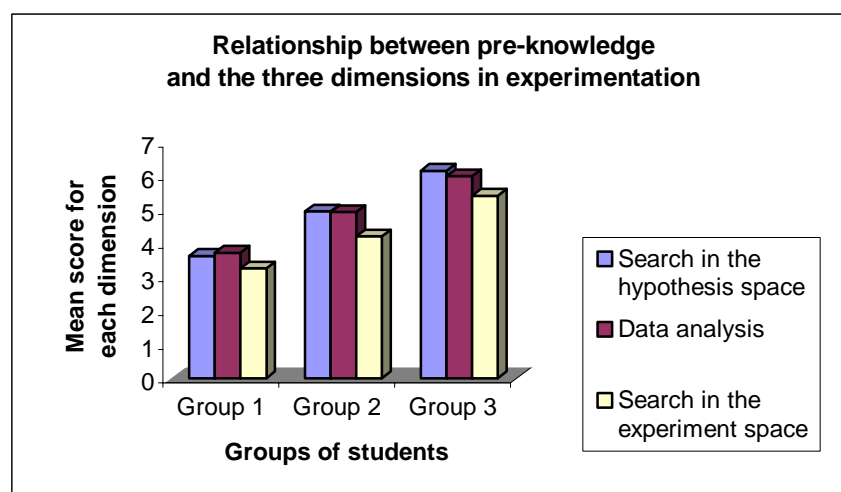


Figure 6.13: Relationship between the students' pre-knowledge and the three dimensions of experimentation for the three groups of students based on total scores in the knowledge test

Figure 6.13 shows that the students in group 1, who had low content knowledge (total scores were 1 to 5 points), solved the tasks in all the three dimensions in experimentation worse than those in group 2 - medium content knowledge (6 to 10 points). The students in group 3 - high content knowledge (11 to 15 points) - solved all the three dimensions better than two other classes.

	Pre-knowledge * Search in the hypotheses space	Pre-knowledge * Data analysis	Pre-knowledge * Search in the experiment space
Group 1 (n = 170)	0.238**	0.227**	0.160*
Group 2 (n = 589)	0.260**	0.215**	0.196**
Group 3 (n = 247)	0.214**	0.163*	0.202**

Table 6.39: Correlations (Spearman) between the students' pre-knowledge and the three dimensions of experimentation for the three classes of students based on total scores in the knowledge test

The correlation coefficients between the students' pre-knowledge and the three dimensions in group 1 ranged from 0.160 to 0.238. In group 2, they ranged from 0.196 to 0.260 and they were from 0.163 to 0.214 in group 3. Among them, in group 1 and group 2, the correlation between pre-knowledge and "search in the hypothesis space" was the lowest, in contrast, in group 3 the correlation between pre-knowledge and "data analysis" was lower than in two other combinations.

Discussion

The correlation coefficients between the students' pre-knowledge and the three dimensions ranged from 0.16 to 0.26 in all three classes of students. In group 1 and in group 2 where the students have low or medium scores in the knowledge test, the relationship between pre-knowledge and "search in the hypothesis space" and "data analysis" was closer than the relationship between pre-knowledge and "search in the experiment space". However, when students have high scores in content knowledge, the correlation between pre-knowledge and "data analysis" was lower than in the two other combinations.

However, figure 6.13 shows that if students have good pre-knowledge, they can also achieve high scores in all the three dimensions of experimentation and vice versa.

2.5.2.3. Relationship between the students' pre-knowledge and the three dimensions in experimentation in the three classes of students based on the three-latent class model

Based on the three latent classes analysis model, we assigned students to three classes of competency in experimentation and calculated the correlation coefficients between the students' pre-knowledge and the three dimensions in experimentation for each class we then had the following table:

	Pre-knowledge * Search in the hypotheses space	Pre-knowledge * Data analysis	Pre-knowledge * Search in the experiment space
Class 1 (n = 245)	0.222**	0.106	0.046
Class 2 (n = 163)	0.049	0.107	0.279**
Class 3 (n = 323)	0.236**	0.109	0.315**

Table 6.40: Correlations (Spearman) between the students' pre-knowledge and the three dimensions of experimentation for the three classes of students based on the three-class model

	(K*H) * (K*D)	(K*H) * (K*E)	(K*D) * (K*E)
Class 1	0.1892	0.0480	0.5066
Class 2	0.6016	0.0336	0.1090
Class 3	0.0972	0.2791	0.0061

Table 6.41: Difference between two independent correlation coefficients (Correlation coefficients between the students' pre-knowledge and the three dimensions in experimentation)

Table 6.40 shows that in class 1, the correlation coefficients between the students' pre-knowledge and the three dimensions ranged from 0.046 to 0.222. Among them, the correlation between pre-knowledge and the "search in the hypothesis space" was much higher than in the two other combinations.

Table 6.41 also shows that this correlation coefficient was significantly different with the correlation between pre-knowledge and the "search in the experiment space" ($p = 0.04$).

In class 2 the correlation coefficient between pre-knowledge and the three dimensions ranged from 0.049 to 0.279, where the correlation between pre-knowledge and "search in the hypothesis space" was lower than in the two others. In contrast, the correlation between pre-knowledge and the "search in the experiment space" was the highest and most significant.

Besides, in class 3, the correlation coefficient between pre-knowledge and the three dimensions ranged from 0.109 to 0.315. Among them, as well as in class 2, the correlation between pre-knowledge and the "search in the experiment space" was higher

than in the two other combinations. It was specially much higher than in the combination of pre-knowledge and the “data analysis” ($p = 0.006$). However, the correlation coefficient between pre-knowledge and the “search in the hypothesis space” was also significant.

Discussion

We expected higher correlation coefficients between the students’ pre-knowledge and those two dimensions in experimentation that depend on knowledge about the science subject matter – i.e. “search in the hypothesis space” and “data analysis” – than the correlation coefficients between the student’s pre-knowledge and the dimension that depends on the students’ methodological knowledge – i.e. “search in the experiment space”. The findings did not support these hypotheses. Though the correlation coefficient between pre-knowledge and the “search in the hypothesis space” was the highest in both of cases (in grade six and in both grades five and six combined) the correlation coefficient between pre-knowledge and the “search in the experiment space” was the lowest. However, the correlation coefficients in all three combinations were not significantly different. That means, if students have good content knowledge, they can also achieve high levels of competency in all three dimensions in experimentation.

Furthermore, based on the sum scores of the students in the knowledge test, we assigned students into three subclasses. The students in group 1 who have low total scores in the knowledge test solved the tasks in the three dimensions worse than in those in group 2 and in group 3 and those in group 3 who have high scores in pre-knowledge (cf. figure 6.2) did the tasks in experimentation better than in the two other classes. This indicated that the students who have good content knowledge are more likely to achieve high scores in all the three dimensions in experimentation than the students who have bad content knowledge.

However, if we base thing on the three latent classes analysis model and assign students to three classes, the correlation coefficients between pre-knowledge and the three dimensions were not similar in different classes of students.

In class 1, the correlation coefficient between pre-knowledge and the “search in the hypothesis space” was higher than in the two other combinations. And it was also significant. The correlations in the two other combinations were very low and not significant. That implicated that for the students who gained the lowest level in experimentation, pre-knowledge importantly influenced experimentation. If they have

good content knowledge, they could remember and use it in doing the tasks in “search in the hypothesis space”.

In contrast, in class 2, the correlation coefficient between pre-knowledge and “search in the experiment space” was higher than in the two other combinations and the correlation between pre-knowledge and the “search in the hypothesis space” was the lowest. This means, for the students who attained the medium level in experimentation the pre-knowledge was very important for the tasks in the “search in the experiment space”; if they have good content knowledge, they can also do well in experimentation.

On the other hand, in class 3, the correlation coefficients between pre-knowledge and “search in the hypothesis space” and between pre-knowledge and the “search in the experiment space” were quite high and significant, in which the second correlation was higher than the first. And the correlation between pre-knowledge and the “data analysis” was low and insignificant. These findings indicated that for students who attained a high level in experimentation, pre-knowledge affected the “search in the hypothesis space” and the “search in the experiment space” but not in the “data analysis”. Thus, if students have good pre-knowledge, they will attain a high level in the “search in the hypothesis space” and in the “search in the experiment space”.

Thus, the finding that if students have good content knowledge, they can also gain high levels in the “search in the hypothesis space” and the “data analysis” rather than in the “search in the experiment space” did not support our hypothesis.

2.5.2.4. Correlations between the students’ knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment (KAF) and the three dimensions in experimentation

	KAF * Search in the hypotheses space	KAF * Data analysis	KAF * Search in the experiment space
Both grades five and six combined (n = 1006)	0.241**	0.225**	0.238**
Grade six (n = 753)	0.255**	0.214**	0.253**

Table 6.42: Correlations (Spearman) between the students’ knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation

The correlations between the students’ knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation ranged from 0.225 to 0.241. Among them, the correlation between the dimension “data analysis” and knowledge about which factors are relevant and

irrelevant for the biological phenomenon examined are the lower than in the two other combinations, but it was not considerable. In grade six, the correlation coefficients the students' knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation ranged from 0.214 to 0.255.

	Search in the hypotheses space	Data analysis	Search in the experiment space
Irrelevant factors	0.237**	0.238**	0.202**
Relevant factors	0.093**	0.075**	0.124**

Table 6.43: Correlations (Spearman) between the students' knowledge about relevant or irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation (n = 1006)

The correlations between the students' knowledge about irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation ranged from 0.202 to 0.238, among them it was slightly lower in combination between irrelevant factors with the "search in the experiment space".

The correlations between the students' knowledge about relevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation were very low; all correlation coefficients were lower than 0.2.

Discussion

The correlations between the students' knowledge about relevant and irrelevant factors for the biological phenomena examined in the experiment (KAF) and the three dimensions in experimentation in grade six as well as in both grades five and 6 combined were similar and higher than 0.2. In pre-test 3, with the smaller number of students in the test, the correlation coefficients were higher than in the main study (they ranged from 0.295 to 0.473, table 5.23).

Similarly, the correlations between the students' knowledge about irrelevant factors for the biological phenomena examined in the experiment and the three dimensions in experimentation were also higher than 0.2. However, the correlation coefficients between relevant factors and the three dimensions were not considerable (cf. table 6.43). This showed that if students have knowledge about either both irrelevant and relevant factors or irrelevant factors, they are more likely to achieve higher scores in the competence test than students who do not have this kind of knowledge or who only have the knowledge about relevant factors for the biological phenomena examined in the experiment.

2.5.2.5. Correlations between the three dimensions of experimentation

	Search in the hypothesis space * Data analysis (H*D)	Data analysis * Search in the experiment space (D*E)	Search in the hypothesis space * Search in the experiment space (H*E)
Both grades 5 and 6 combined (n = 1006)	0.785**	0.615**	0.607**
Grade six (n = 753)	0.782**	0.660**	0.644**

Table 6.44: Correlations (Spearman) between the three dimensions in experimentation

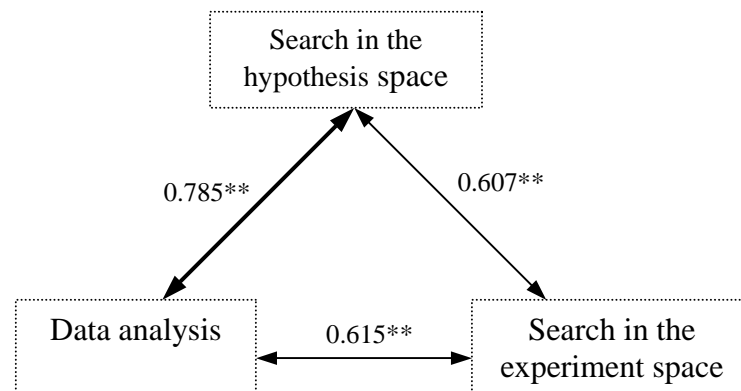


Figure 6.14: Correlations (Spearman) between the three dimensions in experimentation (n = 1006)

	Z-score	1-tail p	2-tail p
(H*D) * (D*E)	7.644	0	0
(H*D) * (H*E)	7.93	0	0
(D*E) * (H*E)	0.286	0.3874	0.7749

Table 6.45: Difference between two independent correlation coefficients
(Correlation coefficients between the three dimensions in experimentation)

Table 6.44 shows that the correlation coefficients between the three dimensions ranged from 0.607 to 0.785. Among them, the correlation coefficient for the relationship between “search in the hypothesis space” and “data analysis” was the highest. On the other hand, the correlation coefficients between “search in the hypothesis space” and “search in the experiment space” and between “data analysis” and “search in the experiment space” were also high.

Similarly, in grade six, the correlations between the three dimensions of experimentation were also high in all the three above relationships, the correlation coefficients ranged from 0.644 to 0.782.

Discussion

The medium and high correlation coefficients for the relationships between the three dimensions are consistent with the findings in pre-test 1 and pre-test 3 where the same correlations were also analysed (cf. tables 3.46 and 5.24). Especially with the large number of students in the test, the findings in the main study indicated that in the three relationships between the three dimensions of experimentation, the correlation between “search in the hypothesis space” and “data analysis” was the highest. On the other hand, comparing the difference between the two independent correlation coefficients, table 6.41 shows that the correlation between “search in the hypothesis space” and “data analysis” was significantly different to the two other correlation coefficients ($p = 0$). This finding supported our hypothesis. We expected high correlation between the three dimensions in experimentation. Especially we assumed that there are higher correlations between the dimensions “search in the hypothesis space” and “data analysis” than between the dimensions “search in the hypothesis space” and “search in the experiment space”. We also assumed that there are higher correlations between the dimensions “search in the hypothesis space” and “data analysis” than between the dimensions “data analysis” and “search in the experiment space”. These hypotheses were based on the assumption that the dimensions “search in the hypothesis space” and “data analysis” are driven by the students’ pre-knowledge about the science contents of the experiment while the dimension “search in the experiment space” should prove more dependent on the students’ methodological knowledge.

On the other hand, the majority of the correlations, however, suggested that the correlations between the three dimensions are high. This is not an unusual finding, however, as high achieving students have high content knowledge as well as high methodological knowledge and do well in all the three dimensions. The opposite is true for low achieving students.

2.5.2.6. The correlations between the three dimensions in experimentation for the different classes of students based on total scores in the knowledge test

We assigned students to three classes with the different total scores in the knowledge test and calculated the correlation coefficients between the three dimensions in experimentation for each group. We then had the following table:

	Search in the hypothesis space * Data analysis	Data analysis * Search in the experiment space	Search in the hypothesis space * Search in the experiment space
Group 1 (n = 170)	0.614**	0.384**	0.373**
Group 2 (n = 589)	0.780**	0.568**	0.561**
Group 3 (n = 247)	0.685**	0.620**	0.591**

Table 6.46: Correlations (Spearman) between the three dimensions in experimentation for the three classes of students based on total scores in the knowledge test

	(H*D) * (D*E)	(H*D) * (H*E)	(D*E) * (H*E)
Group 1	0.0045	0.0031	0.9066
Group 2	0	0	0.8604
Group 3	0.2100	0.0785	0.6129

Table 6.47: Difference between two independent correlation coefficients (2-tail p)
(Correlation coefficients between the three dimensions in experimentation)

Table 6.46 shows that the correlation coefficient between the three dimensions in experimentation in group1 ranged from 0.373 to 0.614, it ranged from 0.561 to 0.780 in group 2 and from 0.591 to 0.685 in group 3. In all three classes, the correlation between “search in the hypothesis space” and “data analysis” was the highest.

Discussion

In group 1 and in group 2, where students have bad or medium scores in content knowledge, the correlation between “search in the hypothesis space” and “data analysis” was much higher than in the two other combinations and it was significantly different with two others (cf. table 6.47, $p = 0$), although, all three combinations were also high and significant. That means, in group 1 and group 2, where students have bad or medium scores in pre-knowledge, if students do well in “search in the hypothesis space” they can do well in “data analysis” rather than in “search in the experiment space” and vice versa.

However, in group 3, where students have good content knowledge, all three correlation coefficients between the three dimensions were quite high and they were not significantly different. This indicated that if students have good content knowledge, they can do well in all the three dimensions in experimentation.

2.5.2.7. Correlations between the three dimensions in experimentation in the different classes of students based on the three-latent-class model

We assigned students to three classes of competency in experimentation and calculated the correlation coefficients between the three dimensions in experimentation for each class. The following table shows this:

	Search in the hypothesis space * Data analysis	Data analysis * Search in the experiment space	Search in the hypothesis space * Search in the experiment space
Class 1 (n = 245)	0.336**	0.198*	0.140**
Class 3 (n = 163)	0.234**	0.325**	0.374**
Class 3 (n = 323)	0.475**	0.442**	0.421**

Table 6.48: Correlations (Spearman) between the three dimensions in experimentation for the three classes of students based on the three latent classes model

	(H*D) * (D*E)	(H*D) * (H*E)	(D*E) * (H*E)
Class 1	0.1013	0.0217	0.5112
Class 2	0.3768	0.1665	0.6174
Class 3	0.5970	0.3925	0.7440

Table 6.49: Difference between two independent correlation coefficients (2-tail p) (Correlation coefficients between the three dimensions in experimentation)

The correlation coefficients between the three dimensions in experimentation in different classes of students were not similar. Table 6.48 shows that in class 1, the correlation coefficients between the three dimensions ranged from 0.140 to 0.336. Among them, the correlation between “search in the hypothesis space” and “data analysis” was higher than in the two other combinations. This correlation was significantly different with the correlation between “search in the hypothesis space” and “search in the experiment space” (cf. table 6.49, $p = 0.02$).

In class 2, the correlation coefficients between the three dimensions ranged from 0.234 to 0.374. Though, the correlation between “search in the hypothesis space” and “search in the experiment space” was higher than in the two other combinations, all three correlations were not significantly different ($p > 0.16$).

In class 3, the correlation coefficients between the three dimensions ranged from 0.421 to 0.475, in which, as well as in class 1, the correlation between “search in the hypothesis space” and “data analysis” was higher than in the two other combinations, however, there was not much difference between these three correlation coefficients ($p > 0.39$).

Discussion

When we assigned students to three classes with the different levels of competency in experimentation, the correlation coefficients between the three dimensions in experimentation were not the same in the different classes.

In class 1, the correlation coefficient between “search in the hypothesis space” and “data analysis” was the highest and it was significantly different with the correlation between “search in the hypothesis space” and “search in the experiment space”. That means, if students solve the tasks in “search in the hypothesis space”, they can also solve the ones in “data analysis” and vice versa, but not in “search in the experiment space”. On the other hand, if students can do well in “data analysis”, most of them can also do well in “search in the experiment space” and vice versa.

In class 2, the correlation coefficient between “search in the hypothesis space” and “search in the experiment space” was the highest and in class 3 it was the highest in the combination between “search in the hypothesis space” and “data analysis”. However, in these two classes, three correlation coefficients between the three dimensions was quite high in all three combinations. It was especially high in class 3 and they were not significant different. In class 2 and class 3, if students do well in one dimension, they can also do well in the others.

GENERAL DISCUSSION

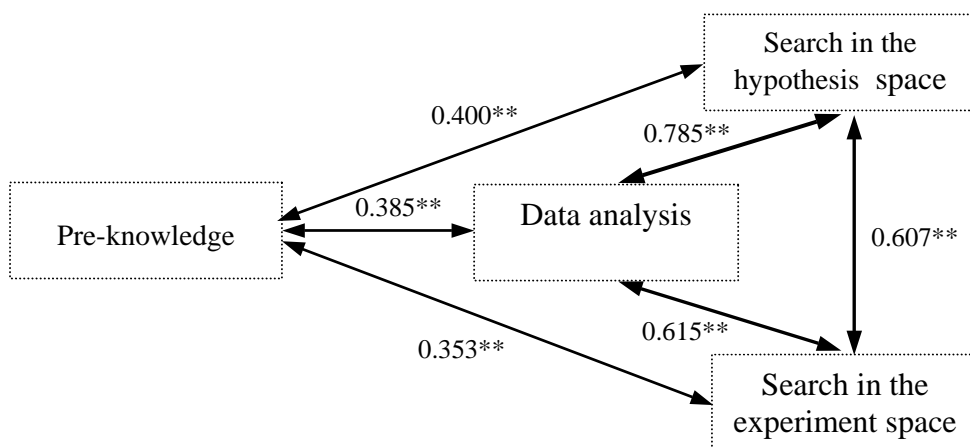


Figure 6.15: Correlations (Spearman) between the students’ pre-knowledge and the three dimensions in experimentation and within the three dimensions (n = 1006)

The correlation coefficients between the students’ pre-knowledge and the three dimensions were low in all three combinations. However, the correlation coefficients between the three dimensions of experimentation were medium and high. The

correlation between “search in the hypothesis space” and “data analysis” was very high. If students do well in one dimension, they can also do well in the two other dimensions. Especially, if they attain a high level in “search in the hypothesis space” they can also attain a high level in “data analysis”. The correlations between the three dimensions were much higher than the correlations between pre-knowledge and the three dimensions, that means, the competence (e.g. search in the experiment space” seems a stronger predictor for other competences than the biological content knowledge.

Similarly, if we assign students to three different classes based on the total scores in the knowledge test, the correlations between pre-knowledge and the three dimensions in experimentation were low in all three classes and not significantly different. Moreover, the correlation between “search in the hypothesis space” and “data analysis” was always higher than in the two other combinations, especially in group 1 and in group 2, this correlation coefficient was significantly different with two others (cf. table 6.47).

Correlations between the students’ pre-knowledge and the three dimensions and within the three dimensions in experimentation for three classes of students basing on sum scores in the knowledge test

Group 1:

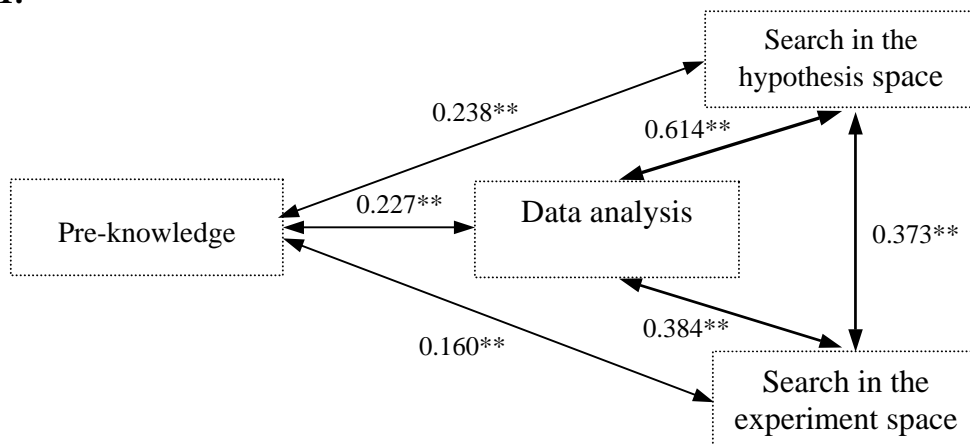


Figure 6.16: Correlations (Spearman) between the students’ pre-knowledge and the three dimensions in experimentation and within the three dimensions (Group 1, sum scores in the knowledge test were from 1 to 5; n = 170)

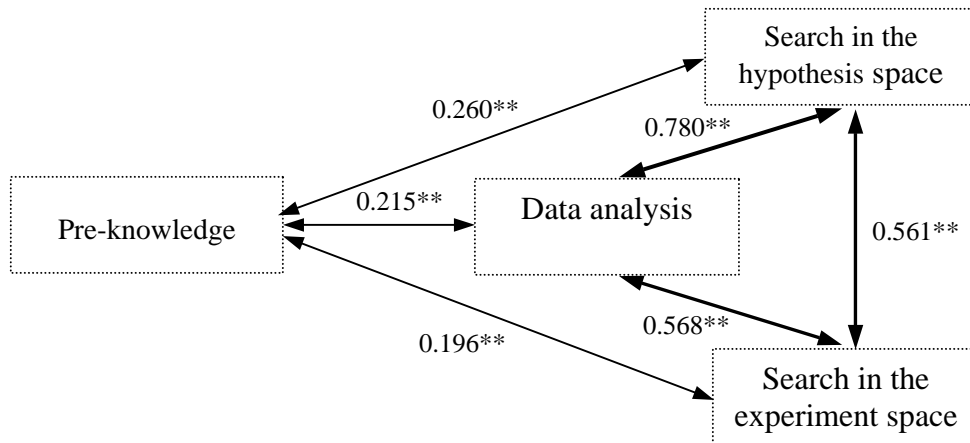
Group 2:

Figure 6.17: Correlations (Spearman) between the students' pre-knowledge and the three dimensions in experimentation and within the three dimensions (Group 2, sum scores in the knowledge test were from 6 to 10; $n = 589$)

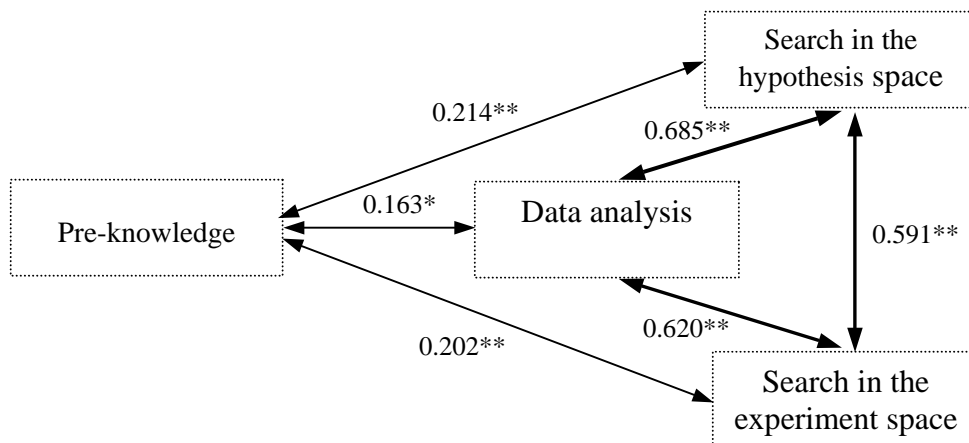
Group 3:

Figure 6.18: Correlations (Spearman) between the students' pre-knowledge and the three dimensions in experimentation and within the three dimensions (Group 3, sum scores in the knowledge test were from 11 to 15; $n = 247$)

Figures 6.16 to 6.18 show that the correlation coefficients between pre-knowledge and the three dimensions of experimentation in all three classes were similar. However, the correlations within the three dimensions were not similar between three classes, especially between group 1, group 2 – where students have either low or medium scores in the content knowledge - and group 3 – where students have high scores in pre-knowledge. The correlation coefficients between the three dimensions were medium and high in all three combinations and in the three groups of students. However, in

group 1 and in group 2, the correlation between “search in the hypothesis space” and “data analysis” was much higher than in the two other combinations and this correlation was significantly different from correlation between “search in the hypothesis space” and “search in the experiment space” and the correlation between “data analysis” and “search in the experiment space” (cf. table 6.47, $p = 0.003$ and $p = 0.004$ or $p = 0$). In group 3, the correlation between the three dimensions were medium and similar in all three combinations. This means if students have good knowledge they can do well in all the three dimensions of experimentation and vice versa. However, if students have low or medium content knowledge, the relationship between “search in the hypothesis space” and “data analysis” was closer than in the two other combinations. If students solve the tasks in “search in the hypothesis space”, they can solve ones in “data analysis” and vice versa.

In contrast, if students have good content knowledge and if they can do well in one dimension, they can also do well in the two other dimensions.

Correlation between the students’ pre-knowledge and the three dimensions and within the three dimensions in experimentation for three classes of students basing on the three latent classes model

In class 1, where students had the lowest level of competency in experimentation, the correlation between the students’ pre-knowledge and the three dimensions in experimentation and between the three dimensions was showed as the following figure:

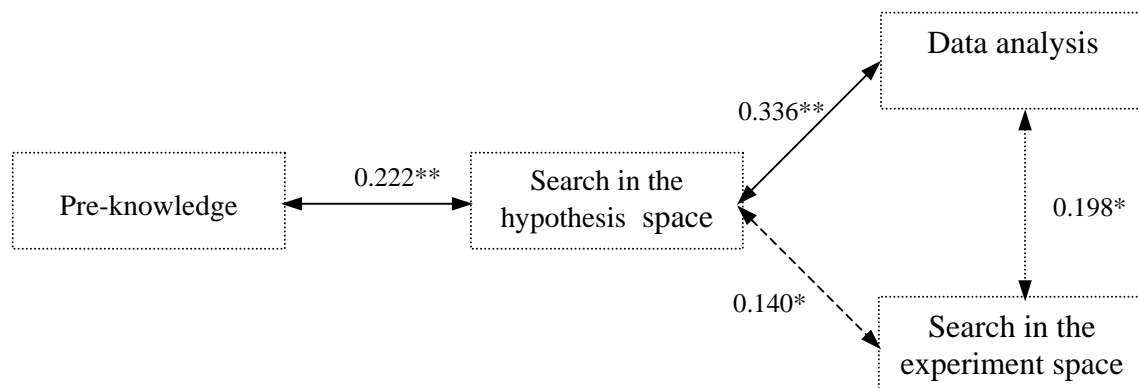


Figure 6.19: Correlations (Spearman) between the students’ pre-knowledge and the three dimensions in experimentation and within the three dimensions (Class 1, $n = 245$)

In class 1, the correlation coefficient between the students' pre-knowledge and the "search in the hypothesis space" was higher than in the two other combinations, but it was still low.

On the other hand, the correlation coefficient between the three dimensions was also low, especially the correlations between "search in the hypothesis space" and "search in the experiment space" and between "search in the experiment space" and "data analysis" were not considerable.

In class 2 (where students had the medium level of competency in experimentation)

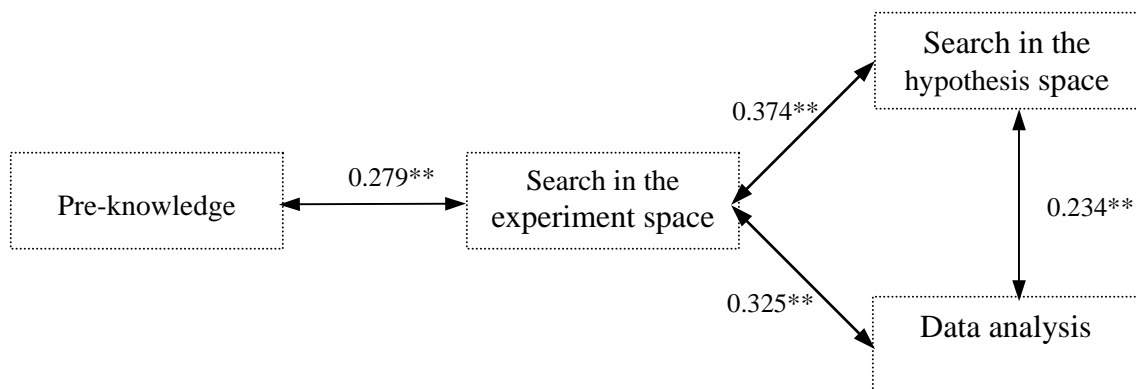


Figure 6.20: Correlations (Spearman) between the students' pre-knowledge and the three dimensions in experimentation and within the three dimensions (Class 2, $n = 163$)

It was not the same as in class 1. In class 2 the correlation coefficient between pre-knowledge and "search in the experiment space" was higher than that in the two other combinations, but as well as in class 1, the correlations between pre-knowledge and the three dimensions were low.

Besides, the correlation coefficient between the three dimensions was also low, it was higher in the combination between "search in the hypothesis space" and "search in the experiment space" than in the two other combinations, however, the correlation coefficients between the three dimensions were not significantly different.

In class 3 (where students had the highest level of competency in experimentation)

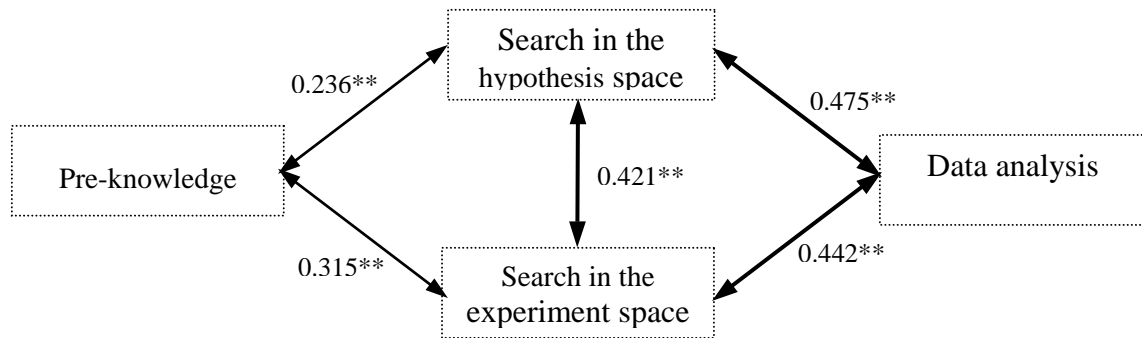


Figure 6.21: Correlations (Spearman) between the students’ pre-knowledge and the three dimensions in experimentation and within the three dimensions (Class 3, $n = 323$)

In class 3, the correlation coefficient between the students’ pre-knowledge and “search in the experiment space” was higher than in the two other combinations, especially much higher than the correlation between pre-knowledge and “data analysis”. However, all correlations were low.

However, the correlation coefficients between the three dimensions were medium in all three combinations. So, if students gain a high level of competency in one dimension, they can also gain a high level of competency in the two other dimensions.

Thus, in all three classes, the correlation between pre-knowledge and “data analysis” was always very low. This indicated that in all three classes the students’ good pre-knowledge did not influence the high scores in the dimension “data analysis”. The reason for this may be, because students have good content knowledge. When they evaluate evidence they only base this on their belief and their pre-knowledge but they do not include present evidence. In other words, students can evaluate evidence based on their subjective thinking but not on objective data.

3. Conclusion

This main study was successful for both the knowledge test and the competency test, especially the competency test with high reliabilities.

In the knowledge test, the reliability coefficient at the booklet level in grade five was only 0.60, but for grade six it was 0.65 and it was 0.63 for both grades five and 6 combined. Thus, the knowledge test at the booklet level either in grade six or in both grades five and six combined was reliable and could be used to assess the relationship between the students' pre-knowledge and their competences in experimentation.

In the competency test, the reliability coefficients at the level of the unit, the scales and the test booklet were high. For example, the Cronbach's alpha for the test booklet was 0.88 and Cronbach's alpha for the scale "search in the hypothesis space" was 0.78.

The correlation coefficients for the relationship between the students' pre-knowledge and the three dimensions of experimentation were equally low in all three combinations. However, the correlation coefficients between the three dimensions were high, especially the correlation between "search in the hypothesis space" and "data analysis". This supported our hypothesis about the closer relationship between "search in the hypothesis space" and "data analysis" than in the two other combinations.

Latent class analysis in the competency test indicated that when students are assigned to three classes, students in class 3 always outdo the two other classes and students in class 2 outdo the students in class 1 in all the three dimensions in experimentation.

Chapter 7: Conclusion

Test development was successful insofar as it was possible to develop reliable scales for the three dimensions “search in the hypothesis space”, “search in the experiment space” and “data analysis”. Since this is a new approach to measuring student achievement in fairly finely-grained sub-dimensions of experimentation, it might prove helpful for teachers and researchers who want to assess students’ competencies in experimentation with a paper-and-pencil test. The reliability coefficient of the competency test was above 0.8 when all items of the three scales are combined. This suggests that the paper-and-pencil items used in this study can yield reliable insights into student’s competences. About the validity of this test, of course, nothing can be said because student achievement in this test was not compared to the performance of the same students in real laboratory work, for example.

Basic competencies in experimentation lie within the reach of students in grade 5 and 6 (age 10-12) and can be tested with the items developed for this study. The item difficulty for items in the competency test ranged from 46% to 74% and was slightly lower for grade five than for grade six. However, the reliability coefficient of the competency test at the unit level for grade 5 was considerably lower than for grade 6. In particular, the reliability coefficient for the two scales “data analysis” and “planning experiment” at the booklet level in grade 5 was 0.63 and 0.55, whereas the reliability coefficient was higher than 0.7 for the scales “forming hypothesis” and “data analysis” for grade six and for grades five and six combined. This suggests that the test used in this study is more suitable for students in grade 6 rather than in grade 5.

The aims of this study consisted in investigating whether it is possible to test different levels of competencies in experimentation. Further, this study investigated the relationships between the students’ pre-knowledge and their competencies in experimentation as well as the relationships between the three dimensions in experimentation. Detailed analyses of the scores of the competency test and the knowledge test – mainly correlation statistics and multivariate methods – point to different directions and leave room for interpretation. It is not quite clear at this point if it is justified to assume an intermediate level of competency: It was possible to demonstrate that the test reliability does not suffer from assuming that an intermediate level exists. Further proof is necessary to illuminate the question of whether levels of competencies exist.

In this study, the question whether or not an intermediate level of competence exists was tackled by latent class analyses.

A three-latent-class model revealed that there were highly competent students who performed well in all three dimensions, whereas other students performed at an intermediate level in the three dimensions. A third class of students was found that performed at the lowest levels of all students in all dimensions. This may be considered as proof that levels of competencies can be measured in order to characterize the differences between different types of students.

Our analysis of the relationship between the students' competences and their biological pre-knowledge suffers from not being able to look at correlations at the unit level due to low reliabilities at that level. Also, perhaps more importantly, low reliabilities of the knowledge test at the unit level and at the test booklet level make it necessary to interpret the findings of this study with some caution. On the basis of our analyses at the test booklet level – i.e. at a level which renders domain-specific knowledge differences meaningless because the score is formed across different units – we found equally low correlations between the students' pre-knowledge and all three dimensions in experimentation. This is in contrast to the main hypotheses of this study, namely the hypothesis that the ability to plan experiments is informed by methodological knowledge, whereas the ability to form hypotheses and interpret data is informed by the students' domain-specific pre-knowledge.

There are three possible reasons for this. First, it is possible, the low reliability of the knowledge test. Second, there may be a general factor that underlies all three dimensions but which could not be identified through in the test instrument of the study, for example the students' intelligence, i.e. their ability to think logically and – on the basis of this – make inferences.

And third, it is possible that solving the items does require different kinds of knowledge than what I tested – i.e., knowledge about the method of experimentation or procedural knowledge about the biological processes which is investigated.

On the other hand, the correlations between the three dimensions of experimentation were medium and high in all three combinations. They are much higher than the correlation coefficients for the relationships between the students' pre-knowledge and the three dimensions. That is, the competence, for example, planning experiments

seems a stronger predictor for other competences in experimentation (e.g., forming hypotheses, analysing data) than knowledge about the biological content of the experiment, for example, knowledge about which factors are relevant for explaining a specific phenomenon under consideration.

Furthermore, the correlation between the dimensions “search in the hypothesis space” and “data analysis” was much higher than in the relationships between the dimensions “search in the hypothesis space” and “search in the experiment space” and between the dimensions “data analysis” and “search in the experiment space”. This proved our hypothesis that interaction between the dimension “search in the hypothesis space” and dimension “data analysis” was slighter than that in two other combinations.

Furthermore, the cross tables between two of three dimensions of experimentation also indicated that if students attain a high level in “search in the experiment space”, they can also attain a high level in “data analysis” and vice versa. Besides, if they have a high level in “search in the experiment space” they also do well in “search in the hypothesis space” and “data analysis”.

Part IV

Appendix

Appendix I: Test instrument

Design of Tests

Nine units corresponding to 9 biological contents:

Unit 1: Seed germination

Unit 2: Chicken eggs

Unit 3: Apple wine

Unit 4: Baking bread

Unit 5: The growth of bean plants

Unit 6: Potatoes

Unit 7: Heart beat

Unit 8: Plant growth (Plant nutrients)

Unit 9: Fish respiration

Version 1

Each unit had 6 questions corresponding to three dimensions in experimentation, each dimension had two questions

We had three booklets:

- Booklet 111 contained: Unit 1, unit 2, unit 3, unit 4
- Booklet 121 contained: Unit 1, unit 5, unit 6, unit 7
- Booklet 131 contained: Unit 1, unit 2, unit 8, unit 9

Version 2

Each unit had 3 questions corresponding to three dimensions in experimentation, each dimension had only one questions

We had four booklets:

- Booklet 211 contained: Unit 1, unit 2, unit 3, unit 4, unit 5
- Booklet 221 contained: Unit 1, unit 6, unit 7, unit 8, unit 9
- Booklet 231 contained: Unit 2, unit 3, unit 5, unit 7, unit 9
- Booklet 241 contained: Unit 1, unit 4, unit 5, unit 6, unit 8

	Pre-test 1	Pre-test 2	Pre-test 3	Main test
Test	Knowledge test Competency test	Knowledge test	Knowledge test Competency test	Knowledge test Competency test
Students	Version 1: n = 252 Version 2: n = 547	Version 1: n = 122	Version 1: n = 77	Version 1: n = 1006
Booklet	111 (n = 73) 121 (n = 82) 131 (n = 97) 211 (n = 131) 221 (n = 216) 231 (n = 94)	111	111	111
Answering format	Knowledge test: CMC Competency test: SMC	Knowledge test: CMC	Knowledge test: SMC Competency test: SMC	Knowledge test: SMC Competency test: SMC

1. Competency test

Version 1

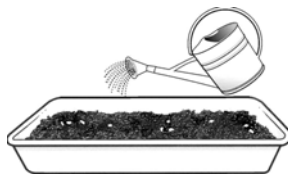
Unit 1: Samenkeimung



Aufgabe 1

Andreas macht ein Experiment zur Samenkeimung. Er verwendet dafür zwei Töpfe mit Erde. Er sät Bohnensamen in die Töpfe aus und sorgt dafür, dass beide Töpfe im Licht bei einer Temperatur von 22°C stehen. Topf 2 erhält kein Wasser (siehe Abbildungen).

Topf 1



**Erde / Wasser/
Licht / 22°C**

Topf 2



**Erde / kein Wasser/
Licht/ 22°C**

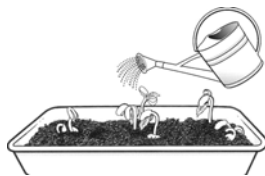
Warum macht Andreas dieses Experiment?

- A Weil er Samen dazu bringen will, schneller auszukeimen.
- B Weil er vermutet, dass Licht und Erde für die Samenkeimung notwendig sind.
- C Weil er vermutet, dass Wasser und Wärme für die Keimung notwendig sind.
- D Weil er vermutet, dass Wasser für die Samenkeimung notwendig ist.

Aufgabe 2

Nach einigen Tagen konnte Andreas folgendes feststellen: Die Samen im Topf 1 waren gekeimt. Im Topf 2 waren die Samen nicht gekeimt (siehe Abbildung).

Topf 1



**Erde / Wasser/
Licht/ 22°C**

Topf 2



**Erde /kein Wasser/
Licht/ 22°C**

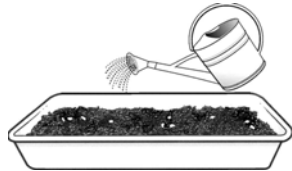
Wie lautet die beste Erklärung für das Ergebnis?

- A Das Experiment zeigte, dass Samen Wasser und Wärme brauchen, um zu keimen.
- B Das Experiment zeigte, dass Samen Wasser brauchen, um zu keimen.
- C Das Experiment klappte nicht, weil die Samen im Topf 2 nicht keimten.
- D Das Experiment zeigte, dass die Samen Licht und Erde brauchen, um zu keimen.

Aufgabe 3

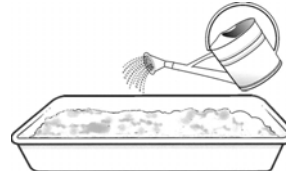
Maria macht auch ein Experiment zur Samenkeimung. Sie benutzt zwei Töpfe: In Topf 1 füllt sie Erde; für Topf 2 nimmt sie Watte aus Baumwolle anstatt Erde. Sie sät Bohnensamen in die beiden Töpfe, gießt die Samen und sorgt für eine Temperatur von 22°C (siehe Abbildungen).

Topf 1



**Erde / Wasser/
Licht / 22°C**

Topf 2



**Keine Erde / Wasser/
Licht/ 22°C**

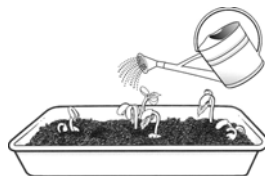
Warum macht Maria dieses Experiment?

- A Weil sie vermutet, dass Wasser und Wärme für die Samenkeimung notwendig sind.
- B Weil sie vermutet, dass Licht und Erde für die Samenkeimung notwendig sind.
- C Weil sie vermutet, dass Erde für die Samenkeimung notwendig ist.
- D Weil sie die Samen dazu bringen will, schneller auszukeimen.

Aufgabe 4

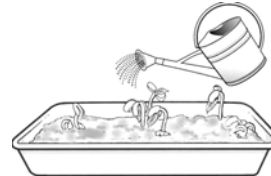
Maria konnte nach einigen Tagen folgendes feststellen: Die Samen waren in beiden Töpfen gekeimt (siehe Abbildungen).

Topf 1



**Erde / Wasser/
Licht/ 22°C**

Topf 2



**Keine Erde / Wasser/
Licht/ 22°C**

Wie lautet die beste Erklärung für das Ergebnis?

- A Das Experiment zeigte, dass Samen ohne Erde keimen können.
- B Das Experiment funktionierte, weil die Samen in beiden Töpfen keimten.
- C Das Experiment zeigte, dass Samen Wasser und Licht brauchen, um zu keimen.
- D Das Experiment zeigte, dass die Samen Wasser benötigen, um zu keimen.

Aufgabe 5

Anna vermutet, **dass Samen kein Licht brauchen, um zu keimen.**

Sie plant ein Experiment, um diese Vermutung zu überprüfen. Sie legt Bohnensamen in einen Topf mit Erde, hält die Erde feucht, stellt den Topf ins Dunkle und sorgt dafür, dass der Topf mit Luft versorgt wird.

Anna braucht aber noch einen zweiten Topf (Topf 2), um diesen mit Topf 1 zu vergleichen.

Topf 1



Wasser/ kein Licht / Luft

Welchen Topf sollte Anna wählen, damit sie ihre Vermutung überprüfen kann?

- | | | | | | |
|---|----------------------|---|-------|---|------------|
| A | Topf 2 : kein Wasser | + | Licht | + | keine Luft |
| B | Topf 2: Wasser | + | Licht | + | keine Luft |
| C | Topf 2: Wasser | + | Licht | + | Luft |
| D | Topf 2: kein Wasser | + | Licht | + | Luft |

Aufgabe 6

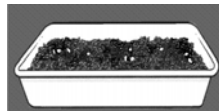
Tobias vermutet, **dass Samen Wärme brauchen, um zu keimen.** Er bereitet vier Töpfe mit Samen vor, um seine Vermutung zu überprüfen. Wie er die vier Töpfe vorbereitet, kannst Du hier sehen.

Topf 1



**Wasser
Licht / 22°C**

Topf 2



**kein Wasser
kein Licht/ 12°C**

Topf 3



**kein Wasser
Licht/ 12°C**

Topf 4



**Wasser
Licht / 12°C**

Welche beiden Töpfe sollte Tobias vergleichen, um sein Vermutung zu überprüfen?

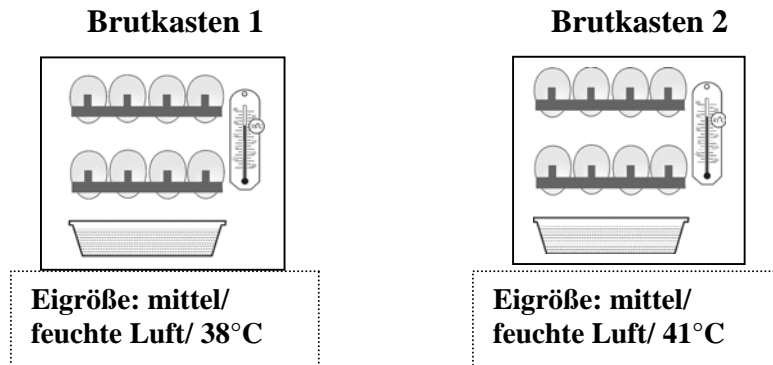
- A Topf 1 + Topf 4
- B Topf 1 + Topf 3
- C Topf 1 + Topf 2
- D Topf 3 + Topf 4



Unit 2: Kleine Küken

Aufgabe 1

Landwirt Bell züchtet Hühner. Er macht ein Experiment mit Eiern und zwei Brutkästen. In beide Brutkästen legt er mittelgroße Hühnereier. In Brutkasten 2 werden die Eier bei einer höheren Temperatur ausgebrütet als in Brutkasten 1. In beiden Brutkästen hat die Luft die gleiche Luftfeuchtigkeit. Dies ist in den Abbildungen zu sehen.

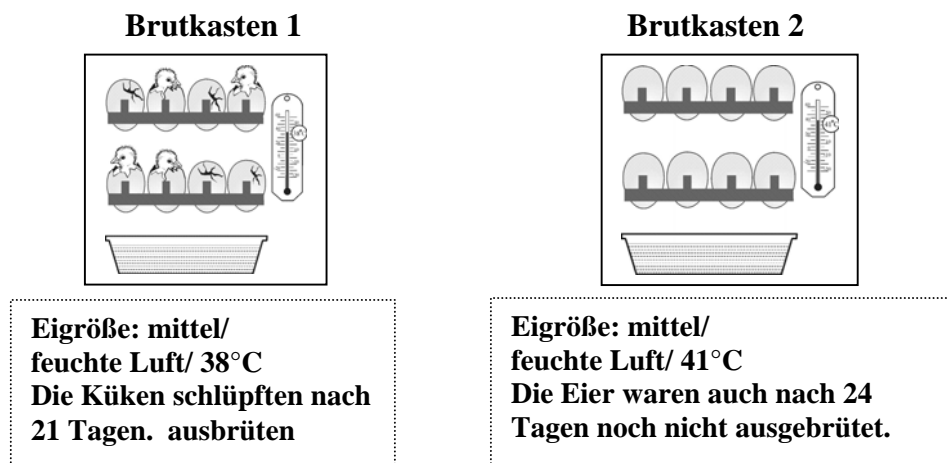


Warum macht Landwirt Bell dieses Experiment?

- A Weil er vermutet, dass sich mittelgroße Eier schneller ausbrüten lassen als große Eier.
- B Weil er vermutet, dass sich Eier schneller ausbrüten lassen, wenn im Brutkasten feuchte Luft ist.
- C Weil er vermutet, dass sich Eier bei höheren Temperaturen besonders schnell ausbrüten lassen.
- D Weil er vermutet, dass viele Bedingungen notwendig sind, um Eier schnell auszubrüten.

Aufgabe 2

Landwirt Bell erzielt das folgende Ergebnisse: Im Brutkasten 1 schlüpften die Küken nach 21 Tagen. Im Gegensatz dazu waren die Eier in Brutkasten 2 auch nach 24 Tagen immer noch nicht ausgebrütet.

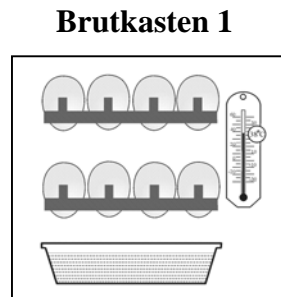


Wie lautet die beste Erklärung für das Ergebnis?

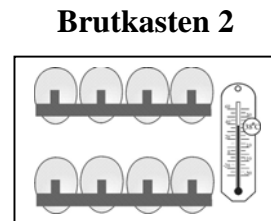
- A Das Experiment funktionierte nicht, weil sich die Eier im Brutkasten 2 nicht ausbrüten ließen.
- B Die Größe der Eier bestimmt, wie schnell die Eier ausgebrütet werden.
- C Wie schnell die Eier ausgebrütet werden, hängt von der Temperatur und der Luftfeuchtigkeit ab.
- D Wie schnell die Eier ausgebrütet werden, hängt von der Temperatur ab.

Aufgabe 3

Seine Nachbarin, Landwirtin Doll macht ein anderes Experiment. Sie legt mittelgroße Eier in zwei Brutkästen bei 38°C. In Brutkasten 1 befindet sich ein kleines Wassergefäß, um die Luft anzufeuchten. In Brutkasten 2 ist trockene Luft.



**Eigröße: mittel/
feuchte Luft/ 38°C**



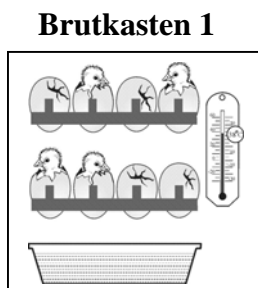
**Eigröße: mittel/
trockene Luft/ 38°C**

Warum macht Landwirtin Doll dieses Experiment?

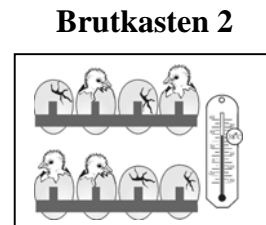
- A Weil sie er vermutet, dass sich Eier in feuchter Luft besonders schnell ausbrüten lassen.
- B Weil sie vermutet, dass die Temperatur wichtig für das Ausbrüten der Eier ist.
- C Weil sie mit den Bedingungen in den beiden Brutkästen bereits gute Erfahrungen gemacht hat.
- D Weil sie vermutet, dass sich Eier mittlerer Größe besonders schnell ausbrüten lassen.

Aufgabe 4

Frau Dolls Experiment hatte die folgenden Ergebnisse: Im Brutkasten 1 schlüpften gesunde Küken, im Brutkasten 2 jedoch verkrüppelte Küken .



**Eigröße: mittel/
feuchte Luft/ 38°C/
Alle geschlüpften Küken
waren gesund.**



**Eigröße: mittel/
trockene Luft/ 38°C/
Ein Teil der Küken kann
verkrüppelt zur Welt.**

Wie lautet die beste Erklärung für das Ergebnis?

- A Die Temperatur und die Luftfeuchtigkeit bestimmen, wie viele Küken gesund zur Welt kommen.
- B Die Zahl der gesunden Küken, hängt von der Luftfeuchtigkeit ab.
- C Die Größe der Eier bestimmt, wie viele Küken gesund zur Welt kommen.
- D Das Experiment funktionierte, weil sich die Eier in beiden Brutkästen ausbrüten ließen.

Aufgabe 5

Landwirt Bell möchte herausfinden, wie wichtig die Eigröße beim Ausbrüten der Eier ist. Er vermutet dass sich **kleinere Eier schneller ausbrüten lassen als große Eier**.

Er plant ein Experiment, um diese Vermutung zu überprüfen. Er muss zwei Brutkästen vorbereiten. Es stehen vier Brutkästen zur Auswahl:

Brutkasten A	Brutkasten B	Brutkasten C	Brutkasten D
Große Eier / 38°C / feuchte Luft	Kleine Eier / 41°C / trockene Luft	Große Eier / 38°C / trockene Luft	Kleine Eier / 38°C / feuchte Luft

Welche beiden Brutkästen sollte Landwirt Bell vergleichen, um seine Vermutung zu überprüfen?

- A Brutkasten B und Brutkasten C
- B Brutkasten A und Brutkasten B
- C Brutkasten A und Brutkasten D
- D Brutkasten C und Brutkasten D

Aufgabe 6

Herr Bell legt mittelgroße Eier in Brutkasten 1, in dem eine Temperatur von 38°C eingestellt ist und feuchte Luft ist. Nach 21 Tagen waren die Eier ausgebrütet.

Brutkasten 1:
mittelgroße Eier / feuchte Luft / 38°C

Er glaubt, dass **die Küken später schlüpfen, wenn im Brutkasten eine niedrigere Temperatur ist**. Er braucht aber noch einen weiteren Brutkasten, um diesen mit dem ersten Brutkasten zu vergleichen. Hierfür stehen vier Brutkästen zur Auswahl:

Brutkasten A	Brutkasten B	Brutkasten C	Brutkasten D
mittelgroße Eier trockene Luft 35°C	kleine Eier feuchte Luft 35°C	kleine Eier trockene Luft 35°C	mittelgroße Eier feuchte Luft 35°C

Welchen Brutkasten sollte Landwirt Bell wählen und mit Brutkasten 1 vergleichen?

- A Brutkasten A
- B Brutkasten B
- C Brutkasten C
- D Brutkasten D

Unit 3: Apfelwein



Aufgabe 1

Dennis interessiert sich für die Herstellung von Apfelwein. Er benutzt zwei gleich große Gefäße, gießt naturtrüben Apfelsaft hinein und fügt Zucker hinzu. Gefäß 1 wird bei einer Temperatur von 20°C aufbewahrt, Gefäß 2 bei einer Temperatur von 45°C. Beide Gefäße besitzen einen speziellen Verschluss, der verhindert, dass Luft von außen in das Gefäß kommt. Die Tabelle zeigt, wie Dennis vorgeht.

Zutaten	Gefäß 1	Gefäß 2
Apfelsaft	naturtrüb	naturtrüb
Temperatur (°C)	20°C	45°C
Zucker (Gramm)	450gr	450gr
Kann von außen Luft in das Gefäß?	nein	nein

Warum macht Dennis dieses Experiment?

- A Weil er herausfinden will, ob man Wein aus naturtrübem Apfelsaft herstellen kann.
- B Weil er herausfinden will, ob man Wein gut herstellen kann, wenn man Zucker zugibt.
- C Weil er sicher gehen will, dass in den Gefäßen guter Apfelwein entsteht.
- D Weil er herausfinden will, ob die Temperatur einen Einfluss auf die Weinherstellung hat.

Aufgabe 2

Nach einem Monat erzielt Dennis in seinem Experiment die folgenden Ergebnisse: die Weinherstellung im Gefäß 1 ist gut gelungen; die Weinherstellung im Gefäß 2 ist fehlgeschlagen.

Zutaten	Gefäß 1	Gefäß 2
Apfelsaft	naturtrüb	naturtrüb
Temperatur (°C)	20°C	45°C
Zucker (Gramm)	450gr	450gr
Kann von außen Luft in das Gefäß?	nein	nein
Ergebnis	Es entstand Wein.	Es entstand Essig.

Wie lautet die beste Erklärung für das Ergebnis?

- A Je niedriger die Temperatur, desto besser gelang die Weinherstellung.
- B Je höher die Zuckermenge, desto besser gelang die Weinherstellung.
- C Je mehr Luft von außen in das Gefäß hinein kann, desto besser gelang die Weinherstellung.
- D Das Experiment schlug fehl, weil die Weinherstellung im Gefäß 2 nicht gelang.

Aufgabe 3

Katrin macht ebenfalls ein Experiment zur Herstellung von Apfelwein. Sie benutzt dafür zwei gleich große Gefäße und gießt naturtrüben Apfelsaft hinein. Sie verwendet in Gefäß 2 mehr Zucker als in Gefäß 1 und bewahrt die beiden Gefäße bei einer Raumtemperatur von 20°C auf. Beide Gefäße besitzen einen speziellen Verschluss, der verhindert, dass Luft von außen in das Gefäß gelangt.

Zutaten	Gefäß 1	Gefäß 2
Apfelsaft	naturtrüb	naturtrüb
Temperatur (°C)	20°C	20°C
Zucker (Gramm)	450gr	900gr
Kann von außen Luft in das Gefäß?	nein	nein

Warum macht Katrin dieses Experiment?

- A Weil sie denkt, dass die Herstellung von Apfelwein bei Raumtemperatur gut gelingt.
- B Weil sie denkt, dass sich Wein umso schneller herstellen lässt, je mehr Zucker man zugibt.
- C Weil sie denkt, dass die Herstellung von Apfelwein besser gelingt, wenn keine Luft von außen in das Gefäß gelangt.
- D Weil sie denkt, dass aus Apfelsaft in jedem Fall Apfelwein wird.

Aufgabe 4

Katrin erzielt das folgende Ergebnis in ihrem Experiment: Der Wein in den Gefäßen 1 und 2 ist gut; der Wein in Gefäß 2 enthält mehr Alkohol als der Wein in Gefäß 1.

Zutaten	Gefäß 1	Gefäß 2
Apfelsaft	naturtrüb	naturtrüb
Temperatur (°C)	20°C	20°C
Zucker (Gramm)	450gr	900gr
Kann von außen Luft in das Gefäß?	nein	nein
Ergebnis	Wein mit wenig Alkohol	Wein mit viel Alkohol

Wie lautet die beste Erklärung für das Ergebnis?

- A Wein enthält viel Alkohol, wenn man naturtrüben Apfelsaft verwendet.
- B Das Experiment gelang in beiden Fällen, weil Alkohol entstand.
- C Je mehr Zucker man zugibt, desto höher ist der Alkoholgehalt des Weins.
- D Es hängt von der Temperatur ab, wie viel Alkohol im Apfelwein enthalten ist.

Aufgabe 5

Dennis Mutter verwendet normalerweise Rohrzucker, wenn sie Apfelwein aus Apfelsaft herstellt.

Aber sie denkt, dass sie genau so gut **Traubenzucker anstatt Rohrzucker für die Herstellung von Apfelwein verwenden kann**. Um ihre Vermutung zu überprüfen, plant sie ein Experiment. Sie gießt Apfelsaft in ein Gefäß, gibt 200 Gramm Rohrzucker hinzu und bewahrt das Gefäß bei einer Raumtemperatur von 20°C auf. Mit einem speziellen Verschluss sorgt sie dafür, dass kein Sauerstoff von außen in das Gefäß gelangen kann. In der Abbildung rechts kannst Du das Gefäß (Gefäß 1) sehen.

Gefäß 1



Rohrzucker: 200g
Temperatur: 20°C,
keine Luft kann
von außen in das
Gefäß

Katrins Mutter braucht aber noch ein zweites Gefäß, um dieses mit dem Gefäß in der Abbildung zu vergleichen. Hierfür stehen 4 Gefäße zur Auswahl:

Gefäß A	Gefäß B	Gefäß C	Gefäß D
Traubenzucker: 300g	Traubenzucker: 300g	Traubenzucker: 200g	Traubenzucker: 200g
20°C	25°C	25°C	20°C
Luft kann von außen in das Gefäß	Luft kann von außen in das Gefäß	keine Luft kann von außen in das Gefäß	keine Luft kann von außen in das Gefäß

Welches der folgenden vier Gefäße sollte sie wählen, damit sie ihre Vermutung überprüfen kann?

- A Gefäß A
- B Gefäß B
- C Gefäß C
- D Gefäß D

Aufgabe 6

Dennis Vater glaubt, dass **die Zugabe von Hefe notwendig ist, um Wein aus Apfelsaft herzustellen**. Um seine Vermutung zu überprüfen, plant er ein Experiment. In der Tabelle sind vier mögliche Experimente mit je zwei Gefäßen aufgeführt.

	Gefäß 1	Gefäß 2
Experiment 1	Rohrzucker/ Hefe / 20°C	Traubenzucker/ keine Hefe /20°C
Experiment 2	Traubenzucker/ Hefe / 20°C	Rohrzucker / keine Hefe /25°C
Experiment 3	Rohrzucker/ keine Hefe / 20°C	Rohrzucker/ Hefe/ 20°C
Experiment 4	Traubenzucker/ Hefe / 25°C	Traubenzucker/ keine Hefe / 20°C

Welches Experiment sollte er durchführen soll, um seine Vermutung zu überprüfen?

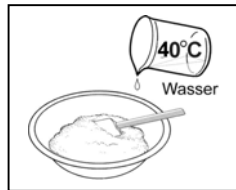
- A Experiment 1
- B Experiment 2
- C Experiment 3
- D Experiment 4

Unit 4: Brot backen

Aufgabe 1

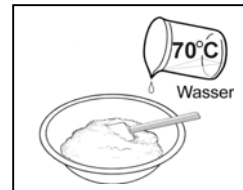
Anne macht ein Experiment zum Brotbacken. Sie vermischt weißes Mehl, Hefe, Zucker, Salz und Butter. Daraufhin teilt sie die Mischung auf 2 Rührschüsseln auf. In jede Schüssel kommt die gleiche Menge der Mischung. Dann gibt sie in Schüssel 1 warmes Wasser mit einer Temperatur von 40°C und rührt den Teig. In Schüssel 2 schüttet sie 70°C warmes Wasser, bevor sie rührt.

Schüssel 1



weißes Mehl/ Hefe/
Zucker/ Butter/ 40°C

Schüssel 2



weißes Mehl/ Hefe/
Zucker, Butter/ 70°C

Warum macht Anne dieses Experiment?

- A Weil sie möchte, dass das Brot in beiden Schüssel gut gelingt.
- B Weil sie vermutet, dass Zucker and Butter notwendig sind, um Brot zu backen.
- C Weil sie vermutet, dass die richtige Temperatur des Wassers wichtig ist.
- D Weil sie vermutet, dass Hefe für das Brot wichtig ist.

Aufgabe 2

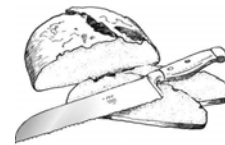
Anne erzielte das folgende Ergebnis in ihrem Experiment: Das Brot aus Schüssel 1 wurde weich und luftig, das Brot aus Schüssel 2 hart und fest.

Schüssel 1



weißes Mehl/ Hefe/
Zucker/ Butter/ 40°C
Das Brot ist weich und luftig.

Schüssel 2



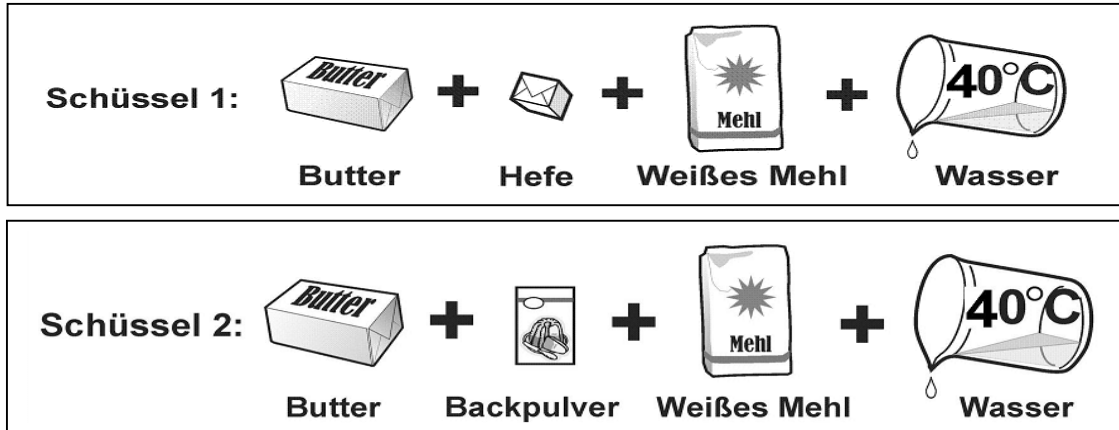
weißes Mehl/ Hefe/
Zucker/ Butter/ 70°C
Das Brot ist hart und fest.

Wie lautet die beste Erklärung für die Ergebnisse?

- A Heißes Wasser bewirkt, dass die Hefe abstirbt und das Brot hart und fest wird.
- B Brot wird weich und luftig, wenn man zum Backen Butter, Zucker und Salz nimmt.
- C Die besten Ergebnisse beim Brotbacken werden erzielt, wenn man Hefe nimmt.
- D Das Experiment funktionierte nicht, weil das Brot in Schüssel 2 hart und fest wurde.

Aufgabe 3

Tobias macht ein anderes Experiment zum Brotbacken. Er benutzt zwei unterschiedliche Teigmischungen. Den einen Brotteig stellt er aus Butter, Hefe und weißem Mehl her. Für den anderen verwendet er Butter, Backpulver und weißes Mehl. Zum Anrühren wird in beiden Fällen Wasser von 40°C verwendet.



Warum macht Tobias dieses Experiment?

- A Weil er vermutet, dass viele Zutaten notwendig sind, um gutes Brot zu backen.
- B Weil er überprüfen will, ob er besonders gutes Brot erhält, wenn man Butter nimmt.
- C Weil er vermutet, dass er besonders gutes Brot erhält, wenn man weißes Mehl nimmt.
- D Weil er überprüfen will, ob das Brotbacken besser mit Hefe oder mit Backpulver gelingt.

Aufgabe 4

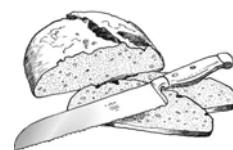
Tobias erzielte das folgende Ergebnis: Die beiden Brote aus den Schüsseln 1 und 2 wurden weich und luftig.

Schüssel 1



Butter/ Hefe/ weißes Mehl/
Das Brot ist weich und luftig.

Schüssel 2



Butter/ Backpulver/ weißes Mehl/
Das Brot ist weich und luftig.

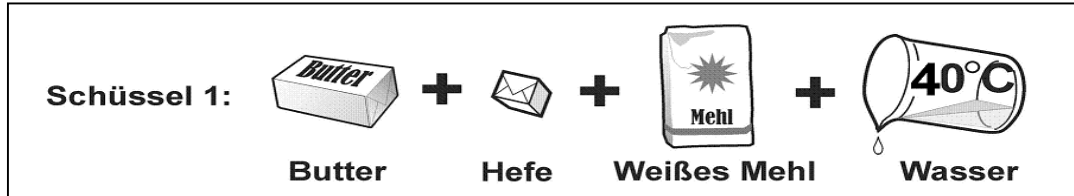
Wie lautet die beste Erklärung für die Ergebnisse?

- A Das Experiment funktionierte, weil das Brot in beiden Schüsseln weich und luftig wurde.
- B Die besten Ergebnisse beim Brotbacken werden erzielt, wenn man Hefe nimmt.
- C Die besten Ergebnisse beim Brotbacken werden erzielt, wenn man Backpulver nimmt.
- D Backpulver kann Hefe beim Brotbacken ersetzen.

Aufgabe 5

Tobias vermutet, dass **Brot mit weißem Mehl besser gelingt als mit Vollkornmehl.**

Er plant ein Experiment, um diese Vermutung zu überprüfen. Er nimmt eine große Schüssel und vermischt weißes Mehl, Hefe und Butter. Den Teig rührt er mit 40°C warmem Wasser an.



Er muss einen weiteren Teig ansetzen, um diesen mit dem ersten zu vergleichen.

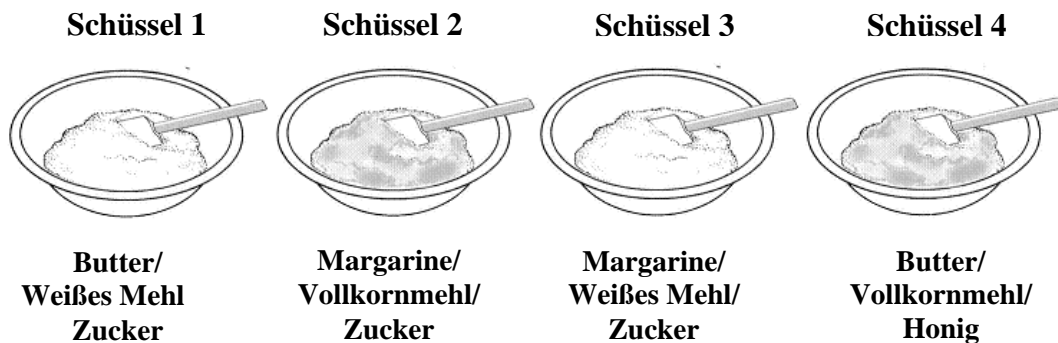
Welchen der folgenden Teige sollte er ansetzen, um seine Vermutung zu überprüfen?

- A Teig aus Vollkornmehl + Backpulver + Margarine + Wasser 60°C.
- B Teig aus Vollkornmehl + Hefe + Butter + Wasser 40°C.
- C Teig aus Vollkornmehl + Backpulver + Butter + Wasser 40°C.
- D Teig aus Vollkornmehl + Hefe + Margarine + Wasser 40°C.

Aufgabe 6

Tobias vermutet, dass er **Margarine anstatt Butter zum Brotbacken verwenden kann und dass das Brot trotzdem gut gelingen wird.**

Er plant ein Experiment, um seine Vermutung zu überprüfen. Hierfür stehen vier verschiedene Teige zur Auswahl:



Welche Teige sollte er anmischen, um seine Vermutung zu überprüfen?

- A Schüssel 2 und Schüssel 4
- B Schüssel 1 und Schüssel 2
- C Schüssel 1 und Schüssel 3
- D Schüssel 3 und Schüssel 4

Unit 5: Wie wachsen Bohnenpflanzen?



Aufgabe 1

Maria möchte mehr darüber wissen, wie Bohnenpflanzen wachsen. Deshalb macht sie ein Experiment. Sie nimmt zwei Töpfe mit jungen Bohnenpflanzen. Den einen Topf stellt sie ins Licht (Topf 1), den anderen ins Dunkle (Topf 2). Beide Töpfe stehen an der frischen Luft bei einer Temperatur von 22°C (siehe Abbildungen).

Topf 1



**Wasser/Licht/
Luft / 22°C**

Topf 2



**Wasser/ kein Licht/
Luft / 22°C**

Warum macht Maria dieses Experiment?

- A Weil sie vermutet, dass Bohnenpflanzen besser wachsen, wenn sie Wärme bekommen.
- B Weil sie vermutet, dass Bohnenpflanzen besser wachsen, wenn sie Licht bekommen.
- C Weil sie vermutet, dass Bohnenpflanzen besser wachsen, wenn sie Wasser bekommen.
- D Damit die Bohnenpflanzen in den beiden Töpfen schneller wachsen.

Aufgabe 2

Malte macht auch ein Experiment zum Bohnenwachstum. Dafür nimmt er wie Maria zwei Töpfe mit jungen Bohnenpflanzen und lässt diese unter den gleichen Bedingungen wachsen. Jedoch stellt er Topf 2 in die Kälte (siehe Abbildungen).

Topf 1



**Wasser/ Licht/
Luft / 22°C**

Topf 2



**Wasser / Licht /
Luft / 10°C**

Warum macht Malte dieses Experiment?

- A Damit die Bohnenpflanzen schneller wachsen.
- B Weil er vermutet, dass Bohnenpflanzen besser wachsen, wenn sie Wasser bekommen.
- C Weil er vermutet, dass Bohnenpflanzen besser wachsen, wenn sie Wärme bekommen.
- D Weil er vermutet, dass Bohnenpflanzen besser wachsen, wenn sie Licht bekommen.

Aufgabe 5

Maria und Malte vermuten, dass **Bohnenpflanzen Luft zum Wachsen benötigen**.

Um diese Vermutung zu überprüfen, planen sie ein Experiment. Hierfür nehmen sie junge Bohnenpflanzen (Topf 1), stellen den Topf an einen sonnigen Platz in der frischen Luft und halten die Erde feucht. Die Temperatur beträgt 22°C.

Topf 1



Licht/ 22°C/Luft

Maria und Malte brauchen aber noch einen zweiten Topf mit Bohnenpflanzen, damit sie ihre Vermutung überprüfen können.

Hier siehst du die Töpfe vor dem Beginn des Experiments.

Topf 1



**kein Licht/ 10°C/
keine Luft /**

Topf 2



**Licht/ 22°C/
keine Luft /**

Topf 3



**Licht/ 10°C/
keine Luft /**

Topf 4



**kein Licht / 22°C/
keine Luft /**

Welchen Topf sollten sie auswählen, um diesen mit Topf 1 zu vergleichen?

- A Topf 1
- B Topf 2
- C Topf 3
- D Topf 4

Aufgabe 6

Maria und Malte vermuten, dass **Bohnenpflanzen schneller in feuchter Erde als in trockener Erde wachsen**. Sie planen ein Experiment, um ihre Vermutung zu überprüfen. Sie pflanzen junge Bohnenpflanzen in vier Töpfe mit Erde.

Hier siehst du die Töpfe vor dem Beginn des Experiments. In den Bildern kannst du sehen, ob sie die Töpfe wässern, bei welcher Temperatur die Töpfe aufbewahrt werden und ob die Töpfe Luft erhalten oder nicht.

Topf 1



**feuchte Erde/
Luft/22°C**

Topf 2



**trockene Erde/
keine Luft/10°C**

Topf 3



**feuchte Erde /
Luft / 10°C**

Topf 4



**trockene Erde /
Luft /22°C**

Welche der vier Töpfe sollten Maria und Malte nach der Durchführung des Experiments vergleichen, um ihre Vermutung zu überprüfen?

- A Topf 1 und Topf 4
- B Topf 2 und Topf 3
- C Topf 3 und Topf 4
- D Topf 1 und Topf 2

Unit 6: Kartoffeln



Aufgabe 1

Sandy und Anna machen ein Experiment mit Kartoffeln. Sie nehmen zwei Töpfe, füllen diese mit feuchtem Sand, legen je zwei Kartoffeln hinein und stellen die Töpfe ins Licht. Die zwei Töpfe werden bei unterschiedlichen Temperaturen aufbewahrt. Auf den beiden Photos siehst du, wie Sandy und Anna vorgehen.

Topf 1



**feuchter Sand/
Licht / 6°C**

Topf 2



**feuchter Sand/
Licht / 20°C**

Warum machen Sandy und Anna dieses Experiment?

- A Weil sie Kartoffeln besonders schnell keimen lassen wollen.
- B Weil sie vermuten, dass Kartoffeln auf feuchtem Sand besonders schnell keimen.
- C Weil sie vermuten, dass Kartoffeln bei Wärme besonders schnell keimen.
- D Weil sie vermuten, dass Kartoffeln bei Licht und Feuchtigkeit besonders schnell keimen.

Aufgabe 2

Auch Christian macht ein Experiment mit Kartoffeln. Er nimmt zwei Töpfe, füllt diese mit feuchtem Sand, legt je zwei Kartoffeln hinein und bewahrt die Töpfe bei 20°C auf. Topf 1 stellt er ins Licht, Topf 2 ins Dunkle.

Topf 1



**feuchter Sand/
Licht / 20°C**

Topf 2



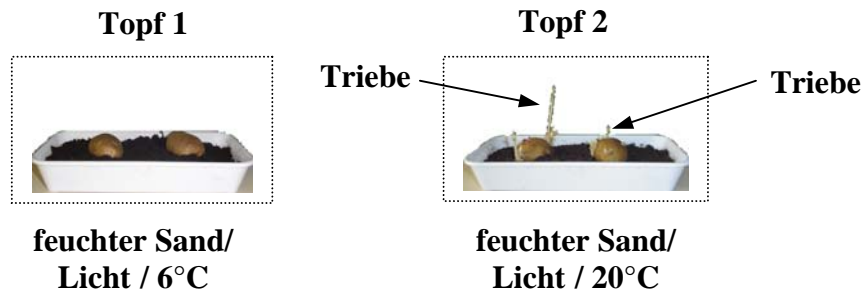
**feuchter Sand/ kein
Licht / 20°C**

Warum macht Christian dieses Experiment?

- A Weil er vermutet, dass die Kartoffeln im Dunkeln schneller keimen als im Licht.
- B Weil er vermutet, dass Kartoffeln Feuchtigkeit brauchen, um zu keimen.
- C Weil er vermutet, dass Kartoffeln im Licht und bei Feuchtigkeit besonders schnell keimen.
- D Damit die Kartoffeln in beiden Töpfen besonders schnell keimen.

Aufgabe 3

Nach 10 Tagen konnten Sandy und Anna sehen, dass nur die Kartoffeln in Topf 2 kleine Triebe hatten. Dies ist auf den Photos zu erkennen. Die Pfeile zeigen auf die Triebe.

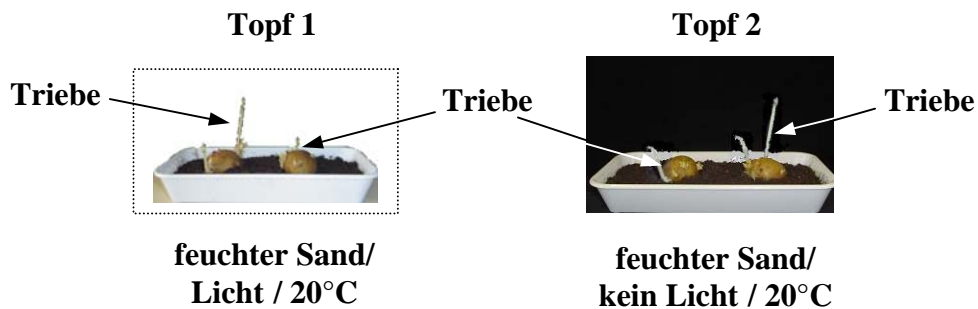


Wie lautet die beste Erklärung für das Ergebnis?

- A Kartoffeln können bei Wärme besser keimen als in Kälte.
- B Das Experiment funktionierte nicht, weil die Kartoffeln in Topf 1 keine Triebe haben.
- C Kartoffeln brauchen Licht und Feuchtigkeit, um zu keimen.
- D Kartoffeln brauchen Wärme und Feuchtigkeit, um zu keimen.

Aufgabe 4

Christian konnte nach 10 Tagen sehen, dass die Kartoffeln in Topf 1 und in Topf 2 kleine Triebe hatten. In beiden Töpfen waren die Triebe gleich lang.



Wie lautet die beste Erklärung für das Ergebnis?

- A Kartoffeln keimen besser auf feuchtem Sand als auf trockenem Sand.
- B Kartoffeln keimen besser bei Wärme als bei Kälte.
- C Das Experiment klappte nicht, weil die Kartoffeln in beiden Töpfen keimten.
- D Kartoffeln keimen gut bei Licht und bei Dunkelheit.

Aufgabe 5

Dennis vermutet, dass **Kartoffeln auch Luft benötigen, um zu keimen**.

Um diese Vermutung zu überprüfen, plant er ein Experiment. Er legt zwei Kartoffeln in den Topf 1 mit feuchtem Sand. Er stellt den Topf bei 20°C an die Luft.

Topf 1: feuchter Sand / 20°C / Luft

Dennis braucht aber noch einen zweiten Topf mit Kartoffeln, um seine Vermutung zu überprüfen.

Welchen der folgenden Töpfe soll er nehmen?

- A Topf A: trockener Sand / 10°C / wenig Luft
- B Topf B: feuchter Sand / 20°C / wenig Luft
- C Topf C: feuchter Sand / 10°C / wenig Luft
- D Topf D: trockener Sand / 20°C / wenig Luft

Aufgabe 6

Sandy, Anna, Christian und Dennis vermuten, dass **Kartoffeln auch Wasser benötigen, um zu keimen**. Jeder von ihnen plant ein eigenes Experiment. Hierfür nimmt jeder 2 Töpfe, befüllt diese mit Erde und legt je zwei Kartoffeln hinein. In der Tabelle siehst du, wie die vier Schüler ihre Experimente angelegt haben.

	Topf 1	Topf 2
Sandy	Licht/ Wasser/ 20°C	kein Licht/ kein Wasser/ 20°C
Anna	Licht/ kein Wasser/ 10°C	kein Licht/ Wasser/ 20°C
Christian	Licht/ Wasser/ 20°C	Licht/ kein Wasser/ 20°C
Dennis	Licht/ Wasser/ 10°C	Licht/ kein Wasser/ 20°C

Entscheide, welcher Schüler sein Experiment so geplant hat, dass man überprüfen kann, ob Kartoffeln Wasser zum Keimen benötigen.

- A Sandy
- B Anna
- C Christian
- D Dennis

Unit 7: Herzschlag



Aufgabe 1

Alex weiß, dass das Herz des Menschen nicht immer gleich schnell schlägt. Er möchte wissen, warum das so ist. Er führt deshalb eine Untersuchung an drei Personen durch: an einem vierjährigen Jungen, einem 10jährigen Jungen und einem 30jährigen Mann. Er untersucht, wie häufig ihr Herz pro Minute schlägt, wenn diese Personen ruhig liegen. Hier siehst Du die Einzelheiten seiner Untersuchung:

	Versuchsperson 1	Versuchsperson 2	Versuchsperson 3
Geschlecht	männlich	männlich	männlich
Alter	4 Jahre	10 Jahre	30 Jahre
Zustand	liegend	liegend	liegend

Warum macht Alex diese Untersuchung?

- A Weil er vermutet, dass es vom Geschlecht eines Menschen abhängt, wie häufig das Herz schlägt.
- B Weil er vermutet, dass es vom Alter der Person abhängt, wie häufig das Herz schlägt.
- C Weil er eine schnelleren Herzschlag anregen will.
- D Weil er vermutet, dass es von der körperlichen Aktivität abhängt, wie häufig das Herz schlägt.

Aufgabe 2

Kira führt eine andere Untersuchung durch. Sie untersucht, wie oft das Herz bei Mädchen im Alter von 10 Jahren pro Minute schlägt. Sie führt die Messung an zwei Mädchen durch. Den Herzschlag des einen Mädchens misst sie, nachdem sich dieses längere Zeit ausgeruht hat. Der Herzschlag des anderen Mädchens wird gemessen, nachdem es intensiv Sport getrieben hat und bevor es sich ausruhen konnte.

	Versuchsperson 1	Versuchsperson 2
Geschlecht	weiblich	weiblich
Alter	10 Jahre	10 Jahre
Was geschah kurz vor der Untersuchung?	Das Mädchen ruhte sich aus.	Das Mädchen trieb Sport.

Warum macht Kira diese Untersuchung?

- A Weil sie vermutet, dass Sport beeinflusst, wie oft das Herz schlägt.
- B Weil sie vermutet, dass das Geschlecht beeinflusst, wie oft das Herz schlägt.
- C Weil sie vermutet, dass das Alter beeinflusst, wie oft das Herz schlägt.
- D Weil sie einen schnelleren Herzschlag anregen will.

Aufgabe 3

Alex erzielte die folgenden Ergebnisse in seiner Untersuchung:

	Versuchsperson 1	Versuchsperson 2	Versuchsperson 3
Geschlecht	männlich	männlich	männlich
Alter	4 Jahre	10 Jahre	30 Jahre
Zustand	liegend	liegend	liegend
Ergebnisse Herzschläge pro Minute	102	90	69

Wie lautet die beste Erklärung für das Ergebnis?

- A Je jünger die Personen sind, desto schneller schlägt das Herz.
- B Alex Experiment ist fehlgeschlagen, weil die Untersuchungsergebnisse in den drei Gruppen unterschiedlich sind.
- C Bei körperlicher Aktivität schlägt das Herz schneller als wenn man sich ausruht.
- D Bei Männern schlägt das Herz schneller als bei Frauen.

Aufgabe 4

Kira erhielt die folgenden Ergebnisse in ihrer Untersuchung:

	Versuchsperson 1	Versuchsperson 2
Geschlecht	weiblich	weiblich
Alter	10 Jahre	10 Jahre
Was geschah kurz vor der Untersuchung?	Das Mädchen ruhte sich aus.	Das Mädchen trieb unmittelbar vor der Untersuchung Sport.
Ergebnisse Herzschläge pro Minute	75	102

Wie lautet die beste Erklärung für das Ergebnis?

- A Kiras Experiment klappte nicht, denn die Untersuchungsergebnisse sind in den beiden Gruppen unterschiedlich.
- B Bei Mädchen schlägt das Herz schneller als bei Jungen.
- C Bei jüngeren Menschen schlägt das Herz schneller als bei älteren Menschen.
- D Das Herz schlägt schneller, wenn man sich intensiv bewegt.

Aufgabe 5

Alex glaubt, dass **das Herz bei Männern und Frauen unterschiedlich schnell schlägt**. Er plant hierzu eine neue Untersuchung. Er untersucht, wie häufig das Herz von 20jährigen Männern schlägt, während sie ruhig liegen (Gruppe 1).

Gruppe 1:

20 Jahre alt / männlich / ruhig liegend

Alex braucht eine zweite Gruppe, um diese mit Gruppe 1 zu vergleichen. Welche Gruppe sollte er wählen, damit er seine Vermutung überprüfen kann?

- A Gruppe A: Alter 10 Jahre / weiblich / ruhig liegend
- B Gruppe B: Alter 20 Jahre / weiblich / nach intensiver Bewegung
- C Gruppe C: Alter 20 Jahre / weiblich / ruhig liegend
- D Gruppe D: Alter 10 Jahre / weiblich / nach intensiver Bewegung

Aufgabe 6

Kira vermutet, **dass das Herz umso schneller schlägt, je länger man Sport treibt**. Sie plant ein Experiment, um ihre Vermutung zu überprüfen. In dem Experiment sollen zwei Gruppen verglichen werden. Gruppe 1 besteht aus 20jährigen Männern, die 5 Minuten Sport treiben. Gruppe 2 besteht aus 40jährigen Frauen, die 10 Minuten Sport treiben. Sie plant also, die folgenden zwei Gruppen zu vergleichen:

Gruppe 1.

- Geschlecht: männlich
- Alter: 20 Jahre
- Die Versuchspersonen treiben 5 Minuten Sport.

Gruppe 2.

- Geschlecht: weiblich
- Alter: 40 Jahre
- Die Versuchspersonen treiben 10 Minuten Sport.

Der Lehrer behauptet, **dass dieses Experiment schlecht geplant sei, so dass Kira ihre Vermutung nicht überprüfen kann**.

Warum ist das Experiment schlecht geplant?

- A Weil die Personen in den beiden Gruppen unterschiedlich lange Sport treiben.
- B Weil die Männer in Gruppe 1 länger Sport treiben sollten, nicht die Frauen in Gruppe 2.
- C Weil die Untersuchung an jüngeren Personen durchgeführt werden sollte.
- D Weil Alter und Geschlecht der Personen in den beiden Gruppen unterschiedlich sind.



Unit 8: Pflanzenwachstum

Aufgabe 1

Björn interessiert sich dafür, wie Pflanzen wachsen. Er plant ein Experiment. Hierfür verwendet er zwei Töpfe mit Sand. In Topf 1 gibt er zusätzlich Pflanzendünger, in Topf 2 nicht. Dann pflanzt er je eine Jungpflanze vom Raps in die Töpfe. Die Pflanzen sind gleich groß. Den Sand in den Töpfen hält er feucht. Beide Töpfe stellte er ins Licht. Nach 14 Tagen misst Björn die Höhe der beiden Pflanzen.

	Topf 1	Topf 2
Art der Erde	Sand	Sand
Wasserzugabe	50 ml /Tag	50 ml /Tag
Düngerzugabe	ja	nein
Belichtung	den ganzen Tag über	den ganzen Tag über

Warum macht Björn dieses Experiment?

- A Weil er vermutet, dass Dünger das Pflanzenwachstum beeinflusst.
- B Weil er vermutet, dass Licht und Dünger das Pflanzenwachstum beeinflussen.
- C Weil er vermutet, dass die Art der Erde und Wasser das Pflanzenwachstum beeinflussen.
- D Weil er möchte, dass die Jungpflanzen in beiden Töpfen schnell wachsen.

Aufgabe 2

Nach 14 Tagen erhält Björn das folgende Ergebnis:

	Topf 1	Topf 2
Art der Erde	Sand	Sand
Wasserzugabe	50 ml / Tag	50 ml / Tag
Düngerzugabe	ja	nein
Belichtung	den ganzen Tag über	den ganzen Tag über
Ergebnis: Höhe der Pflanzen	30 cm	20 cm

Wie lautet die beste Erklärung für das Ergebnis?

- A Pflanzen wachsen besonders schnell, wenn sie an das Licht gestellt werden.
- B Das Experiment klappte nicht, denn die Pflanze in Topf 2 wuchs zu langsam.
- C Pflanzen wachsen besonders schnell nach Zugabe von Pflanzendünger.
- D Pflanzen wachsen besonders schnell, wenn der Sand, in dem sie wachsen, feucht gehalten wird.

Aufgabe 3

Ann-Christin führt auch ein Experiment mit Pflanzen durch. Sie verwendet 2 Töpfe. In den einen Topf füllt sie Sand (Topf 1), in den anderen Blumenerde (Topf 2). Dann gibt sie die gleiche Düngermenge in beide Töpfe, pflanzt gleich große junge Rapspflanzen hinein und stellt die Töpfe ins Licht. In der folgenden Zeit hält sie den Sand und die Erde in den beiden Töpfen feucht. Auch Ann-Christin misst nach 14 Tagen die Höhe der Pflanzen.

	Topf 1	Topf 2
Art der Erde	Sand	Blumenerde
Wasserzugabe	50 ml / Tag	50 ml / Tag
Düngerzugabe	ja	ja
Belichtung	den ganzen Tag über	den ganzen Tag über

Warum macht Ann-Christin dieses Experiment?

- A Weil sie vermutet, dass die Düngermenge und Wassermenge einen Einfluss auf das Pflanzenwachstum haben.
- B Damit die Rapspflanzen in den beiden Töpfen besonders gut wachsen.
- C Weil sie vermutet, dass die Art der Erde und die Düngermenge einen Einfluss auf das Pflanzenwachstum haben.
- D Weil sie vermutet, dass die Art der Erde einen Einfluss auf das Wachstum der Rapspflanzen hat.

Aufgabe 4

Nach 14 Tagen erzielt Ann-Christin das folgenden Ergebnisse:

	Topf 1	Topf 2
Art der Erde	Sand	Blumenerde
Wasserzugabe	50 ml / Tag	50 ml / Tag
Belichtung	den ganzen Tag über	den ganzen Tag über
Dünger	ja	ja
Ergebnis: das durchschnittliche Wachstum der Pflanzen	30 cm	43 cm

Wie lautet die beste Erklärung für das Ergebnis?

- A Pflanzen wachsen besonders gut, wenn sie den ganzen Tag Licht bekommen.
- B Pflanzen wachsen besonders gut, wenn sie gedüngt werden.
- C Pflanzen wachsen besonders gut, wenn sie in Blumenerde gepflanzt werden.
- D Das Experiment klappte, denn die Pflanze in Topf 2 wuchs zu schnell.

Aufgabe 5

Björn glaubt, dass **organischer Dünger für das Wachstum der Pflanzen besser ist als chemischer Dünger**. Er plant hierzu ein Experiment. Er pflanzt je eine Jungpflanze vom Raps in zwei Töpfe. Der eine Topf enthält Sand und organischen Dünger. Björn hält den Topf feucht und stellt ihn ins Licht.

Topf 1 erhält Sand + organischen Dünger + Wasser

Um seine Vermutung zu überprüfen, dass organischer Dünger für das Wachstum von Rapspflanzen besser sei als chemischer Dünger, benötigt Björn einen zweiten Topf zum Vergleich.

Welchen der folgenden Töpfe sollte Björn wählen?

- A Topf A: Blumenerde + organischer Dünger + Wasser
- B Topf B: Blumenerde (statt Sand) + chemischer Dünger + Wasser
- C Topf C: Sand + chemischer Dünger + kein Wasser
- D Topf D: Sand + chemischer Dünger + Wasser

Aufgabe 6

Ann-Christin glaubt, dass **Humuserde für das Pflanzenwachstum besser ist als Lehm**. Sie plant hierzu ein Experiment. Sie bereitet vier Töpfe mit jungen Rapspflanzen vor. Wie sie die Töpfe vorbereitet, kannst Du hier sehen:

Topf 1



**Humus/ kein Wasser/
Dünger**

Topf 2



**Lehm/ Wasser/
Dünger**

Topf 3



**Lehm/ Wasser/
kein Dünger**

Topf 4

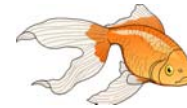


**Humus/ Wasser/
Dünger**

Welche beiden Töpfe sollte Ann-Christin vergleichen, um ihre Vermutung zu überprüfen?

- A Topf 3 + Topf 4
- B Topf 2 + Topf 4
- C Topf 1 + Topf 2
- D Topf 1 + Topf 3

Unit 9: Atmung der Fische



Aufgabe 1

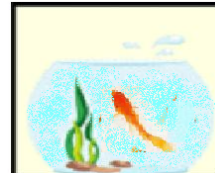
Daniel interessiert sich sehr für Fische. Eines Tages macht er ein Experiment zur Atmung von Fischen. Dabei verwendet er zwei gleiche Aquarien. Er setzt je einen Goldfisch in ein Aquarium mit einer Pflanze. Nach einiger Zeit beobachtet er, wie häufig die Fische pro Minute atmen. Dies erkennt er daran, wie schnell sich die Kiemendeckel der Fische bewegen. In dem einen Aquarium herrscht eine Wassertemperatur von 20°C, in dem anderen von 10°C.

Aquarium 1



20°C / eine Pflanze /
1 Fisch

Aquarium 2



10°C / eine Pflanze /
1 Fisch

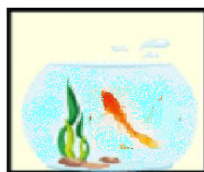
Warum macht Daniel dieses Experiment?

- A Weil er vermutet, dass es sowohl von der Anzahl der Fische im Aquarium als auch von der Wassertemperatur abhängt, wie häufig die Fische pro Minute atmen.
- B Weil er die Fische in den beiden Aquarien dazu bringen will, schneller zu atmen.
- C Weil er vermutet, dass es von der Wassertemperatur abhängt, wie häufig die Fische pro Minute atmen.
- D Weil er vermutet, dass es von der Anzahl der Pflanzen abhängt, wie häufig die Fische pro Minute atmen.

Aufgabe 2

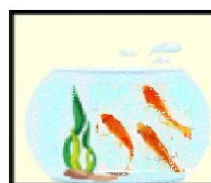
Jana macht auch ein Experiment zur Fischatmung. Sie benutzt zwei gleiche Aquarien. In das eine Aquarium setzt sie einen Goldfisch und in das andere drei Goldfische. Alle Fische sind gleich groß. Nach einiger Zeit beobachtet sie, wie häufig der Goldfische im Aquarium 1 pro Minute atmet. Anschließend misst sie, wie oft pro Minute einer der drei Fische im Aquarium 2 atmet. Auch Jana zählt die Bewegungen der Kiemendeckel pro Minute.

Aquarium 1



20°C / eine Pflanze /
1 Fisch

Aquarium 2



20°C / eine Pflanze /
3 Fische

Warum macht Jana dieses Experiment?

- A Weil sie die Fische dazu bringen möchte, schneller zu atmen.
- B Weil sie vermutet, dass es von der Zahl der Pflanzen und der Temperatur abhängt, wie häufig die Fische pro Minute atmen.
- C Weil sie vermutet, dass es von der Temperatur und die Anzahl der Fische abhängt, wie häufig die Fische pro Minute atmen.
- D Weil sie vermutet, dass es von der Anzahl der Fische im Aquarium abhängt, wie häufig die Fische pro Minute atmen.

Aufgabe 3

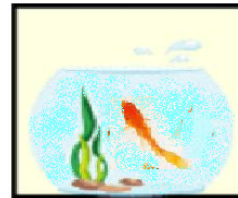
Daniel schreibt auf, wie häufig die beiden Fische in seinem Experiment pro Minute atmen. Hier ist sein Ergebnis: Im Aquarium 1 bewegen sich die Kiemendeckel 96 Mal pro Minute; im Aquarium 2 bewegen sich die Kiemendeckel 73 Mal pro Minute.

Aquarium 1



**20°C / 1 Fisch /
96 Kiemendeckel-
bewegungen pro Minute**

Aquarium 2



**10°C / 1 Fisch /
73 Kiemendeckel-
bewegungen pro Minute**

Wie lautet die beste Erklärung für Daniels Ergebnis?

- A Fische atmen besonders schnell, wenn mehrere Fische und Pflanzen im Aquarium sind.
- B Fische atmen besonders schnell, wenn das Wasser warm ist und wenn mehrere Fische im Aquarium sind.
- C Fische atmen schneller in einem Aquarium mit warmem Wasser als im kalten Wasser.
- D Das Experiment funktionierte nicht, weil die Fische in den beiden Aquarien verschieden schnell atmeten.

Aufgabe 4

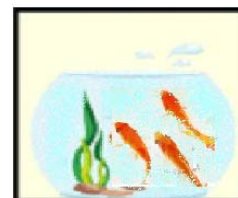
Janas Experiment ergab die folgenden Ergebnisse: Der große Fisch im Aquarium 1 atmete 96 Mal pro Minute; der eine große Fisch in Aquarium 2 atmete 108 Mal pro Minute.

Aquarium 1



**20°C / 1 Fisch /
96 Kiemendeckel-
bewegungen pro Minute**

Aquarium 1



**20°C / 3 Fische /
108 Kiemendeckel-
bewegungen pro Minute**

Wie lautet die beste Erklärung für Janas Ergebnis?

- A Fische atmen besonders schnell, wenn die Wassertemperatur hoch ist und mehrere Fische im Aquarium sind.
- B Fische atmen besonders schnell, wenn mehrere Fische im Aquarium sind.
- C Das Experiment funktionierte nicht, weil die Fische in den beiden Aquarien verschieden schnell atmeten.
- D Fische atmen besonders schnell, wenn man sie in ein Aquarium mit warmen Wasser setzt.

Aufgabe 5

Daniel glaubt, dass **die Größe eines Fisches einen Einfluss darauf hat, wie häufig er atmet**. Er plant ein Experiment, um seine Vermutung zu überprüfen. Er verwendet dafür zwei gleiche Aquarien und setzt je einen Goldfisch in die beiden Aquarien.

Die Tabelle zeigt 4 mögliche Experimente:

Experiment	Aquarium 1	Aquarium 2
Experiment 1	ein Fisch / 7cm 20°C	ein Fisch / 10cm 20°C
Experiment 2	ein Fisch / 10cm 10°C	zwei Fische / 7cm 20°C
Experiment 3	ein Fisch / 7cm 20°C	ein Fisch / 10cm 10°C
Experiment 4	zwei Fische / 10cm 20°C	ein Fisch / 7cm 20°C

Wähle ein Experiment, das Daniel verwenden kann, um seine Vermutung zu überprüfen.

- A Experiment 1
- B Experiment 2
- C Experiment 3
- D Experiment 4

Aufgabe 6

Jana glaubt, dass **Fische bei höheren Wassertemperaturen besonders schnell atmen**. Sie plant ein Experiment. Dazu verwendet sie 3 gleiche Aquarien, setzt zwei Fische in Aquarium 1, drei Fische in Aquarium 2 und einen Fisch in Aquarium 3. Alle Fische sind gleich groß. Die Temperatur im Aquarium 1 beträgt 20°C; in Aquarium 2 ist 10°C kaltes Wasser und Aquarium 3 enthält 30°C warmes Wasser. Sie gibt kein Futter in die drei Aquarien.

Aquarium 1

**20°C / 2 Fische
kein Futter**

Aquarium 2

**10°C / 3 Fische
kein Futter**

Aquarium 3

**30°C / 1 Fisch
kein Futter**

Ihr Lehrer sagt, dass sie das Experiment schlecht geplant habe und ihre Vermutung nicht überprüfen kann. Warum?

- A Weil sie die Größe der drei Aquarien verändern sollte.
- B Weil die Wassertemperatur in den drei Aquarien verschieden ist.
- C Weil sich in allen drei Aquarien kein Futter befindet.
- D Weil die Anzahl der Fische in den drei Aquarien verschieden ist.

Version 2

Unit 1: Samenkeimung



Aufgabe 1

Jan macht ein Experiment zur Samenkeimung. Er verwendet dafür zwei Töpfe mit Erde (Topf 1 und Topf 2) und einen Topf mit Watte aus Baumwolle anstatt Erde (Topf 3). Dann sät er Bohnensamen in die Töpfe und sorgt dafür, dass alle drei Töpfe eine Temperatur von 22°C erhalten. Er wässert Topf 1 und Topf 3, nicht aber Topf 2.

Topf 1



**Erde / Wasser /
Licht / 22°C**

Topf 2



**Erde / kein Wasser /
Licht / 22°C**

Topf 3



**Keine Erde / Wasser /
Licht / 22°C**

Warum macht Jan dieses Experiment?

- A Weil er vermutet, dass Wärme und Licht für die Samenkeimung notwendig sind.
- B Weil er vermutet, dass Erde und Wasser für die Samenkeimung notwendig sind.
- C Weil er alle Samen dazu bringen will auszukeimen.
- D Weil er vermutet, dass Wasser, Wärme, Licht und Erde für die Samenkeimung notwendig sind.

Aufgabe 2

Nach einigen Tagen konnte Jan folgendes feststellen: Die Samen im Topf 1 und 3 waren gekeimt. Aber in Topf 2 waren die Samen nicht gekeimt.

Topf 1



**Erde / Wasser /
Licht / 22°C**

Topf 2



**Erde / kein Wasser /
Licht / 22°C**

Topf 3



**Keine Erde / Wasser /
Licht / 22°C**

Wie lautet die beste Erklärung für dieses Ergebnis?

- A Das Experiment zeigt, dass Samen Wasser und Erde zur Keimung brauchen.
- B Das Experiment klappte nicht, weil die Samen im Topf 2 nicht keimten.
- C Das Experiment zeigt, dass Samen Wärme und Licht zur Keimung brauchen.
- D Das Experiment zeigt, dass Samen keine Erde, aber Wasser zum Keimen brauchen.

Aufgabe 3

Jan vermutet, dass **Samen besser keimen, wenn es warm ist.**

Er plant ein Experiment, um diese Vermutung zu überprüfen. Dies ist einer von Jans Töpfen (Topf 1). Er sät Bohnensamen in Erde, gießt die Samen und sorgt für eine Temperatur von 22 °C und Licht.

Topf 1



Erde / Licht/ 22°C

Jan braucht aber noch einen zweiten Topf mit Bohnensamen, damit er diesen mit Topf 1 vergleichen und seine Vermutung überprüfen kann. Es stehen vier Töpfe zur Auswahl:

Topf A



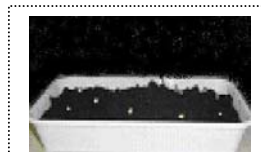
**Erde /
Licht/ 10°C**

Topf B



**Keine Erde /
Licht/ 10°C**

Topf C



**Erde /
kein Licht/ 10°C**

Topf D



**Keine Erde /
kein Licht/ 10°C**

Welchen Topf soll Jan nehmen, um seine Vermutung zu überprüfen?

- A Topf A
- B Topf B
- C Topf C
- D Topf D

Unit 2: Kleine Küken



Aufgabe 1

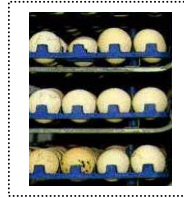
Landwirt Bell züchtet Hühner. Er macht ein Experiment mit Eiern und drei Brutkästen. In alle Brutkästen legt er mittelgroße Hühnereier. Die Temperatur in Brutkasten 1 und 3 beträgt 38°C. In Brutkasten 2 beträgt die Temperatur 41°C. Außerdem ist in Brutkasten 1 und 2 feuchte Luft. In Brutkasten 3 ist trockene Luft. Dies ist in den Abbildungen zu sehen.

Brutkasten 1



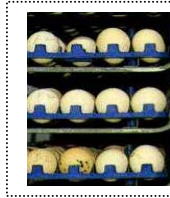
**Eigröße: mittel/
feuchte Luft/ 38°C**

Brutkasten 2



**Eigröße: mittel/
feuchte Luft/ 41°C**

Brutkasten 3



**Eigröße: mittel/
trockene Luft/ 38°C**

Warum macht Landwirt Bell dieses Experiment?

- A Weil er vermutet, dass viele Bedingungen wichtig sind, um Eier schnell auszubrüten.
- B Weil er vermutet, dass die Temperatur und Luftfeuchtigkeit für das schnelle Ausbrüten wichtig sind.
- C Weil er vermutet, dass die Temperatur, Luftfeuchtigkeit und Eigröße für das schnelle Ausbrüten wichtig sind.
- D Weil er vermutet, dass sich mittelgroße Eier schneller ausbrüten lassen als große Eier.

Aufgabe 2

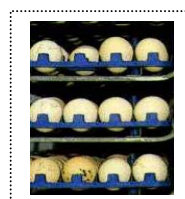
Landwirt Bell erhält die folgenden Ergebnisse: Im Brutkasten 1 schlüpften die Küken nach 21 Tagen, die Küken im Brutkasten 3 nach 19 Tagen. In Brutkasten 2 waren die Eier auch nach 24 Tagen immer noch nicht ausgebrütet.

Brutkasten 1



**Eigröße: mittel/
feuchte Luft/ 38°C
Die Küken schlüpften
nach 21 Tagen.**

Brutkasten 2



**Eigröße: mittel/
feuchte Luft/ 41°C
Die Eier waren auch nach 24
Tagen noch nicht ausgebrütet.**

Brutkasten 3



**Eigröße: mittel/
trockene Luft/ 38°C
Die Küken schlüpften
nach 19 Tagen.**

Wie lautet die beste Erklärung für die Ergebnisse?

- A Die Luftfeuchtigkeit und die Temperatur bestimmen die Länge der Brutzeit.
- B Das Experiment klappte nicht, weil sich die Eier im Brutkasten 2 nicht ausbrüten ließen.
- C Die Temperatur bestimmt die Länge der Brutzeit.
- D Die Temperatur und die Größe der Hühnereier bestimmen die Länge der Brutzeit.

Aufgabe 3

Landwirt Bell vermutet, dass sich **Hühnereier bei hohen Temperaturen nicht ausbrüten lassen**. Um diese Vermutung zu überprüfen, plant er ein weiteres Experiment: Er legt mittelgroße Eier in zwei Brutkästen mit einer Temperatur von 37° C (Brutkasten 1) und 39° C (Brutkasten 2). In Brutkasten 1 ist die Luft feucht, in dem anderen trocken (siehe Abbildung).

Brutkasten 1



**Eigröße: mittel
feuchte Luft/ 37° C**

Brutkasten 2



**Eigröße: groß
trockene Luft/ 39° C**

Ein benachbarter Landwirt sagt, dass Landwirt Bell **das Experiment schlecht geplant habe und deshalb seine Vermutung nicht überprüfen kann**. Warum sagt der benachbarte Landwirt dies?

- A Weil in den beiden Brutkästen unterschiedliche Temperaturen sind.
- B Weil in beiden Brutkästen die Luftfeuchtigkeit unterschiedlich ist.
- C Weil in beiden Brutkästen die Eigröße und die Luftfeuchtigkeit unterschiedlich sind.
- D Weil sich Hühnereier bei 39° C nicht ausbrüten lassen.



Unit 3: Apfelwein

Aufgabe 1

Dennis interessiert sich für die Herstellung von Apfelwein. Er benutzt drei gleich große Gefäße, gießt naturtrüben Apfelsaft hinein und fügt Zucker hinzu. In die Gefäße 1 und 2 füllt er weniger Zucker als in das Gefäß 3. Gefäß 1 und 3 werden bei einer Raumtemperatur von 20°C aufbewahrt, Gefäß 2 bei 45°C. Alle drei Gefäße besitzen einen speziellen Verschluss, der verhindert, dass Luft von außen in das Gefäß kommt. Die Tabelle verdeutlicht, wie Dennis vorgeht.

Zutaten	Gefäß 1	Gefäß 2	Gefäß 3
Apfelsaft	naturtrüb	naturtrüb	naturtrüb
Temperatur (°C)	20°C	45°C	20°C
Zucker (Gramm)	450gr	450gr	900gr
Kann Luft von außen in das Gefäß gelangen?	nein	nein	nein

Warum macht Dennis dieses Experiment?

- A Weil er herausfinden will, ob Apfelwein gelingt, wenn man viel Zucker hinzufügt.
- B Weil er denkt, dass Apfelwein gut gelingt, wenn keine Luft von außen in das Gefäß gelangen kann.
- C Weil er sicher gehen will, dass in den drei Gefäßen guter Apfelwein entsteht.
- D Weil er glaubt, dass die Weinherstellung von der Temperatur und der Zuckerzugabe abhängt.

Aufgabe 2

Nach einem Monat erhält Dennis in seinem Experiment die folgenden Ergebnisse: In Gefäß 1 entstand Wein, der wenig Alkohol enthält. In Gefäß 3 entstand Wein mit einem hohen Alkoholgehalt. In Gefäß 2 entstand Essig.

Zutaten	Gefäß 1	Gefäß 2	Gefäß 3
Apfelsaft	naturtrüb	naturtrüb	naturtrüb
Temperatur (°C)	20°C	45°C	20°C
Rohrzucker (Gramm)	450gr	450gr	900gr
Kann Luft von außen in das Gefäß gelangen?	nein	nein	nein
Ergebnis	Wein mit wenig Alkohol	Essig	Wein mit viel Alkohol

Wie lautet die beste Erklärung für die Ergebnisse?

- A Die Temperatur und die zugegebene Zuckermenge sind wichtig für die Weinherstellung.
- B Gut schmeckender Apfelwein entsteht, wenn man Hefe verwendet.
- C Die Ergebnisse zeigen, dass keine Luft von außen in das Gefäß kommen darf.
- D Apfelsaft wird zu Apfelwein, wenn man eine große Menge Zucker zum Apfelsaft gibt.

Aufgabe 3

Dennis verwendet normalerweise Rohrzucker, wenn er Apfelwein aus Apfelsaft herstellt. Aber er glaubt, dass er genau so gut **Traubenzucker anstatt Rohrzucker verwenden kann**. Er plant ein Experiment, um seine Vermutung zu überprüfen.

Er nimmt zwei Gefäße und gießt die gleiche Menge naturtrüben Apfelsaft in jedes Gefäß. Dann fügt er 200 Gramm Rohrzucker zum Apfelsaft in Gefäß 1 und bewahrt das Gefäß bei einer Raumtemperatur von 20°C auf. Gefäß 2 enthält 400 Gramm Traubenzucker und wird an einen warmen Ort mit einer Temperatur von 25°C gestellt. Beide Gefäße haben einen speziellen Verschluss, so dass kein Sauerstoff von außen in das Gefäß gelangt. Die Fotos zeigen, wie Dennis vorgeht.

Gefäß 1



**Rohrzucker: 200g / 20°C /
Es gelangt keine Luft von
außen in das Gefäß.**

Gefäß 2



**Traubenzucker: 400g /25°C/
Es gelangt keine Luft von
außen in das Gefäß.**

Dennis Vater schaut sich das Experiment an. Es sagt, dass Dennis das Experiment schlecht geplant habe und seine Vermutung nicht überprüfen kann. Warum?












- A Weil die beiden Gefäße an Orten mit unterschiedlicher Temperatur aufbewahrt werden.
- B Weil in den beiden Gefäßen die Zuckermenge und die Temperatur unterschiedlich sind.
- C Weil in das zweite Gefäß ebenfalls Rohrzucker hinzugefügt werden sollte.
- D Weil in beiden Gefäß durch einen speziellen Verschluss verhindert wurde, dass Luft von außen in das Gefäß gelangt.

Unit 4: Brot backen



Aufgabe 1

Tobias macht ein Experiment zum Brotbacken. Er benutzt drei unterschiedliche Schüsseln. In Schüssel 1 und Schüssel 3 rührt er Teig aus Butter, Hefe und weißem Mehl an. Für den Teig in Schüssel 2 verwendet er keine Hefe. Zum Anrühren wird in Schüssel 1 und Schüssel 2 Wasser mit einer Temperatur von 30°C verwendet. In Schüssel 3 verwendet er Wasser mit einer Temperatur von 70°C.

Schüssel 1:		+		+		+	
	Butter		Hefe		Weißes Mehl		Wasser 30°C
Schüssel 2:			+			+	
	Butter				Weißes Mehl		Wasser 30°C
Schüssel 3:		+		+		+	
	Butter		Hefe		Weißes Mehl		Wasser 70°C

Warum macht Tobias dieses Experiment?

- A Weil er vermutet, dass er besonders gutes Brot erhält, wenn man weißes Mehl und Wasser mit einer Temperatur von 30°C nimmt.
- B Weil er vermutet, dass er wichtig ist, ob man Hefe nimmt und welche Temperatur das Wasser hat.
- C Weil er gutes Brot backen will.
- D Weil er überprüfen will, ob er besonders gutes Brot erhält, wenn man Hefe nimmt.

Aufgabe 2

Tobias erzielte das folgende Ergebnis in seinem Experiment: Das Brot aus Schüssel 1 wurde weich und luftig, das Brot aus Schüssel 2 und Schüssel 3 hart und fest.

Schüssel 1



weißes Mehl/ Hefe/ Zucker/
Butter/ 30°C

Das Brot ist weich und luftig.

Schüssel 2



weißes Mehl/ Zucker/
Butter/ 30°C

Das Brot ist hart und fest.

Schüssel 2



weißes Mehl/ Hefe/ Zucker/
Butter/ 70°C

Das Brot ist hart und fest.

Wie lautet die beste Erklärung für die Ergebnisse?

- A Brot wird weich und luftig, wenn man Butter und Hefe nimmt.
- B Das Experiment klappte nicht, weil das Brot in Schüssel 2 und 3 hart und fest wurde.
- C Die besten Ergebnisse beim Brotbacken werden erzielt, wenn man Hefe und Wasser mit einer Temperatur von 30°C nimmt.
- D Die besten Ergebnisse beim Brotbacken werden erzielt, wenn man Hefe nimmt.

Aufgabe 3

Tobias vermutet, dass er **Margarine anstatt Butter zum Brotbacken verwenden kann und dass das Brot trotzdem gut gelingen wird.**

Er plant ein Experiment, um seine Vermutung zu überprüfen. Hierfür stehen vier verschiedene Teige zur Auswahl:

Schüssel 1



**Butter/
Weißes Mehl
Zucker**

Schüssel 2



**Margarine/
Vollkornmehl/
Zucker**

Schüssel 3



**Margarine/
Weißes Mehl/
Zucker**

Schüssel 4



**Butter/
Vollkornmehl/
Honig**

Welche Teige sollte er anmischen, um seine Vermutung zu überprüfen?

- A Schüssel 2 und Schüssel 4
- B Schüssel 1 und Schüssel 2
- C Schüssel 1 und Schüssel 3
- D Schüssel 3 und Schüssel 4



Unit 5: Wie wachsen Bohnenpflanzen?

Aufgabe 1

Maria möchte mehr darüber wissen, wie Bohnenpflanzen wachsen. Deshalb führt sie ein Experiment durch. Sie sät Bohnensamen in drei Töpfe mit Erde und lässt sie keimen. Anschließend lässt sie die jungen Bohnenpflanzen in den drei Töpfen weiter wachsen. Sie achtet darauf, dass alle drei Töpfe Wasser und frische Luft bekommen. Sie stellt aber Topf 2 ins Dunkle und Topf 3 ins Kühle (siehe Abbildungen).

Topf 1



**Wasser/Licht/
Luft / 22°C**

Topf 2



**Wasser/kein Licht /
Luft / 22°C**

Topf 3



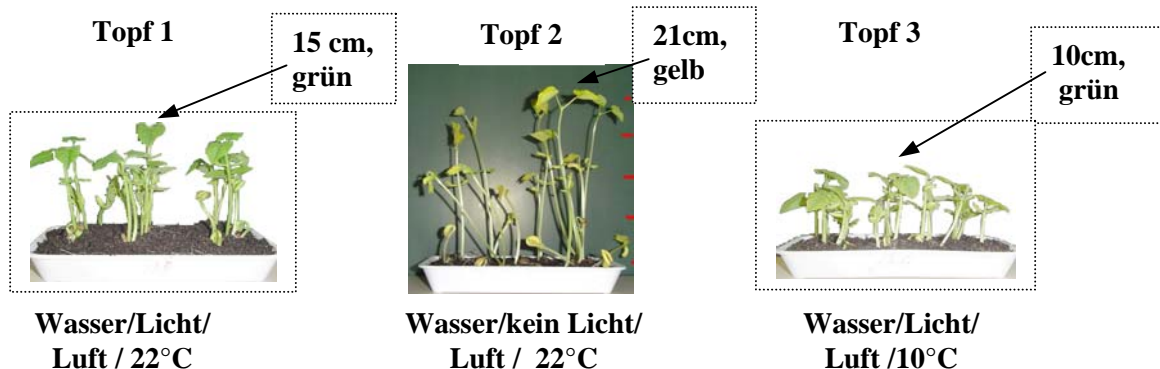
**Wasser/ Licht/
Luft / 10°C**

Warum macht Maria dieses Experiment?

- A Weil sie vermutet, dass die Temperatur und das Licht für das Bohnenwachstum wichtig sind.
- B Damit die Bohnenpflanzen schneller wachsen.
- C Weil sie vermutet, dass Bohnenpflanzen besonders gut wachsen, wenn sie Wasser und frische Luft bekommen.
- D Weil sie vermutet, dass Bohnenpflanzen besonders gut wachsen, wenn man sie gießt.

Aufgabe 2

Nach 3 Tagen erzielt Maria die folgenden Ergebnisse. Die Bohnenpflanzen in Topf 1 waren 15 cm hoch und ihre Farbe war grün. Die Bohnenpflanzen in Topf 2 waren 21 cm hoch und gelb. In Topf 3 waren die Bohnenpflanzen 10 cm hoch und grün.



Wie lautet die beste Erklärung für die Ergebnisse?

- A Das Experiment klappte nicht, da in Topf 2 etwas mit dem Pflanzenwachstum nicht stimmt.
- B Licht und Temperatur bestimmen das Wachstum der Bohnenpflanzen.
- C Bohnenpflanzen wachsen besonders schnell, wenn man sie gießt.
- D Frische Luft und Wasser beeinflussen das Wachstum der Bohnenpflanzen.

Aufgabe 3

Maria vermutet, dass Bohnenpflanzen **schneller in feuchter Erde als in trockener Erde wachsen**. Sie plant ein Experiment, um ihre Vermutung zu überprüfen. Sie nimmt drei Töpfe mit jungen Bohnenpflanzen und lässt diese unter den gleichen Bedingungen wachsen. Jedoch stellt sie Topf 3 in die Kälte und Topf 2 ins Dunkle. Sie gießt die drei Töpfe unterschiedlich stark.

Hier siehst du die drei Töpfe, die Maria verwendet:

Topf 1



**Licht/ 22°C/
10 ml Wasser pro Tag**

Topf 2



**kein Licht/ 22°C/
20 ml Wasser pro Tag**

Topf 3



**Licht/ 10°C/
30 ml Wasser pro Tag**

Marias Lehrer sagte, dass sie das Experiment schlecht geplant habe und dass sie ihre Vermutung nicht überprüfen kann. Warum?

- A Topf 2 erhält kein Licht, aber Topf 1 und Topf 3 erhalten Licht.
- B Topf 1 erhält weniger Wasser als Topf 2 und Topf 3.
- C Topf 3 wird bei 10°C aufbewahrt, Topf 1 und 2 bei 22°C.
- D Topf 2 erhält kein Licht und Topf 3 wird bei 10°C aufbewahrt.

Unit 6: Kartoffeln

Aufgabe 1

Sandy und Anna machen ein Experiment mit Kartoffeln. Sie nehmen drei Töpfe, füllen diese mit Erde, legen Kartoffeln hinein und stellen die Töpfe an gut belüftete Plätze. Topf 1 und 3 werden ins Licht gestellt; Topf 2 ins Dunkle. Topf 1 und 2 werden bei einer Temperatur von 20°C aufbewahrt; Topf 3 bei einer Temperatur von 6°C.

Topf 1



**Erde/ Licht/
Luft /20°C**

Topf 2



**Erde/ kein Licht/
Luft /20°C**

Topf 3



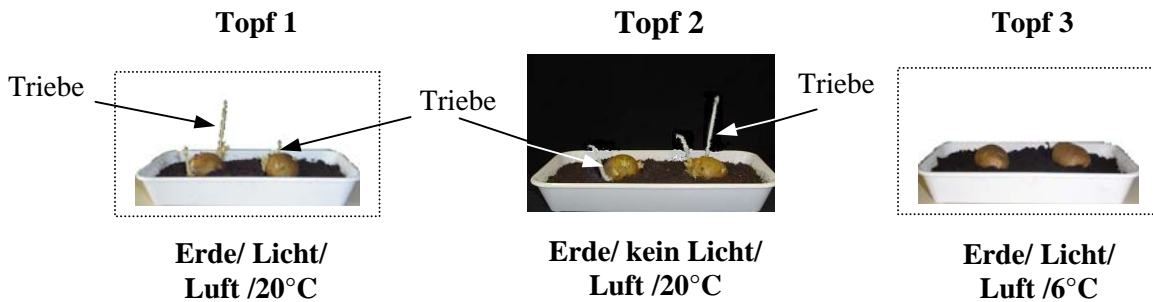
**Erde/ Licht/
Luft /6°C**

Warum machen Sandy und Anna dieses Experiment?

- A Weil sie vermuten, dass Kartoffeln Wärme und Licht brauchen, um zu keimen.
- B Weil sie vermuten, dass Kartoffeln Wärme und Luft brauchen, um zu keimen.
- C Weil sie vermuten, dass Kartoffeln Erde und Luft brauchen, um zu keimen.
- D Weil sie Kartoffeln besonders schnell keimen lassen wollen.

Aufgabe 2

Nach 10 Tagen konnten Sandy und Anna sehen, dass die Kartoffeln in Topf 1 und Topf 2 kleine Triebe hatten. Die Kartoffeln in Topf 3 waren nicht gekeimt. Dies ist auf den Fotos zu erkennen. Die Pfeile zeigen auf die Triebe.



Wie lautet die beste Erklärung für das Ergebnis?

- A Das Experiment klappte nicht, weil die Kartoffeln in Topf 3 nicht keimten.
- B Kartoffeln keimen besonders gut, wenn man sie ins Licht und in die Wärme stellt.
- C Kartoffeln keimen besonders gut, wenn man sie in die Erde legt und ihnen Luft gibt.
- D Kartoffeln keimen bei Wärme besonders gut; Licht ist für die Keimung nicht nötig.

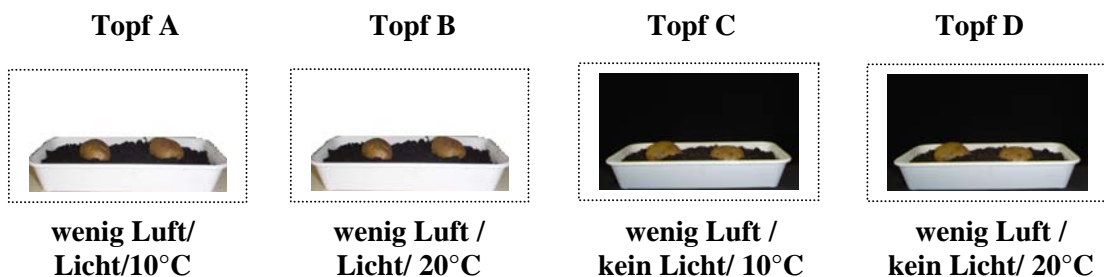
Aufgabe 3

Sandy und Anna vermuten, dass **Luft für die Kartoffelkeimung notwendig ist**.

Um diese Vermutung zu überprüfen, planen sie ein Experiment. Sie legen Kartoffeln in einen Topf mit Erde, stellen den Topf ins Licht und sorgen für eine Temperatur von 20°C (Topf 1).



Sandy und Anna brauchen aber noch einen zweiten Topf mit Kartoffeln, damit sie ihre Vermutung überprüfen können. Es stehen 4 Töpfe zur Auswahl.



Welchen der folgenden Töpfe sollen sie mit Topf 1 vergleichen, um ihre Vermutung zu überprüfen?

- A Topf A
- B Topf B
- C Topf C
- D Topf D

Unit 7: Herzschlag



Aufgabe 1

Alex weiß, dass das Herz des Menschen nicht immer gleich schnell schlägt. Er möchte wissen, warum das so ist. Er führt deshalb eine Untersuchung an drei Personen durch: an zwei 20jährigen Männern und an einem 10jährigen Jungen. Den Herzschlag des einen Mannes misst er, nachdem dieser 10 Minuten zügig gegangen ist. Den Herzschlag der anderen beiden Versuchspersonen misst er, nachdem sich diese eine Zeit lang ausgeruht haben. Hier siehst du die Einzelheiten seiner Untersuchung:

	Versuchsperson 1	Versuchsperson 2	Versuchsperson 3
Geschlecht	männlich	männlich	männlich
Alter (Jahre)	20	20	10
Was geschah kurz vor der Untersuchung?	Der Mann ging 10 Minuten zügig.	Der Mann ruhte sich aus.	Der Junge ruhte sich aus.

Warum macht Alex diese Untersuchung

- A Weil er einen schnelleren Herzschlag anregen möchte.
- B Weil er vermutet, dass es vom Alter und von der körperlichen Aktivität abhängt, wie häufig das Herz schlägt.
- C Weil er vermutet, dass das Alter und das Geschlecht beeinflussen, wie schnell das Herz schlägt.
- D Weil er vermutet, dass die körperliche Aktivität beeinflusst, wie schnell das Herz schlägt.

Aufgabe 2

Alex erzielte die folgenden Ergebnisse in seiner Untersuchung:

	Versuchsperson 1	Versuchsperson 2	Versuchsperson 3
Geschlecht	männlich	männlich	männlich
Alter (Jahre)	20	20	10
Was geschah kurz vor der Untersuchung?	Der Mann ging 10 Minuten zügig.	Der Mann ruhte sich aus.	Der Junge ruhte sich aus.
Ergebnisse (Herzschläge/Minute)	100	72	90

Wie lautet die beste Erklärung für die Ergebnisse?

- A Alex Experiment klappte nicht, denn die Untersuchungsergebnisse sind in den drei Gruppen unterschiedlich.
- B Bei 20jährigen Männern schlägt das Herz schneller als bei 10jährigen Jungen.
- C Das Ausführen von Übungen und das Alter beeinflussen, wie häufig das Herz schlägt.
- D Je älter die Leute sind, desto schneller schlägt das Herz.

Aufgabe 3

Alex glaubt, dass das Herz umso schneller schlägt, je intensiver man Sport treibt. Er plant ein Experiment, um seine Vermutung zu überprüfen. In dem Experiment sollen zwei Personen verglichen werden.

Er hat dabei die Wahl zwischen fünf verschiedenen Personen. Hier siehst du, wie alt die Versuchspersonen sind, welche körperliche Aktivität sie ausüben und ob die Personen männlich oder weiblich sind.

	Alter	Körperliche Aktivität	Geschlecht
Versuchsperson 1	10	laufen	weiblich
Versuchsperson 2	18	ruhig liegend	männlich
Versuchsperson 3	30	gehen	männlich
Versuchsperson 4	18	laufen	männlich

Bitte wähle aus, welche Gruppen er vergleichen soll, um seine Vermutung zu überprüfen.

- A Versuchsperson 1 + Versuchsperson 2
- B Versuchsperson 2 + Versuchsperson 4
- C Versuchsperson 3 + Versuchsperson 4
- D Versuchsperson 2 + Versuchsperson 3

**Unit 8: Pflanzenwachstum****Aufgabe 1**

Björn interessiert sich dafür, wie Pflanzen wachsen. Er plant hierzu ein Experiment. Dabei verwendet er drei Töpfe: In Topf 1 und Topf 2 füllt er Sand, in Topf 3 Blumenerde. Für die Töpfe 2 und 3 verwendet er Pflanzendünger, für Topf 1 verwendet er keinen Dünger. Dann pflanzt er je eine Jungpflanze vom Raps in die drei Töpfe und stellt sie ins Licht. Den Sand und die Blumenerde hält er feucht. Nach 14 Tagen misst Björn die Höhe der drei Pflanzen.

	Topf 1	Topf 2	Topf 3
Art der Erde	Sand	Sand	Blumenerde
Wasserzugabe	ja	ja	ja
Belichtung	den ganzen Tag über	den ganzen Tag über	den ganzen Tag über
Düngerzugabe	nein	ja	Ja

Warum macht Björn dieses Experiment?

- A Weil er vermutet, dass Licht und Wasser das Pflanzenwachstum beeinflussen.
- B Weil er möchte, dass die Jungpflanzen in den drei Töpfen besonders schnell wachsen.
- C Weil er vermutet, dass die Art der Erde und Dünger das Pflanzenwachstum beeinflussen.
- D Weil er vermutet, dass Licht und Dünger das Pflanzenwachstum beeinflussen.

Aufgabe 2

Nach 14 Tagen erhält Björn das folgende Ergebnis:

	Topf 1	Topf 2	Topf 3
Art der Erde	Sand	Sand	Blumenerde
Wasserzugabe	ja	ja	ja
Belichtung	den ganzen Tag über	den ganzen Tag über	den ganzen Tag über
Dünger	nein	ja	ja
Ergebnis: Höhe der Pflanzen	20 cm	30 cm	43 cm

Wie lautet die beste Erklärung für das Ergebnis?

- A Pflanzen wachsen besonders schnell, wenn sie gedüngt und gewässert werden.
- B Pflanzen wachsen besonders schnell, wenn sie in Blumenerde gepflanzt und gedüngt werden.
- C Das Experiment klappte nicht, denn die Pflanzen in Topf 1 wuchsen zu langsam.
- D Pflanzen wachsen besonders schnell, wenn sie gedüngt werden.

Aufgabe 3

Björn glaubt, **dass organischer Dünger für das Wachstum der Pflanzen besser sei als chemischer Dünger.**

Um seine Vermutung zu überprüfen, plant er ein Experiment. Er pflanzt je eine Jungpflanze vom Raps in zwei Töpfe. In Topf 1 befindet sich Sand und chemischer Dünger. Topf 2 enthält Blumenerde und organischen Dünger. Er wässert beide Töpfe und stellt sie ans Licht. Nach 14 Tagen soll die Höhe der Pflanzen gemessen werden.

Topf 1

**Sand/ Wasser/ Licht/
chemischer Dünger**

Topf 2

**Blumenerde/ Wasser/ Licht/
organischer Dünger**

Björn berichtet seinem Biologielehrer von seiner Planung. Dieser sagt, Björn habe sein **Experiment nicht so geplant, dass er seine Vermutung überprüfen kann.**

Warum ist Björns Experiment schlecht geplant?

- A Weil die Jungpflanzen in Topf 1 und Topf 2 mit unterschiedlichen Düngern gedüngt wurden.
- B Weil die beiden Töpfe mit Jungpflanzen unterschiedliche Arten von Erde enthalten.
- C Weil beide Töpfe mit den Jungpflanzen ins Licht gestellt und gewässert wurden.
- D Weil die Pflanzen in Topf 1 auch organischen Dünger erhalten müssen.

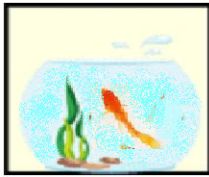
Unit 9: Atmung der Fische



Aufgabe 1

Laura macht ein Experiment zur Atmung der Fische. Hierzu verwendet sie drei gleichgroße Aquarien. Sie setzt Goldfische in drei Aquarien, und zwar je einen Fisch in Aquarium 1 und 3 und drei Fische in das Aquarium 2. Alle Fische sind gleich groß. Die Wassertemperatur beträgt 20°C in den Aquarien 1 und 2. Im Aquarium 3 ist 10°C kaltes Wasser. In jedem Aquarium befindet sich eine Wasserpflanze. Nach einiger Zeit beobachtet Laura, wie häufig die Fische pro Minute atmen. Dies erkennt sie daran, wie schnell sich die Kiemendeckel der Fische bewegen.

Aquarium 1



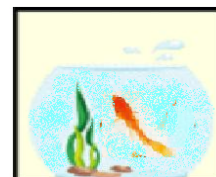
**1 Wasserpflanze /
1 Goldfisch / 20°C**

Aquarium 2



**1 Wasserpflanze /
3 Goldfische / 20°C**

Aquarium 3



**1 Wasserpflanze /
1 Goldfisch / 10°C**

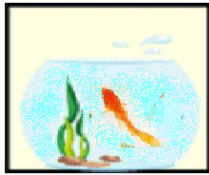
Warum machte Laura dieses Experiment?

- A Weil sie vermutet, dass Wasserpflanzen im Aquarium einen Einfluss darauf haben, wie schnell die Fische atmen.
- B Weil sie vermutet, dass die Zahl der Fische und die Wassertemperatur im Aquarium einen Einfluss darauf haben, wie schnell die Fische atmen.
- C Weil sie die Fische in den drei Aquarien dazu bringen möchte, schneller zu atmen.
- D Weil sie vermutet, dass die Zahl der Fische und die Größe des Aquariums einen Einfluss darauf haben, wie schnell die Fische atmen.

Aufgabe 2

Lauras Experiment brachte das folgende Ergebnis: In Aquarium 1 bewegen sich die Kiemendeckel des Goldfisches pro Minute 96 Mal, im Aquarium 2 pro Minute 108 Mal und im Aquarium 3 pro Minute 73 Mal.

Aquarium 1



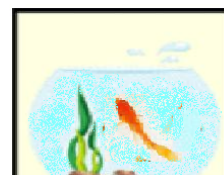
**1 Goldfisch / 20°C /
96 Kiemendeckel-
bewegungen pro Minute**

Aquarium 2



**3 Goldfische / 20°C /
108 Kiemendeckel-
bewegungen pro Minute**

Aquarium 3



**1 Goldfisch / 10°C /
73 Kiemendeckel-
bewegungen pro Minute**

Wie lautet die beste Erklärung für das Ergebnis?

- A Fische atmen besonders schnell, wenn mehrere Fische und Pflanzen im Aquarium sind.
- B Fische atmen besonders schnell, wenn das Wasser warm ist und mehrere Fische im Aquarium sind.
- C Fische atmen besonders schnell, wenn sie in einem Aquarium mit warmem Wasser sind.
- D Das Experiment funktionierte nicht, weil die Fische in den drei Aquarien verschieden schnell atmeten.

Aufgabe 3

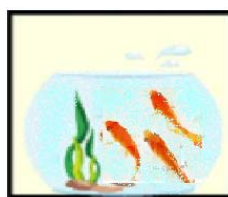
Laura glaubt, dass **Goldfische bei einer höheren Wassertemperatur schneller atmen als bei einer niedrigen Wassertemperatur**. Sie plant ein Experiment, um ihre Vermutung zu überprüfen. Sie verwendet drei gleich große Aquarien. In die Aquarien 1 und 2 füllt sie je 5 Liter Wasser und 3 Liter in Aquarium 3. Sie setzt zwei Fische in Aquarium 1, drei Fische in Aquarium 2 und einen Fisch in Aquarium 3. Die Temperatur im Aquarium 1 beträgt 20°C, 10°C im Aquarium 2 und 30°C im Aquarium 3.

Aquarium 1



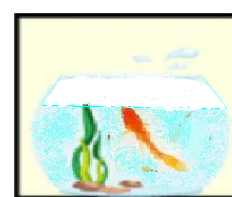
**20°C / 2 Fische/
5 Liter Wasser**

Aquarium 2



**10°C / 3 Fische/
5 Liter Wasser**

Aquarium 3



**30°C / 1 Fisch/
3 Liter Wasser**

Ihr Biologielehrer sagt, dass sie das Experiment falsch geplant habe und ihre Vermutung nicht überprüfen kann. Warum?

- A Weil Laura unterschiedlich viele Fische in die 3 Aquarien gesetzt hat.
- B Weil das Wasser in den Aquarien unterschiedlich warm ist.
- C Weil Laura unterschiedlich viel Wasser in die drei Aquarien gefüllt hat.
- D Weil die Zahl der Fische und die Wassermenge in den 3 Aquarien unterschiedlich sind.

2. Knowledge test

Unit 1: Samenkeimung

1)

Was ist in einem Samen enthalten?	Ja oder Nein?
Nährstoffvorräte	Ja / Nein
Blüten	Ja / Nein
einen ruhenden Pflanzenembryo	Ja / Nein
weitere kleine Samen	Ja / Nein
Wasser	Ja / Nein
zusammensetzbare Pflanzenteile	Ja / Nein

2)

Woher stammen Samen?	Ja oder Nein?
von den Wurzeln	Ja / Nein
von den Stängeln	Ja / Nein
von den Blättern	Ja / Nein
von den Blüten	Ja / Nein

3)

Welche Aufgaben haben Samen?	Ja oder Nein?
Sie helfen bei der Ausbreitung der Pflanze.	Ja / Nein
Sie helfen bei der Vermehrung der Pflanze.	Ja / Nein
Sie ermöglichen die Überwinterung der Pflanze.	Ja / Nein
Sie halten Pflanzen gesund.	Ja / Nein
Sie dienen dem Pflanzenwachstum.	Ja / Nein

4)

Was passiert bei der Samenkeimung?	Ja oder Nein?
Der Same nimmt Wasser auf.	Ja / Nein
Der Same wird grün.	Ja / Nein
Aus dem Samen wächst eine Pflanze heraus.	Ja / Nein
Der Same stirbt ab.	Ja / Nein

5)

Was kann man beobachten, nachdem man Samen für einige Stunden ins Wasser gelegt hat?	Ja oder Nein?
Der Same wird leichter.	Ja / Nein
Der Same wird größer.	Ja / Nein
Der Same wird weicher.	Ja / Nein
Der Same verändert seine Farbe.	Ja / Nein

6) In ägyptischen Pyramiden hat man sehr alte Samen gefunden, die immer noch keimen können.

Welche Aussage erklärt die lange Keimfähigkeit von Samen?	Ja oder Nein?
Die Samen sind tot und werden durch Wasser zum Leben erweckt.	Ja / Nein
In den Samen ruhen winzige Pflanzen, die sehr lange leben können.	Ja / Nein
Samen haben eine Schutzschicht und bleiben deshalb so lange frisch.	Ja / Nein
Samen besitzen Stoffe, die das Altern verhindern.	Ja / Nein

Unit 2: Hühnereier

1)

Was ist in einem Hühnerei enthalten, wenn es frisch gelegt wird?	Ja oder Nein?
Eigelb	Ja / Nein
Eiklar	Ja / Nein
Küken	Ja / Nein
Luft	Ja / Nein
Wasser	Ja / Nein

2)

Sind die folgenden Aussagen richtig oder falsch?	Richtig oder Falsch?
Eier dienen zur Vermehrung von Hühnern.	richtig / falsch
Sauerstoff gelangt durch die Kalkschale ins Ei.	richtig / falsch
Die Henne hilft dem Küken beim Schlüpfen.	richtig / falsch
Nur aus befruchteten Eiern entstehen Küken.	richtig / falsch
Die Kalkschale schützt das Hühnerei.	richtig / falsch

3)

Was geschieht, wenn ein Hühnerei ausgebrütet wird?	Ja oder Nein?
Ein Hühnerembryo entwickelt sich.	Ja / Nein
Zellen des Hühnerembryos teilen sich.	Ja / Nein
Samenzelle und Eizelle verschmelzen.	Ja / Nein
Die Kalkschale bildet sich.	Ja / Nein
Das Eidotter bildet sich.	Ja / Nein

4)

Wo bildet sich das Küken?	Ja oder Nein?
im Eileiter	Ja / Nein
im Eiweiß	Ja / Nein
im Eidotter	Ja / Nein
In den Hagelschnüren	Ja / Nein

5)

Welche Tiere legen Eier?	Ja oder Nein?
Schnecke	Ja / Nein
Amsel	Ja / Nein
Laubfrosch	Ja / Nein
Käfer	Ja / Nein
Biber	Ja / Nein

6)

Bei welcher Temperatur lassen sich Hühnereier am besten ausbrüten?	Ja oder Nein?
28°C	Ja / Nein
38°C	Ja / Nein
46°C	Ja / Nein
55°C	Ja / Nein

7)

Was ist notwendig, damit Eier ausgebrütet werden?	Ja oder Nein?
Wärme	Ja / Nein
Luft	Ja / Nein
Licht	Ja / Nein
Luftfeuchtigkeit	Ja / Nein
Drehen der Eier	Ja / Nein

Unit 3: Apfelwein

1)

Was entsteht bei der Weinherstellung?	Ja oder Nein?
Zucker	Ja / Nein
Hefe	Ja / Nein
Sauerstoff	Ja / Nein
Kohlenstoffdioxid	Ja / Nein
Alkohol	Ja / Nein
Essig	Ja / Nein

2)

Was ist notwendig, um Wein herzustellen?	Ja oder Nein?
Fruchtsaft	Ja / Nein
Hefe	Ja / Nein
Luft	Ja / Nein
Licht	Ja / Nein

3)

Wie würdest du die Herstellung von Wein beschreiben?	Ja oder Nein?
Verdampfung von Fruchtsaft	Ja / Nein
Vergärung von Fruchtsaft	Ja / Nein
Verunreinigung von Fruchtsaft	Ja / Nein
Veratmung von Fruchtsaft	Ja / Nein

4)

Für die Herstellung welcher Lebensmittel braucht man Hefe?	Ja oder Nein?
Jogurt	Ja / Nein
Bier	Ja / Nein
Wein	Ja / Nein
Brot	Ja / Nein
Sauerkraut	Ja / Nein

5)

Wann enthält Wein besonders viel Alkohol?	Ja oder Nein?
Wenn man den Wein besonders lang lagert.	Ja / Nein
Wenn der Wein besonders süß ist.	Ja / Nein
Wenn der Fruchtsaft zur Weinherstellung besonders süß ist.	Ja / Nein
Wenn der Wein besonders viel Wasser enthält.	Ja / Nein

6)

Wie kommt der Alkohol in den Wein?	Ja oder Nein?
Alkohol entsteht bei der Lagerung von fertigem Wein.	Ja / Nein
Alkohol wird dem Wein vor der Abfüllung zugesetzt.	Ja / Nein
Alkohol entsteht bei der Weinherstellung.	Ja / Nein
Alkohol ist schon vor der Weinherstellung vorhanden.	Ja / Nein

Unit 4: Brotbacken

1)

Wodurch wird Brot luftig?	Ja oder Nein?
Butter	Ja / Nein
Zucker	Ja / Nein
Hefe	Ja / Nein
Salz	Ja / Nein
Backpulver	Ja / Nein

2)

Was ist Hefe?	Ja oder Nein?
eine chemische Substanz	Ja / Nein
ein lebender Organismus	Ja / Nein
Backpulver	Ja / Nein
eine Backmischung	Ja / Nein
ein Enzym	Ja / Nein
eine Art Pilz	Ja / Nein

3)

Warum soll man beim Anmischen von Hefeteig kein kochendes Wasser verwenden?	Ja oder Nein?
Weil sich die Butter zu schnell auflöst.	Ja / Nein
Weil man das Mehl verbrühen würde.	Ja / Nein
Weil der Teig verklumpen würde.	Ja / Nein
Weil man die Hefe schädigen würde.	Ja / Nein
Weil man den Zucker zu schnell auflöst.	Ja / Nein

4)

Was passiert, wenn der Hefeteig „geht“?	Ja oder Nein?
Der Hefeteig wird schwerer.	Ja / Nein
Der Hefeteig wird größer.	Ja / Nein
Der Hefeteig wird leichter.	Ja / Nein
Im Hefeteig entstehen Blasen.	Ja / Nein
Der Hefeteig bekommt Füße.	Ja / Nein

5)

Bei welcher Temperatur wird Brot gebacken?	Ja oder Nein?
50°C – 79°C	Ja / Nein
80°C – 149°C	Ja / Nein
150°C – 249°C	Ja / Nein
250°C - 500°C	Ja / Nein

Unit 5: Bohnenwachstum

1)

Was nehmen Bohnenpflanzen über ihre Wurzeln auf?	Ja oder Nein?
Wasser	Ja / Nein
Licht	Ja / Nein
Vitamine	Ja / Nein
Pflanzennährstoffe	Ja / Nein
Luft	Ja / Nein

2)

Was geschieht, wenn junge Bohnenpflanzen einige Tage ins Dunkle gestellt werden?	Ja oder Nein?
Sie sterben sofort.	Ja / Nein
Sie wachsen genauso schnell weiter wie vorher.	Ja / Nein
Sie treiben Fotosynthese.	Ja / Nein
Sie werden gelblich.	Ja / Nein

3)

Was passiert mit jungen Bohnenpflanzen, wenn sie einige Tage von einem warmen Platz (20°C) an einen kalten Platz (10°C) gestellt werden?	Ja oder Nein?
Sie erfrieren.	Ja / Nein
Ihr Wachstum wird langsamer.	Ja / Nein
Sie produzieren mehr Wärme.	Ja / Nein
Ihr Stoffwechsel wird schneller.	Ja / Nein

4)

Was brauchen Bohnenpflanzen zum Wachstum?	Ja oder Nein?
Wasser	Ja / Nein
Luft	Ja / Nein
Zuckerlösung	Ja / Nein
Licht	Ja / Nein
Pflanzennährstoffe	Ja / Nein

5)

Sind die folgenden Aussagen richtig oder falsch?	Richtig oder Falsch?
Pflanzen produzieren Sauerstoff.	richtig / falsch
Pflanzen treiben Fotosynthese.	richtig / falsch
Pflanzen müssen mit den Wurzeln Erde aufnehmen, um wachsen zu können.	richtig / falsch
Pflanzen wachsen alleine durch die Stoffe in der Luft.	richtig / falsch

Unit 6: Kartoffeln

1)

Welche Teile der Kartoffelpflanze sind essbar?	Ja oder Nein?
Laubblätter	Ja / Nein
Beeren	Ja / Nein
Blüten	Ja / Nein
Knollen	Ja / Nein

2)

Welche Aufgabe haben Kartoffelknollen?	Ja oder Nein?
Samen verbreiten	Ja / Nein
Stoffe speichern	Ja / Nein
Blüten bilden	Ja / Nein
überwintern	Ja / Nein
Beeren bilden	Ja / Nein

3)

Was ist sind notwendig, damit Kartoffeln keimen?	Ja oder Nein?
Wasser	Ja / Nein
Licht	Ja / Nein
Wärme	Ja / Nein
Erde	Ja / Nein
Dünger	Ja / Nein

4)

Kartoffeln sind Teile von Erdsprossen und keine Wurzeln. Woran kann man dies erkennen?	Ja oder Nein?
Kartoffeln werden an der Erdoberfläche grün.	Ja / Nein
Kartoffeln wachsen an Erdsprossen.	Ja / Nein
Kartoffel bilden neue Pflanzen.	Ja / Nein
Kartoffeln sind essbar.	Ja / Nein

5)

Was passiert bei der Kartoffelkeimung?	Ja oder Nein?
Die keimende Kartoffel wird grün.	Ja / Nein
Die Triebe wachsen vom Licht weg in die Erde.	Ja / Nein
Aus den „Augen“ der Kartoffel wachsen Triebe.	Ja / Nein
Die Kartoffel schrumpft.	Ja / Nein

6)

Wie kann man verhindern, dass Kartoffel zu früh keimen?	Ja oder Nein?
Lagerung bei Temperaturen unter 10°C	Ja / Nein
einfrieren	Ja / Nein
Lagerung bei Licht	Ja / Nein
häufiges Drehen	Ja / Nein
Kartoffeln und Äpfel getrennt lagern	Ja / Nein

Unit 7: Herzschlag

1) Das Blut fließt auch in den Daumen.

Was passiert, wenn das Blut im Daumen angekommen ist?	Ja oder Nein?
Das Blut bleibt dort.	Ja / Nein
Das Blut fließt auf gleichem Weg wieder zurück.	Ja / Nein
Das Blut bewirkt, dass der Daumen sich bewegt	Ja / Nein
Das Blut wird dort verbraucht.	Ja / Nein
Das Blut fließt auf anderem Weg wieder zurück.	Ja / Nein

2)

Wie bekommt das Gehirn Sauerstoff?	Ja oder Nein?
durch Luftkanäle	Ja / Nein
durch Blutadern	Ja / Nein
durch die Haare	Ja / Nein
durch das Rückenmark	Ja / Nein
durch die Haut	Ja / Nein

3)

Wie bekommen die Muskeln Nährstoffe?	Ja oder Nein?
durch Blutadern	Ja / Nein
durch Nährstoffkanäle	Ja / Nein
durch Knochen	Ja / Nein
durch Sehnen	Ja / Nein
durch Luftkanäle	Ja / Nein
durch Nerven	Ja / Nein

4)

Welche Aufgaben hat das Herz?	Ja oder Nein?
Das Herz pumpt Blut.	Ja / Nein
Das Herz säubert Blut.	Ja / Nein
Das Herz bildet Blut.	Ja / Nein
Das Herz durchmischt Blut und Atemluft.	Ja / Nein
Das Herz durchmischt Blut und Nährstoffe.	Ja / Nein

5)

Wann schlägt das Herz schneller als normal?	Ja oder Nein?
wenn man aufgeregt ist,	Ja / Nein
wenn man schläft,	Ja / Nein
wenn man Sport treibt,	Ja / Nein
wenn man hungrig ist,	Ja / Nein
wenn man ruhig im Bett liegt.	Ja / Nein

6)

Was passiert, wenn man Sport treibt?	Ja oder Nein?
Man fängt an zu schwitzen.	Ja / Nein
Die Muskeln werden stärker durchblutet.	Ja / Nein
Das Herz schlägt schneller.	Ja / Nein
Man atmet häufiger.	Ja / Nein
Das Blut wird besser gereinigt.	Ja / Nein

Unit 8: Pflanzenwachstum

1)

Was nehmen Pflanzen über ihre Wurzeln auf?	Ja oder Nein?
Wasser	Ja / Nein
Kohlenstoffdioxid	Ja / Nein
Erde	Ja / Nein
Pflanzennährstoffe	Ja / Nein

2)

Was findet man in chemischen Düngermitteln für Pflanzen?	Ja oder Nein?
Pflanzennährstoffe	Ja / Nein
Fette	Ja / Nein
Zucker	Ja / Nein
Stickstoff	Ja / Nein
Sauerstoff	Ja / Nein

3)

Findet bei diesen Dingen Pflanzenwachstum statt?	Ja oder Nein?
Zellen einer Pflanze teilen sich und werden größer.	Ja / Nein
Ein Same nimmt Wasser auf.	Ja / Nein
Eine Biene bestäubt eine Blüte.	Ja / Nein
Ein Windstoß verteilt die Samen vom Löwenzahn.	Ja / Nein

4)

Welche Stoffe müssen Pflanzen aufnehmen, damit sie gut wachsen?	Ja oder Nein?
Sauerstoff	Ja / Nein
Mineralien	Ja / Nein
Vitamine	Ja / Nein
Kohlenstoffdioxid	Ja / Nein

5)

Woraus bestehen Pflanzen?	Ja oder Nein?
Muskeln	Ja / Nein
Zellen	Ja / Nein
Nerven	Ja / Nein
aus vielen Geweben	Ja / Nein

Unit 9: Atmung der Fische

1)

Womit atmen Fische?	Ja oder Nein?
mit der Lunge	Ja / Nein
mit den Kiemen	Ja / Nein
mit dem Herzen	Ja / Nein
mit den Schuppen	Ja / Nein

2)

Was passiert, wenn Fische atmen?	Ja oder Nein?
Sie nehmen Wasserstoff auf.	Ja / Nein
Sie nehmen Sauerstoff auf.	Ja / Nein
Sie nehmen Nährstoffe auf.	Ja / Nein
Sie nehmen Kohlenstoffdioxid auf.	Ja / Nein

3)

Wann beschleunigt sich die Fischatmung?	Ja oder Nein?
Bei Licht	Ja / Nein
Bei steigenden Wassertemperaturen	Ja / Nein
Wenn Fische schnell schwimmen	Ja / Nein
Im Winter	Ja / Nein

4)

Welche Ausrüstung sorgt für einen höheren Sauerstoffgehalt im Aquarium?	Ja oder Nein?
Wasserfilter	Ja / Nein
Aquariumpumpe	Ja / Nein
Wasserpflanzen	Ja / Nein
Wasserschnecken	Ja / Nein

5)

Wozu dient die Schwimmblase der Fische?	Ja oder Nein?
Zum Auftauchen	Ja / Nein
Zum Atmen	Ja / Nein
Zum Schweben	Ja / Nein
Zum rückwärts Schwimmen	Ja / Nein

Knowledge test for main test

Unit 1: Samenkeimung

1. Wie wichtig sind die folgenden Dinge, damit Samen keimen?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Licht				
Wärme				
Luft				
Erde				
Wasser				

2. Was ist in einem Samen enthalten?

- kleine Früchte
- Nährstoffe und Keimling
- weitere kleine Samen
- Pflanze mit Blüten

3. Woher stammen Samen?

- aus den Wurzeln
- aus den Stängeln
- aus den Blättern
- aus den Blüten

4. Was passiert bei der Samenkeimung?

- Der Same nimmt Wasser auf.
- Der Same nimmt Nährstoffe auf.
- Der Same nimmt Erde auf.
- Der Same nimmt Licht auf.

5. Was kann man beobachten, wenn Bohnensamen keimen?

- Zuerst erscheinen die Keimblätter.
- Zuerst erscheint die Keimwurzel.
- Zuerst erscheint die Spitze des Keimlings.
- Zuerst erscheint der Stängel des Keimlings.

6. In ägyptischen Pyramiden hat man sehr alte Samen gefunden, die immer noch keimen können. Wie sollten man Samen aufbewahren, damit sie lange Zeit überdauern, ohne zu keimen?

- bei hoher Luftfeuchtigkeit
- bei großer Trockenheit
- bei absoluter Sauberkeit
- bei absoluter Dunkelheit

7. Welche Aussage ist richtig?

- Samen brauchen Erde, um zu keimen.
- Samen müssen Nährstoffe aufnehmen, um zu keimen.
- Samen keimen schneller, wenn sie gedüngt werden.
- Samen können im Dunkeln keimen.

Unit 2: Hühnereier

1. Wie wichtig sind die folgenden Dinge, damit Hühnereier so schnell wie möglich ausgebrütet werden?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Licht				
Wärme				
feuchte Luft				
Eigröße				
Eifarbe				

2. Wozu gehören die Hühner

- zu den Pflanzenfressern
- zu den Körnerfressern
- zu den Insektenfressern
- zu den Allesfressern

3. Kreuze die falsche Antwort an.

- Ein frisch gelegtes Hühnerei enthält ein Eigelb.
- Ein frisch gelegtes Hühnerei enthält Eiklar.
- Ein frisch gelegtes Hühnerei enthält ein Küken.
- Ein frisch gelegtes Hühnerei besteht zu einem bestimmten Anteil aus Wasser.

4. Welches Tier legt keine Eier?

- Bienenkönigin
- Amsel
- Biber
- Laubfrosch

5. Bei welcher Temperatur lassen sich Hühnereier am besten ausbrüten?

- bei 8°C
- bei 18°C
- bei 38°C
- bei 58°C

6. Welche Aussage ist falsch?

- Hühner haben einen Kropf.
- Hühner haben Schuppen.
- Hühner haben Zähne.
- Hühner haben einen Kaumagen.

7. Wo legen Hühner ihre Eier?

- im Geäst
- am Boden
- in Baumhöhlen
- in Felsnischen

8. Welche Aussage ist falsch? ?

- Beim Ausbrüten brauchen Hühnereier gleichbleibende Wärme.
- Beim Ausbrüten brauchen Hühnereier genügend Luftfeuchtigkeit.
- Beim Ausbrüten müssen Hühnereier häufig gewendet werden.
- Beim Ausbrüten brauchen Hühnereier eine bestimmte Anzahl an Lichtstunden.

Unit 3: Apfelwein

1. Wie wichtig sind die folgenden Dinge, damit die Weinherstellung gelingt?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Licht				
Wärme				
ein gut verschließbares Gefäß, damit keine Luft hinein kommt				
Zuckermenge				
Hefe				

2. Was schadet der Weinherstellung?

- a. Hefe
- b. Licht
- c. Sauerstoff
- d. Fruchtsaft

3. Wie passiert bei der Herstellung von Wein?

- a. Fruchtsaft wird verdampft.
- b. Fruchtsaft wird vergoren.
- c. Fruchtsaft wird verdünnt.
- d. Fruchtsaft wird gekocht.

4. Welches Lebensmittel wird mit Hefe hergestellt?

- a. Apfelsaft
- b. Sauerkraut
- c. Wein
- d. Jogurt

5. Wann enthält Wein besonders viel Alkohol?

- a. Wenn man den Wein besonders lange lagert.
- b. Wenn man zusätzlichen Zucker zum Apfelsaft hinzu gibt.
- c. Wenn er aus besonders süßen Weintrauben gemacht wird.
- d. Wenn der Wein besonders viel Wasser enthält.

6. Wie kommt der Alkohol in den Wein?

- a. Alkohol entsteht bei der Lagerung von fertigem Wein.
- b. Alkohol wird dem Wein vor der Abfüllung zugesetzt.
- c. Alkohol entsteht beim Gären des Traubensaftes.
- d. Alkohol ist schon vor der Weinherstellung vorhanden.

7. Welche folgende Aussage ist falsch?

- a. Alkohol kann aus Obst hergestellt werden.
- b. Alkohol kann aus Reis hergestellt werden.
- c. Alkohol kann aus Fleisch hergestellt werden.
- d. Alkohol kann aus Zuckerrohr hergestellt werden.

8. Welche Aussage ist falsch?

- a. Alkohol ist eine Droge.
- b. Viel Wein zu trinken ist gut für die Gesundheit.
- c. Viel Wein zu trinken beeinflusst die Aktivität des Gehirns.
- d. Viel Wein zu trinken kann der Leber schaden.

Unit 4: Brotbacken

1. Wie wichtig sind die folgenden Dinge , damit ein Hefeteig aufgeht?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Hefe				
Mehl				
Zucker				
Butter				
Wassertemperatur				

2. Wodurch wird Brot luftig?

- a) durch Butter
- b) durch Hefe
- c) durch Zucker
- d) durch Mehl

3. Eine Frau möchte Brot backen, sie hat aber kein Backpulver mehr. Welche der folgenden Zutaten kann sie verwenden, um Backpulver zu ersetzen?

- a. Zucker
- b. Salz
- c. Hefe
- d. Butter

4. Warum darf man beim Anmischen von Hefeteig kein kochendes Wasser verwenden?

- a. Weil die Butter zu schnell flüssig wird.
- b. Weil man das Mehl verbrühen würde.
- c. Weil man die Hefe schädigen würde.
- d. Weil sicher Zucker zu schnell auflösen würde.

5. Was passiert, wenn der Hefeteig „geht“?

- a. Der Hefeteig wird schwerer.
- b. Der Hefeteig wird größer.
- c. Der Hefeteig wird leichter.
- d. Der Hefeteig wird kleiner.

6. Welche Wassertemperatur ist am besten geeignet, um einen Brotteig anzusetzen?

- a. 10°C
- b. 40°C
- c. 70°C
- d. 100°C

7. Welche der folgenden Faktoren beeinflussen das Wachstum der Hefe?

- a. Temperatur
- b. Licht
- c. Mehl
- d. Butter

8. Wofür braucht man keine Hefe?

- a) zum Brotbacken
- b) zum Kochen von Grießbrei
- c) zum Herstellen von Wein
- d) zum Herstellen von Bier

3. Opinion test for pre-test 3

Unit 1: Samenkeimung

1. Wie wichtig sind die folgenden Dinge, damit Samen keimen?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Licht				
Wärme				
Luft				
Erde				
Wasser				

2. Sind die folgenden Aussagen richtig oder falsch? Male in jeder Zeile einen Kreis um richtig oder falsch.

Aussage	richtig oder falsch?
Samen muss man warm stellen, damit sie keimen.	richtig/falsch
Samen muss man ins Helle stellen, damit sie keimen.	richtig/falsch
Samen müssen Luft bekommen, damit sie keimen.	richtig/falsch
Trockene Samen muss man wässern, damit sie keimen.	richtig/falsch
Samen können nur in Erde, nicht aber in einem anderen Material, (wie z.B. Baumwolle) keimen.	richtig/falsch

Unit 2: Hühnereier

1. Wie wichtig sind die folgenden Dinge, damit Hühnereier so schnell wie möglich ausgebrütet werden?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Licht				
Wärme				
feuchte Luft				
Eigröße				
Eifarbe				

2. Sind die folgenden Aussagen richtig oder falsch? Male in jeder Zeile einen Kreis um richtig oder falsch.

Aussage	richtig oder falsch?
Kleine Hühnereier werden genau so schnell ausgebrütet wie große Hühnereier.	richtig/falsch
Hühnereier werden in trockener Luft genau so schnell ausgebrütet wie in feuchter Luft.	richtig/falsch
Hühnereier brauchen Licht, um ausgebrütet werden zu können.	richtig/falsch
Hühnereier brauchen Wärme, damit kleine Küken schlüpfen können.	richtig/falsch

Unit 3: Apfelwein

1. Wie wichtig sind die folgenden Dinge, damit die Weinherstellung gelingt?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Licht				
Wärme				
ein gut verschließbares Gefäß, damit keine Luft hinein kommt				
Zuckermenge				

2. Sind die folgenden Aussagen richtig oder falsch? Male in jeder Zeile einen Kreis um richtig oder falsch.

Aussage	Richtig oder falsch?
Wein enthält viel Alkohol, wenn man zusätzlichen Zucker zum Apfelsaft hinzu gibt.	richtig/falsch
Es hängt von der Temperatur ab, ob die Weinherstellung gelingt.	richtig/falsch
Die Weinherstellung gelingt nicht, wenn man das Gefäß ins Dunkle stellt.	richtig/falsch
Die Weinherstellung gelingt nicht, wenn Luft von außen in das Gefäß kann.	richtig/falsch

Unit 4: Brotbacken

1. Wie wichtig sind die folgenden Dinge, damit ein Hefeteig aufgeht?

	sehr wichtig	wichtig	nicht so wichtig	unwichtig
Hefe				
Mehl				
Zucker				
Butter				
Wassertemperatur				

2. Sind die folgenden Aussagen richtig oder falsch? Male in jeder Zeile einen Kreis um richtig oder falsch.

Welche Aussage ist richtig?	Richtig oder falsch?
Brot wird hart und fest, wenn man beim Backen die Butter vergisst.	richtig/falsch
Brot wird hart und fest, wenn man zum Backen die Hefe vergisst.	richtig/falsch
Brot wird hart und fest, wenn man den Hefeteig mit kochendem Wasser anrührt.	richtig/falsch
Brot wird hart und fest, wenn man beim Backen den Zucker vergisst.	richtig/falsch

Appendix II: Tables and formula

Table 1: Students' answers for each item (3 levels: 0, 1 and 2) in the dimension "search in the hypothesis space" _ Version 1, Cognitive laboratory

Name	Unit 1		Unit 2		Unit 3		Unit 4	Unit 5	
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Kim	0	1	2	1	1	2	2	0	0
Nils	1	0	0	2	2	2	2	2	0
Milena	2	2	1	1	1	1	2	2	2

Table 2: Students' answers for each item (3 levels: 0, 1 and 2) in the dimension "data analysis" _Version 1, Cognitive laboratory

Name	Unit 1		Unit 2		Unit 3		Unit 4	Unit 5	
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Kim	1	1	2	0	2	2	2	1	2
Nils	1	2	2	2	2	2	2	2	2
Milena	2	2	2	2	2	2	2	2	2

Table 3: Students' answers for each item (3 levels: 0, 1 and 2) in the dimension "search in the experiment space" _ Version 1, Cognitive laboratory

Name	Unit 1	Unit 2	Unit 3	Unit 4		Unit 5
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Kim	1	2	1	2	0	0
Nils	0	1	2	2	1	1
Milena	2	1	2	2	2	2

Table 4: Students' answers for each item (3 levels: 0, 1 and 2) in the dimension "search in the hypothesis space" _ Version 2, Cognitive laboratory

Name	Unit 1 (item 1)	Unit 2 (item 2)	Unit 3 (item 3)	Unit 4 (item 4)	Unit 5 (item 5)
Jacob	0	0	2	2	2
Tobias	2	0	2	2	1
Sandra	1	2	0	0	2

Table 5: Students' answers for each item (3 levels: 0, 1 and 2) in the dimension "data analysis" _Version 2, Cognitive laboratory

Name	Unit 1 (item 1)	Unit 2 (item 2)	Unit 3 (item 3)	Unit 4 (item 4)	Unit 5 (item 5)
Jacob	2	2	2	2	2
Tobias	1		2	2	
Sandra	1	2	2	2	2

Table 6: Students' answers for each item (3 levels: 0, 1 and 2) in the dimension "search in the experiment space" _ Version 2, Cognitive laboratory

Name	Unit 1 (item 1)	Unit 2 (item 2 + item 3)		Unit 3 (item 4)	Unit 4 (item 5)	Unit 5 (item 6)
Jacob	0	0	2	1	2	1
Tobias	0	0	1	0	0	2
Sandra	1	0	1	1	2	1

Table 7: Item difficulty for the sub-questions of Unit 1 (Seed germination) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	65.4	71.5	77.0	61.2	82.5	66.8
2	65.9	82.1	79.7	64.7	66.1	50.5
3	46.5	89.1	22.4	92.2	70.6	37.3
4	57.7	77.2	78.8	81.1	64.4	63.0
5	64.5		50.0			
6	62.3					

Table 8: Item difficulty for the sub-questions of Unit 2 (Chicken eggs) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
1	83.0	78.7	68.5	58.3	41.3	72.8	96.4
2	59.2	37.8	23.3	72.6	88.8	57.6	29.2
3	66.6	70.3	56.8	52.1	71.8	74.9	50.0
4	29.0	84.5	59.3	81.7	47.2	86.2	33.7
5	22.1	85.5	54.7		85.8		15.4

Table 9: Item difficulty for the sub-questions of Unit 3 (Apple wine) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	60.3	83.4	63.4	86.9	29.5	57.5
2	87.3	24.1	56.6	55.4	72.4	71.5
3	80.1	58.6	84.8	27.1	36.0	60.4
4	33.3	66.9	82.2	88.7	85.1	80.8
5	96.5			93.2		
6	78.1					

Table 10: Item difficulty for the sub-questions of Unit 4 (Baking bread) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5
1	67.8	88.4	72.3	62.1	88.2
2	91.9	8.3	71.5	83.0	61.0
3	87.4	67.7	35.1	76.3	59.6
4	87.9	33.5	48.7	29.4	81.8
5	56.5	85.2	75.9	96.2	
6		14.0			

Table 11: Item difficulty for the sub-questions of Unit 5 (The growth of bean plants) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5
1	95.3	68.6	74.7	97.0	85.8
2	71.3	80.0	82.2	68.5	54.9
3	42.6	71.2	78.9	93.2	52.8
4	76.0	69.5	62.4	86.0	72.7
5	28.5			69.9	

Table 12: Item difficulty for the sub-questions of Unit 6 (Potatoes) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	92.7	53.0	14.0	31.6	66.1	63.9
2	73.2	60.3	36.4	49.4	58.9	72.1
3	92.9	58.9	62.4	49.7	56.1	65.6
4	83.8	38.4	16.7	15.6	32.3	78.1
5		75.3	71.2			36.6

Table 13: Item difficulty for the sub-questions of Unit 7 (Heart beat) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	88.2	57.3	70.2	91.0	93.0	97.9
2	69.5	71.6	60.8	61.5	97.1	70.6
3	59.3	90.8	78.6	65.5	97.1	96.6
4	77.1	90.1	69.2	64.7	92.8	91.7
5	65.6	66.9	84.6	70.6	95.4	73.3
6			65.5			

Table 14: Item difficulty for the sub-questions of Unit 8 (Plant growth) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5
1	95.0	88.7	54.6	19.5	91.6
2	74.4	84.5	22.7	70.4	75.9
3	77.8	84.9	41.7	50.7	85.5
4	74.2	31.5	34.1	51.4	73.4
5		39.7			

Table 15: Item difficulty for the sub-questions of Unit 9 (Fish respiration) in the knowledge test, Pre-test 1

Sub-question	Item 1	Item 2	Item 3	Item 4	Item 5
1	85.9	54.1	86.9	47.3	60.0
2	92.9	72.0	50.1	65.5	56.3
3	90.9	73.1	58.6	72.7	27.6
4	90.7	82.3	64.1	86.6	78.5

Table 16: Reliability at the booklet level in the knowledge test_ Booklet 111, Pre-test 1

Unit	Item	Corrected item-total correlation (All items)	Corrected item-total correlation (Items selection)
Seed germination	1	0.2648	0.3067
	2	0.3339	0.5368
	3	0.1416	0.1735
	4	0.0607	
	5	0.1322	
	6	0.0722	
Chicken eggs	1	0.1624	
	2	-0.0331	
	3	0.4002	0.3542
	4	0.4656	0.3340
	5	0.2062	
	6	0.4187	0.2819
	7	0.0345	
Apple wine	1	0.3390	0.3810
	2	0.3446	0.3910
	3	0.3497	0.4227
	4	0.1872	0.1590
	5	0.0777	
	6	0.3179	0.3671
Baking bread	1	0.0635	
	2	0.4816	0.5806
	3	0.1046	0.1554
	4	0.3873	0.4933
	5	0.1565	0.1763
Cronbach's alpha		0.6705	0.7389

Table 17: Reliability at the booklet level in the knowledge test_ Booklet 121, Pre-test 1

Unit	Item	Corrected item-total correlation (All items)	Corrected item-total correlation (Items selection)
Seed germination	1	0.5751	0.6065
	2	0.5465	0.5027
	3	0.3710	0.3411
	4	0.0378	
	5	0.0826	
	6	0.0351	
Bean plants	1	0.3728	0.2821
	2	0.1674	0.2242
	3	0.2830	0.3253
	4	-0.1088	
	5	0.4984	0.5257
Potatoes	1	0.2531	0.2769
	2	0.3897	0.2973
	3	0.1560	0.3288
	4	-0.0286	
	5	0.1892	0.2201
	6	0.0356	
Heart beat	1	0.4219	0.3962
	2	0.2921	0.2894
	3	0.3212	0.3194
	4	0.2921	0.4011
	5	0.1295	
	6	0.0561	
Cronbach's alpha		0.6681	0.7555

Table 18: Reliability at the booklet level in the knowledge test_ Booklet 131, Pre-test 1

Unit	Item	Corrected item-total correlation (All items)	Corrected item-total correlation (Items selection)
Seed germination	1	0.2754	0.3098
	2	0.3316	0.3988
	3	0.3945	0.4609
	4	0.2381	0.3159
	5	0.2885	0.3839
	6	-0.0099	
Chicken eggs	1	0.3444	0.3968
	2	0.2823	0.2394
	3	0.0366	
	4	0.1907	0.2356
	5	0.2660	0.3105
	6	0.1918	0.1995
	7	0.1269	
Plant growth	1	0.2505	0.2485
	2	-0.1738	
	3	-0.1410	
	4	0.2181	0.2522
	5	0.1671	0.2531
Fish respiration	1	0.3374	0.3536
	2	0.1434	0.1616
	3	0.0216	
	4	0.0685	
	5	0.1637	
Cronbach's alpha		0.5589	0.6933

Table 19: Reliability at the booklet level in the knowledge test_ Booklet 211, Pre-test 1

Unit	Item	Corrected item-total correlation (All items)	Corrected item-total correlation (Items selection)
Seed germination	1	0.4190	0.4160
	2	0.5374	0.5813
	3	0.2971	0.3380
	4	0.1016	
	5	0.3292	0.3230
	6	0.0965	
Chicken eggs	1	0.3314	0.3685
	2	0.3877	0.3509
	3	0.3449	0.3266
	4	0.1884	0.2046
	5	0.1149	
	6	0.1389	
	7	0.1155	
Apple wine	1	0.0725	
	2	0.1531	
	3	0.5155	0.5089
	4	0.4269	0.4276
	5	0.0646	
	6	0.2934	0.2461
Baking bread	1	0.2911	0.2630
	2	0.2517	0.2737
	3	0.2584	0.2932
	4	0.3332	0.3368
	5	0.2380	0.2632
Bean plants	1	0.0909	
	2	0.2584	0.2884
	3	0.2647	0.2931
	4	0.2184	0.2723
	5	0.3055	0.3321
Cronbach's alpha		0.7392	0.7749

Table 20: Reliability at the booklet level in the knowledge test_ Booklet 221, Pre-test 1

Unit	Item	Corrected item-total correlation (All items)	Corrected item-total correlation (Items selection)
Seed germination	1	0.3898	0.3972
	2	0.4077	0.4194
	3	0.2784	0.2813
	4	0.2254	0.2028
	5	0.2624	0.2837
	6	0.1302	
Potatoes	1	0.4615	0.4909
	2	0.2924	0.3369
	3	0.2084	0.2099
	4	-0.0691	
	5	0.2795	0.3035
	6	0.1893	0.1805
Heart beat	1	0.3753	0.3957
	2	0.3099	0.3154
	3	0.3734	0.4172
	4	0.3840	0.4010
	5	0.4693	0.4840
	6	0.2305	0.2030
Plant growth	1	0.4258	0.4188
	2	0.1887	0.1591
	3	0.0678	
	4	0.3396	0.3365
	5	0.1026	
Heart beat	1	0.4435	0.4729
	2	0.3313	0.3282
	3	0.1931	0.2153
	4	0.4178	0.4226
	5	0.3411	0.3767
Cronbach's alpha		0.7707	0.7990

Table 21: Reliability at the booklet level in the knowledge test_ Booklet 231, Pre-test 1

Unit	Item	Corrected item-total correlation (All items)	Corrected item-total correlation (Items selection)
Chicken eggs	1	0.3448	0.3491
	2	0.2019	0.2090
	3	0.1464	
	4	0.2528	0.2561
	5	0.3925	0.4252
	6	0.2948	0.2561
	7	0.1339	
Apple wine	1	0.1637	0.1536
	2	0.1387	
	3	0.3133	0.2809
	4	0.3003	0.2811
	5	0.1101	
	6	0.1197	
Bean plants	1	0.1847	0.1880
	2	0.2930	0.3002
	3	0.2697	0.2432
	4	0.2370	0.2344
	5	0.2975	0.2915
Heart beat	1	0.4445	0.4788
	2	0.1078	
	3	0.3710	0.3811
	4	0.2545	0.2443
	5	0.3717	0.3699
	6	0.1958	0.2201
Fish respiration	1	0.3911	0.3710
	2	0.2807	0.3023
	3	0.4885	0.4983
	4	0.3904	0.3734
	5	0.3919	0.4033
Cronbach's alpha		0.7584	0.7669

Table 22: Reliability at the booklet level in the knowledge test_ Booklet 241, Pre-test 1

Unit	Item	Corrected item-total correlation (All items)	Corrected item-total correlation (Items selection)
Seed germination	1	0.3413	0.3993
	2	0.3300	0.4410
	3	0.0343	
	4	-0.0226	
	5	0.1627	0.1804
	6	0.0195	
Baking bread	1	0.0792	0.1641
	2	0.3634	0.3851
	3	0.3482	0.3945
	4	0.2094	0.1957
	5	0.0831	
Bean plants	1	0.1140	
	2	-0.0281	
	3	0.0363	
	4	0.1048	
	5	0.3790	0.4409
Potatoes	1	0.3991	0.4328
	2	0.2931	0.4267
	3	0.2262	0.3153
	4	-0.0756	
	5	0.2263	0.3709
	6	0.1751	0.1854
Plant growth	1	0.1474	0.2014
	2	0.1982	0.2532
	3	-0.1708	
	4	-0.0026	
	5	0.1109	0.1590
Cronbach's alpha		0.5351	0.7119

Table 23: Reliability for the scales “forming hypothesis”, “data analysis” & “planning experiment” at the booklet level in the competency test: Corrected item-total correlation and Cronbach’s alphas, Pre-test 1

Booklet	Units	Item	Search in the hypothesis space		Data analysis		Search in the experiment space	
			CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
111	Seed germination	1	0.4595		0.3768		0.3931	
	Chicken eggs	2	0.3617		0.1890		0.3277	
	Apple wine	1	0.1478		0.5932		0.4733	
	Baking bread	2	0.5766	0.7543	0.4339	0.7328	0.2749	0.6788
	Bean plants	1	0.4708		0.4303		0.3645	
	Potatoes	2	0.6894		0.4986		0.4389	
	Heart beat	1	0.4620		0.5892		0.4088	
	Fish respiration	2	0.4762		0.3182		0.2770	
121	Seed germination	1	0.4361		0.4873		0.0465	
	Chicken eggs	2	0.4482		0.4770		0.1827	
	Apple wine	1	0.4875		0.3900		0.3753	
	Baking bread	2	0.5813	0.6176	0.4175	0.7082	0.2606	0.5102
	Bean plants	1	0.3755		0.3286		0.4048	
	Potatoes	2	0.0508		0.3914		0.2303	
	Heart beat	1	0.3110		0.2767		0.1547	
	Fish respiration	2	0.2290		0.4184		0.2441	
131	Seed germination	1	0.4133		0.3881		0.2878	
	Chicken eggs	2	0.5125		0.4092		0.3417	
	Apple wine	1	0.2986		0.3818		0.4735	
	Baking bread	2	0.4363	0.6977	0.4001	0.7385	0.3786	0.6745
	Bean plants	1	0.3953		0.5264		0.4262	
	Potatoes	2	0.4281		0.4847		0.3786	
	Heart beat	1	0.3011		0.4627		0.2234	
	Fish respiration	2	0.3245		0.3954		0.4262	
211	Seed germination	-	0.0765		0.1426		0.0171	
	Chicken eggs		0.2113	0.3594	0.0849	0.2934	0.2721	0.2896
	Apple wine		0.2927		0.1917		0.1968	
	Baking bread		0.2292		0.0869		0.1362	
	Bean plants		0.2503		0.1898		0.0811	
221	Seed germination		0.1148		0.3365		0.2374	
	Potatoes		0.1165		0.3390		0.2393	
	Heart beat		0.2295	0.3647	0.2798	0.5377	0.3212	0.5188
	Plant growth		0.1213		0.2768		0.3413	
	Fish respiration		0.2940		0.2769		0.2968	
231	Chicken eggs		0.3408		0.3288		0.3114	
	Apple wine		0.3458		0.3608		0.2835	
	Bean plants		0.3282	0.5770	0.2278	0.5437	0.1262	0.5042
	Heart beat		0.3458		0.3516		0.3222	
	Fish respiration		0.3117		0.2666		0.3322	
241	Seed germination		0.1750		0.3882		0.2021	
	Baking bread		0.3033		0.3544		0.2724	
	Bean plants		0.3386	0.4884	0.2651	0.5817	0.3098	0.4399
	Potatoes		0.2930		0.3653		0.2898	
	Plant growth		0.2095		.3239		.0881	

Table 24: Reliability for the scales “forming hypothesis”, “data analysis” and “planning experiment” at the booklet level: After taking some low items out, Pre-test 1

Booklet	Unit	Item	Search in the hypothesis space		Data analysis		Search in the experiment space	
			CITC	Cronbach's alpha	CITC	Cronbach's alpha	CITC	Cronbach's alpha
111	Seed germination	1	0.3501		0.3338		0.3931	
		2	0.3776				0.3277	
	Chicken eggs	1			0.5934		0.4733	
		2	0.5281	0.7818	0.4286	0.7484	0.2749	0.6788
	Apple wine	1	0.4807		0.4499		0.3645	
		2	0.7083		0.4684		0.4389	
	Baking bread	1	0.5153		0.6284		0.4088	
		2	0.4971		0.3595		0.2770	
121	Seed germination	1	0.4013		0.4873			
		2	0.4856		0.4770			
	Bean plants	1	0.4811		0.3900		0.4656	
		2	0.5726	0.7116	0.4175	0.7082	0.1502	0.5619
	Potatoes	1	0.3485		0.3286		0.4195	
		2			0.3914		0.2665	
	Heart beat	1	0.3474		0.2767			
		2	0.3128		0.4184		0.3291	
131	Seed germination	1	0.4133		0.3881		0.2878	
		2	0.5125		0.4092		0.3417	
	Chicken eggs	1	0.2986		0.3818		0.4735	
		2	0.4363	0.6977	0.4001	0.7385	0.3786	0.6745
	Plant growth	1	0.3953		0.5264		0.4262	
		2	0.4281		0.4847		0.3786	
	Fish respiration	1	0.3011		0.4627		0.2234	
		2	0.3245		0.3954		0.4262	
211	Seed germination							
	Chicken eggs		0.2787				0.1889	
	Apple wine		0.3151	0.4886	0.2054	0.3408	0.1889	0.3172
	Baking bread		0.2616					
	Bean plants		0.2803		0.2054			
221	Seed germination				0.3365		0.2374	
	Potatoes				0.3390		0.2393	
	Heart beat		0.2025	0.3366	0.2798	0.5377	0.3212	0.5188
	Plant growth				0.2768		0.3413	
	Fish respiration		0.2025		0.2769		0.2968	
231	Chicken eggs		0.3408		0.3288		0.3567	
	Apple wine		0.3458		0.3608		0.3229	
	Bean plants		0.3282	0.5770	0.2278	0.5437		0.5297
	Heart beat		0.3458		0.3516		0.2603	
	Fish respiration		0.3117		0.2666		0.3308	
241	Seed germination				0.3882		0.2316	
	Baking bread		0.2476		0.3544		0.3369	
	Bean plants		0.3876	0.4911	0.2651	0.5817	0.3057	0.4821
	Potatoes		0.2467		0.3653		0.2456	
	Plant growth		.2667		.3239			

Table 25: Reliability of the booklet (all items of the four units combined): Corrected item-total correlation and Cronbach's alphas. Booklet 111, Pre-test 1

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H1	0.6245	Apple wine	H1	0.5037
	H2	0.3981		H2	0.5948
	D1	0.4578		D1	0.5104
	D2	0.2269		D2	0.5167
	E1	0.4498		E1	0.5397
	E2	0.4459		E2	0.5331
Chicken eggs	H1	0.0510	Baking bread	H1	0.6619
	H2	0.5713		H2	0.5167
	D1	0.6776		D1	0.6961
	D2	0.5037		D2	0.4823
	E1	0.4751		E1	0.4652
	E2	0.3228		E2	0.2816

Cronbach's alpha = 0.8885

Table 26: Reliability of the booklet (all items of the four units combined): Corrected item-total correlation and Cronbach's alphas. Booklet 121, Pre-test 1

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H1	0.4525	Potatoes	H1	0.4184
	H2	0.5612		H2	0.1081
	D1	0.5359		D1	0.3822
	D2	0.5790		D2	0.3978
	E1	0.2069		E1	0.4828
	E2	0.1243		E2	0.2129
Bean plants	H1	0.3978	Heart beat	H1	0.4223
	H2	0.6058		H2	0.3296
	D1	0.3986		D1	0.3282
	D2	0.3396		D2	0.4737
	E1	0.3837		E1	0.4222
	E2	0.4141		E2	0.2096

Cronbach's alpha = 0.8366

Table 27: Reliability of the booklet (all items of the four units combined): Corrected item-total correlation and Cronbach's alphas. Booklet 131, Pre-test 1

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H1	0.5505	Plant growth	H1	0.4427
	H2	0.4726		H2	0.4760
	D1	0.4296		D1	0.5312
	D2	0.5150		D2	0.5001
	E1	0.1748		E1	0.3921
	E2	0.2288		E2	0.4714
Chicken eggs	H1	0.3361	Fish respiration	H1	0.3415
	H2	0.4857		H2	0.3258
	D1	0.4606		D1	0.4608
	D2	0.3483		D2	0.5332
	E1	0.4199		E1	0.3102
	E2	0.3593		E2	0.2749

Cronbach's alpha = 0.8549

Table 28: Reliability of the booklet (all items of the five units combined): Corrected item-total correlation and Cronbach's alphas. Booklet 211, Pre-test 1

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H	0.0477	Baking bread	H	0.3078
	D	0.2831		D	0.1464
	E	0.0455		E	0.1493
Chicken eggs	H	0.2427	Bean plants	H	0.2927
	D	0.1025		D	0.2987
	E	0.2478		E	0.1657
Apple wine	H	0.4408			
	D	0.2949			
	E	0.3069			

Cronbach's alpha = 0.5887

Table 29: Reliability of the booklet (all items of the five units combined): Corrected item-total correlation and Cronbach's alphas. Booklet 221, Pre-test 1

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H	-0.0014	Plant growth	H	0.3709
	D	0.2987		D	0.2599
	E	0.0943		E	0.4101
Potatoes	H	0.1105	Fish respiration	H	0.3513
	D	0.3432		D	0.4091
	E	0.2327		E	0.2693
Heart beat	H	0.1780			
	D	0.2599			
	E	0.3844			

Cronbach's alpha = 0.6522

Table 30: Reliability of the booklet (all items of the five units combined): Corrected item-total correlation and Cronbach's alphas. Booklet 231, Pre-test 1

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Chicken eggs	H	0.2897	Heart beat	H	0.2234
	D	0.2979		D	0.4694
	E	0.3045		E	0.3504
Apple wine	H	0.3492	Fish respiration	H	0.3909
	D	0.4033		D	0.4177
	E	0.3266		E	0.4286
Bean plants	H	0.3833			
	D	0.1968			
	E	0.2674			

Cronbach's alpha = 0.7392

Table 31: Reliability of the booklet (all items of the five units combined): Corrected item-total correlation and Cronbach's alphas. Booklet 241, Pre-test 1

Unit	Item	Corrected item-total correlation	Unit	Item	Corrected item-total correlation
Seed germination	H	0.1837	Potatoes	H	0.1796
	D	0.4484		D	0.4573
	E	0.1608		E	0.1997
Baking bread	H	0.4553	Plant growth	H	0.3097
	D	0.4305		D	0.3207
	E	0.3775		E	0.3910
Bean plants	H	0.4009			
	D	0.3052			
	E	0.2647			

Cronbach's alpha = 0.7237

Table 32: Reliability at the unit level in the knowledge test: Corrected item-total correlation and Cronbach's alphas. Unit 1: Seed germination, Pre-test 3

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (item selection)	Cronbach's alpha (item selection)
1	-0.0112	0.1756		0.4488
2	-0.0360			
3	-0.0245			
4	0.1919		0.2309	
5	0.0109			
6	0.0384		0.2429	
7	0.4278		0.3829	
8	-0.0360			
9	0.0308		0.1857	

Table 33: Reliability at the unit level in the knowledge test: Corrected item-total correlation and Cronbach's alphas. Unit 2: Chicken eggs, Pre-test 3

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (item selection)	Cronbach's alpha (item selection)
1	0.0814			
2	0.1724		0.2088	
3	0.2626		0.2376	
4	0.0182			
5	0.1832	0.4185		0.5024
6	0.3226		0.2543	
7	-0.0226			
8	0.3495		0.2902	
9	0.1298			
10	0.0175			
11	0.1770		0.2930	
12	0.1852		0.2610	

Table 34: Reliability at the unit level in the knowledge test: Corrected item-total correlation and Cronbach's alphas. Unit 3: Apple wine, Pre-test 3

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (item selection)	Cronbach's alpha (item selection)
1	-0.0503			
2	0.2224		0.4170	
3	0.3330		0.4036	
4	0.1499	0.4240	0.2049	0.5491
5	-0.2591			
6	0.2418		0.2922	
7	0.4259		0.3160	
8	0.1654			
9	0.0970			
10	0.0209			
11	0.2528		0.1534	

Table 35: Reliability at the unit level in the knowledge test: Corrected item-total correlation and Cronbach's alphas. Unit 4: Baking bread, Pre-test 3

Item	Corrected item-total correlation (all items)	Cronbach's alpha (all items)	Corrected item-total correlation (item selection)	Cronbach's alpha (item selection)
1	-0.0207			
2	0.1047		0.1692	
3	0.2516		0.3268	
4	-0.0387			
5	0.0279	0.1948		0.4541
6	-0.1565			
7	-0.0826			
8	0.0301		0.2248	
9	-0.0083			
10	0.2741		0.2988	
11	0.2400		0.1902	

Table 36: Reliability at the booklet level in the knowledge test: Corrected item-total correlation and Cronbach's alphas, Pre-test 3

Unit	Items	Corrected item-total correlation (all items)	Corrected item-total correlation (item selection)
Seed germination	1	0.3300	0.2534
	2	0.0051	
	3	0.1587	
	4	0.1622	0.1633
	5	0.1022	
	6	-0.2327	
	7	0.2290	0.1783
	8	0.2419	0.3765
	9	-0.0143	
Chicken eggs	1	-0.0412	
	2	0.1498	0.2277
	3	0.1678	
	4	0.0217	
	5	0.1686	
	6	0.3800	0.4493
	7	0.0860	
	8	0.4299	0.4003
	9	0.3349	0.2883
	10	0.1093	
	11	0.0317	
	12	0.1579	0.3147
Apple wine	1	-0.1300	
	2	0.2609	0.3781
	3	0.4024	0.5037
	4	0.1739	0.3441
	5	-0.1944	
	6	0.4024	0.3818
	7	0.3934	0.2893
	8	0.1692	0.2344
	9	-0.1015	
	10	0.0068	
	11	0.1272	
Baking bread	1	0.1306	0.2368
	2	0.0565	
	3	0.2880	0.4728
	4	0.0565	
	5	-0.1580	
	6	0.0922	
	7	-0.1422	
	8	0.3520	0.4221
	9	0.1265	
	10	0.0577	
	11	0.2091	0.3289
Cronbach's alpha for booklet		0.5796	0.7605

Table 37: Item difficulty for the items in the knowledge test (in grade 5), Main test

Question	Seed germination	Chicken eggs	Apple wine	Baking bread
2	57.8	84.3	46.0	60.3
3	75.4	40.7	36.4	70.7
4	39.2	80.2	44.4	46.7
5	29.9	57.6	8.1	77.5
6	30.8	40.1	54.3	40.5
7	8.6	58.6	43.6	54.3
8		15.0	48.0	40.3

Table 38: Item difficulty for the items in the knowledge test (in grade 6), Main test

Question	Seed germination	Chicken eggs	Apple wine	Baking bread
2	66.2	79.9	48.2	73.6
3	79.4	48.3	52.7	65.9
4	36.7	85.7	55.0	59.0
5	35.0	67.1	11.1	87.0
6	42.7	42.7	66.3	41.5
7	13.0	60.4	44.9	55.2
8		20.9	58.9	50.4

Table 39: Item difficulty for the items in the competency test (in grade 5), Main test

Unit	Search in the hypothesis space		Data analysis		Search in the experiment space	
	H1	H2	D1	D2	E1	E2
Seed germination	44.0	45.1	46.7	65.0	67.4	43.6
Chicken eggs	52.6	52.9	42.0	39.8	52.5	39.2
Apple wine	50.5	55.7	74.8	60.0	44.5	37.1
Baking bread	63.5	69.1	70.3	61.7	57.1	48.1

Table 40: Item difficulty for the items in the competency test (in grade 6), Main test

Unit	Search in the hypothesis space		Data analysis		Search in the experiment space	
	H1	H2	D1	D2	E1	E2
Seed germination	57.7	60.7	59.1	77.6	70.4	59.9
Chicken eggs	64.4	69.3	47.6	48.2	65.4	49.1
Apple wine	63.3	70.8	78.2	77.0	58.3	54.6
Baking bread	78.1	80.4	74.1	71.6	69.2	60.5

Table 41: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total correlation. Unit 1: Seed germination (Students in grade 5), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	0.1103		0.2011	
3	0.1144		0.2011	
4	-0.1376	0.0827		0.3323
5	-0.0595			
6	0.1368			
7	0.0991			

Table 42: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total correlation. Unit 1: Seed germination (Students in grade 6), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	0.1287			
3	0.1006			
4	0.2341	0.3750	0.2407	0.3844
5	0.1501		0.1500	
6	0.2768		0.2619	
7	0.1584		0.1899	

Table 43: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total correlation. Unit 2: Chicken eggs (Students in grade 5), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	0.0924			
3	0.1067			
4	0.1404		0.2529	
5	0.1981	0.2390	0.2181	0.3586
6	0.0516			
7	0.1124		0.1614	
8	-0.0456			

Table 44: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total correlation in Unit 2: Chicken eggs (Students in grade 6), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	-0.0825			
3	0.0957		0.1319	
4	0.1778		0.1815	
5	0.0485	0.1798		0.2953
6	0.0553			
7	0.1137		0.1971	
8	0.0967		0.1049	

Table 45: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total-correlation. Unit 3: Apple wine (Students in grade 5), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	0.2477		0.3184	
3	0.2328		0.2016	
4	0.2743		0.2748	
5	-0.0870	0.3268		0.4561
6	0.2215		0.2426	
7	-0.0960			
8	0.1433			

Table 46: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total correlation. Unit 3: Apple wine (Students in grade 6), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	0.2365		0.2584	
3	0.2362		0.2754	
4	0.2857		0.2962	
5	0.1054	0.4447		0.4868
6	0.2895		0.3007	
7	0.1619			
8	0.1144			

Table 47: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total correlation. Unit 4: Baking bread (Students in grade 5), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	0.2460		0.2667	
3	0.0503			
4	0.0282		0.1961	
5	0.0188	0.1807		0.3601
6	-0.1067			
7	0.1256			
8	0.1496		0.1635	

Table 48: Reliability at the unit level in the knowledge test – Cronbach’s alpha and corrected item-total-correlation. Unit 4: Baking bread (Students in grade 6), Main test

Item	Corrected item-total correlation (all items)	Cronbach’s alpha (all items)	Corrected item-total correlation (Item selection)	Cronbach’s alpha (Item selection)
2	0.1428			
3	0.2137		0.2216	
4	0.2389		0.2325	
5	0.2207	0.3844	0.2090	0.4126
6	0.0215			
7	0.1580		0.1649	
8	0.2232		0.2392	

Table 49: Reliability at the booklet level in the knowledge test – Corrected item-total correlation (CITC). Main test

Units	Item	All students		Students in grade 5 th		Students in grade 6 th	
		CITC (All items)	CITC (Item selection)	CITC (All items)	CITC (Item selection)	CITC (All items)	CITC (Item selection)
Seed germination	2	0.2570	0.2099	0.3332	0.2972	0.2270	0.1907
	3	0.2637	0.3107	0.3701	0.3017	0.2206	0.3004
	4	0.0110		-0.1482		0.0824	
	5	0.0365		-0.1409		0.0855	
	6	0.1394		-0.0905		0.1813	
	7	0.1036		0.0169		0.1181	
	Chicken eggs	2	-0.0687		0.0037		-0.0795
3		0.1649		0.0828		0.1679	
4		0.2015	0.2803	0.2575	0.2773	0.1734	0.2635
5		0.1224		0.2026		0.0738	
6		0.0117		0.0131		0.0058	
7		0.2058	0.2005	0.0729		0.2465	0.2324
8		0.1687	0.1606	-0.1020		0.2443	0.2480
Apple wine		2	0.1967	0.2219	0.2059	0.2638	0.1947
	3	0.3278	0.3649	0.2150	0.2610	0.3293	0.3783
	4	0.2977	0.3539	0.2718	0.2824	0.2967	0.3825
	5	0.0567		-0.1728		0.1220	
	6	0.2674	0.2791	0.2098	0.3060	0.2642	0.2827
	7	0.0874		0.0177		0.1031	
	8	0.1833		0.2113	0.2173	0.1552	
	Baking bread	2	0.1723	0.2068	0.3565	0.2877	0.0808
3		0.1197	0.1650	-0.0119		0.1838	0.2290
4		0.2106	0.2718	-0.0074		0.2821	0.3036
5		0.1471	0.2399	0.0979		0.1353	0.2514
6		0.0388		-0.0194		0.0639	
7		0.1502	0.1733	0.1632	0.1939	0.1544	0.1908
8		0.2872	0.3381	0.2467	0.2865	0.2722	0.3396

Table 50: Reliability at the booklet level in the knowledge test – Cronbach's alpha. Main test

	All students		Students in grade 5 th		Students in grade 6 th	
	All items	Item selection	All items	Item selection	All items	Item selection
Cronbach's alpha	0.5534	0.6307	0.4115	0.6083	0.5704	0.6471

Table 51: Reliability at the unit level in the competency test: Corrected item-total correlation and Cronbach's alphas (In grade 5). Main test

Item	Seed germination	Chicken eggs	Apple wine	Baking bread
H1	0.3698	0.2543	0.5226	0.4845
H2	0.3972	0.3207	0.4835	0.4018
D1	0.3422	0.3243	0.4582	0.3529
D2	0.2903	0.1698	0.5645	0.4546
E1	0.2732	0.2196	0.3942	0.1971
E2	0.1487	0.2387	0.2190	0.1733
Cronbach's alpha for unit	0.5617	0.4983	0.7058	0.6050

Table 52: Reliability at the unit level in the competency test: Corrected item-total correlation and Cronbach's alphas (In grade 6). Main test

Item	Seed germination	Chicken eggs	Apple wine	Baking bread
H1	0.5665	0.4430	0.6066	0.4933
H2	0.4835	0.4327	0.5804	0.5594
D1	0.5378	0.3899	0.5350	0.4968
D2	0.3639	0.3371	0.6349	0.5119
E1	0.3936	0.3322	0.4754	0.4115
E2	0.3224	0.3725	0.3981	0.3721
Cronbach's alpha for unit	0.7118	0.6545	0.7841	0.7339

Table 53: Reliability for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and Cronbach's alpha (In grade 5). Main test

Units	Item	Scale “forming hypothesis”		Scale “data analysis”		Scale “planning experiment”	
		CITC	Cronbach's alpha	CITC	Cronbach's alpha	CITC	Cronbach's alpha
Seed germination	1	0.4166	0.7554	0.2227	0.6303	0.2837	0.5590
	2	0.3854		0.2677		0.2286	
Chicken eggs	1	0.3454	0.7554	0.2881	0.6303	0.3234	0.5590
	2	0.4624		0.2122		0.2015	
Apple wine	1	0.5125	0.7554	0.4302	0.6303	0.2738	0.5590
	2	0.4759		0.4691		0.2251	
Baking bread	1	0.4753	0.7554	0.2960	0.6303	0.2598	0.5590
	2	0.5476		0.4403		0.3540	

Table 54: Reliability for the scales “forming hypotheses”, “data analysis” and “planning experiments” at the booklet level: Corrected item-total correlation (CITC) and Cronbach's alpha (In grade 6). Main test

Units	Item	Scale “forming hypothesis”		Scale “data analysis”		Scale “planning experiment”	
		CITC	Cronbach’s alpha	CITC	Cronbach’s alpha	CITC	Cronbach’s alpha
Seed germination	1	0.4644		0.3646		0.3558	
	2	0.4431		0.3332		0.4480	
Chicken eggs	1	0.3591	0.7793	0.3178	0.6971	0.4231	0.7372
	2	0.4965		0.2512		0.4704	
Apple wine	1	0.5730		0.4927		0.4671	
	2	0.5011		0.4847		0.4424	
Baking bread	1	0.5141		0.4745		0.3808	
	2	0.5279		0.4311		0.4464	

Table 55: Reliability of the booklet (all items of the four units combined): Corrected item-total correlation. Main test

Unit	Item	Corrected item-total correlation (all students)	Corrected item-total correlation (students in grade 5)	Corrected item-total correlation (students in grade 6)
Seed germination	H1	0.5149	0.5266	0.4946
	H2	0.4775	0.4362	0.4677
	D1	0.4306	0.3359	0.4488
	D2	0.4085	0.3713	0.3998
	E1	0.3653	0.2189	0.4132
	E2	0.3861	0.1581	0.4360
Chicken eggs	H1	0.4184	0.3022	0.4428
	H2	0.5636	0.4725	0.5747
	D1	0.3684	0.3293	0.3766
	D2	0.2888	0.1799	0.3135
	E1	0.4238	0.3203	0.4343
	E2	0.4053	0.2387	0.4577
Apple wine	H1	0.5828	0.5290	0.5827
	H2	0.5482	0.5212	0.5372
	D1	0.5184	0.4731	0.5317
	D2	0.5878	0.5090	0.5971
	E1	0.5201	0.4495	0.5255
	E2	0.4584	0.3190	0.4675
Baking bread	H1	0.5236	0.4782	0.5183
	H2	0.5401	0.5121	0.5364
	D1	0.4703	0.3549	0.5071
	D2	0.5233	0.5425	0.5132
	E1	0.3799	0.2356	0.4121
	E2	0.4246	0.3515	0.4329

Table 56: Reliability of the booklet (all items of the four units combined): Coefficient Cronbach’s alphas, Main test

	All students	Students in grade 5	Students in grade 6
Cronbach's alpha	0.8840	0.8350	0.8894

Table 57: Value of the probable tested models in the dimension “Search in the hypothesis space”, Main test

Model	BIC-Index	CAIC-Index
Classified model with two latent classes (LCA 2)	10682.54	10717.54
Classified model with three latent classes (LCA 3)	10746.37	10799.37
Classified model with four latent classes (LCA 4)	11066.06	11137.06

Table 58: Value of the probable tested models in the dimension “Data analysis”, Main test

Model	BIC-Index	CAIC-Index
Classified model with two latent classes (LCA 2)	10382.50	10417.50
Classified model with three latent classes (LCA 3)	10462.24	10515.24
Classified model with four latent classes (LCA 4)	10538.53	10609.53

Table 59: Value of the probable tested models in the dimension “Search in the experiment space”. Main test

Model	BIC-Index	CAIC-Index
Classified model with two latent classes (LCA 2)	11056.22	11091.22
Classified model with three latent classes (LCA 3)	11137.16	11190.16
Classified model with four latent classes (LCA 4)	11234.10	11305.10

Table 60: Cross table between the students' pre-knowledge and "Search in the hypothesis space", Main test

			Search in the hypothesis space			
			1	2	3	Total
Pre-knowledge	1	Number	83	14	38	135
		% of Knowledge	61.5%	10.4%	28.1%	100.0%
		% of Hypothesis	25.9%	10.3%	8.8%	15.2%
		% of Total	9.3%	1.6%	4.3%	15.2%
	2	Number	201	84	240	525
		% of Knowledge	38.3%	16.0%	45.7%	100.0%
		% of Hypothesis	62.6%	61.8%	55.4%	59.0%
		% of Total	22.6%	9.4%	27.0%	59.0%
	3	Number	37	38	155	230
		% of Knowledge	16.1%	16.5%	67.4%	100.0%
		% of Hypothesis	11.5%	27.9%	35.8%	25.8%
		% of Total	4.2%	4.3%	17.4%	25.8%
Total	Number	321	136	433	890	
	% of Knowledge	36.1%	15.3%	48.7%	100.0%	
	% of Hypothesis	100.0%	100.0%	100.0%	100.0%	
	% of Total	36.1%	15.3%	48.7%	100.0%	

Table 61: Cross table between the students' pre-knowledge and "Data analysis". Main test

			Data analysis			
			1	2	3	Total
Pre-knowledge	1	Number	70	36	31	137
		% of Knowledge	51.1%	26.3%	22.6%	100.0%
		% of Data analysis	30.4%	14.6%	7.9%	15.7%
		% of Total	8.0%	4.1%	3.6%	15.7%
	2	Number	140	159	207	506
		% of Knowledge	27.7%	31.4%	40.9%	100.0%
		% of Data analysis	60.9%	64.4%	52.5%	58.1%
		% of Total	16.1%	18.3%	23.8%	58.1%
	3	Number	20	52	156	228
		% of Knowledge	8.8%	22.8%	68.4%	100.0%
		% of Data analysis	8.7%	21.1%	39.6%	26.2%
		% of Total	2.3%	6.0%	17.9%	26.2%
Total	Number	230	247	394	871	
	% of Knowledge	26.4%	28.4%	45.2%	100.0%	
	% of Data analysis	100.0%	100.0%	100.0%	100.0%	
	% of Total	26.4%	28.4%	45.2%	100.0%	

Table 62: Cross table between the students' pre-knowledge and "Search in the experiment space". Main test

			Search in the experiment space			
			1	2	3	Total
Pre-knowledge	1	Number	66	37	20	123
		% of Knowledge	53.7%	30.1%	16.3%	100.0%
		% of Experiment	22.1%	21.1%	6.3%	15.6%
		% of Total	8.4%	4.7%	2.5%	15.6%
	2	Number	186	105	174	465
		% of Knowledge	40.0%	22.6%	37.4%	100.0%
		% of Experiment	62.4%	60.0%	55.1%	58.9%
		% of Total	23.6%	13.3%	22.1%	58.9%
	3	Number	46	33	122	201
		% of Knowledge	22.9%	16.4%	60.7%	100.0%
		% of Experiment	15.4%	18.9%	38.6%	25.5%
		% of Total	5.8%	4.2%	15.5%	25.5%
Total	Number	298	175	316	789	
	% of Knowledge	37.8%	22.2%	40.1%	100.0%	
	% of Experiment	100.0%	100.0%	100.0%	100.0%	
	% of Total	37.8%	22.2%	40.1%	100.0%	

$$r'_{tt} = \frac{\frac{n'}{n} \cdot r_{tt}}{1 + \left(\frac{n'}{n} - 1\right) \cdot r_{tt}}$$

r'_{tt} = Reliability of changed test

r_{tt} = Reliability of the original test

n = Number of items in original test

n' = Number of the items in the changed test

Spearman- Brown formula

Literatures

- Adi, H., Karplus, R. Lawson, S. & Pulo, R. (1978). Intellectual development beyond elementary school: VI. Correlational reasoning. *School Science and Mathematics* 78: 675-683.
- Antonietti, A., Ignazi, S., and Perego, P. (2000). Metacognitive knowledge about problem-solving methods. *British Journal of Educational Psychology* 70: 1-16.
- Aznar, M.M. (2005). Solving problems in genetics. *International Journal of Science Education* 27: 101-121.
- Baumert, J. at al. (1998). Testaufgaben Naturwissenschaften TIMSS 7./8. Klasse. Berlin: Max-Planck-Institut für Bildungsforschung.
- Baumert, J., Lehmann, R. (1997). TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich – Opladen: Leske+ Budrich, 86.
- Bayrhuber, H., Brinkman, F. (Eds.). (1998). What-Why-How? Research in Didaktik of Biology. *Proceedings of the First Conference of European Researchers in Didaktik of Biology* (ERIDOB).
- Beishuizen, J.J. Hof, E., van Putten, C.M., Bouwmeester, S. and Asscher, J.J. (2001). Students' and teachers' cognitions about good teachers. *British Journal of Educational Psychology* 71: 185-201.
- Bell, R.L., Blair, L.M., Crawford, B.A., Lederman, N.G. (2003). Just Do It? Impact of a Science Apprenticeship Program on High School Students' Understanding of the Nature of Science and Scientific Inquiry. *Journal of research in science teaching* 40: 487-509.
- Bourne, F.C., Jr., and Restle, F. (1959). Mathematical theory of concept identification. *Psychological review* 66: 278-296.
- Brown, A.L. (1990). Domain-specific principles affect learning and transfer in children. *Cognitive Science* 14: 107-133.
- Bruner, J.S., Goodnow, J.J., Austin, G.A. (1956). *A study of thinking*. New York: NY Science Editions.
- Bybee, R.W. (1997). *Achieving Scientific Literacy: From Promise to Practice*. Portsmouth: Heinemann
- Bybee, R.W. (2002). Scientific Literacy – Mythos oder Realität? In: W. Gräber, P. Nentwig, T. Koballa, and R. Evans (Hg.): *Scientific Literacy*. Opladen: Leske + Budrich, 21-43.
- Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, Mass. : Bradford Books/ MIT Press.
- Carey, S. (1985a). Are Children Fundamentally Different Kinds of Thinker Than Adults? – In: J. Segal, R. Glaser (Hg.): *Thinking and Learning Skills 2 Research and Open Questions*. Hilldale, NJ: Erlbaum, 485-517.

- Carey, S., Evans, R., Honda, M., Jay, E., Unger, C. (1989). An Experiment Is When You Try It and See If It Works: A Study of Grade 7 Students' Understanding of the Construction of Scientific Knowledge. *Int. Journal of Science Education* 11: 514–529.
- Champagne, A.B., Gunstone, R.F., Klopfer, L.E. (1985). Instructional consequences of students' knowledge about physical phenomena. In : L. H. T. West, A. L. Pines (Hg.): *Cognitive structure and conceptual change*. Orlando: Academic.
- Champagne, A.B., Klopfer, L.E., & Chaiklin, S.D. (1984). *Structure of naive knowledge : A quilt or collection of patches*. Paper presented at the Annual Meeting of the American Association for the Advancement of Science, New York.
- Chen, Z., Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of variables Strategy. *Child Development* 70: 1098-1120.
- Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science* 5: 121-152.
- Chinn, C., Brewer, W. (1998). An Empirical Test of a Taxonomy of responses to Anomalous Data in Science – *Journal of Research in Science Teaching* 35 (6): 623-654.
- Dawes, R.M. (2001). *Everyday irrationality*. Boulder: West view Press.
- Demie, F. (2002). Pupil mobility and educational achievement in schools: an empirical analysis. *Educational Research* 44: 197-215.
- Deutsches PISA-KONSORTIUM (Hg.). (2001). *Pisa 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich – Opladen: Leske + Budrich*.
- Dunbar, K. (1993). Concept Discovery in a Scientific Domain. *Cognitive Science* 17: 397-434.
- Dunbar, K. (1997). How scientists think: on-line creativity and conceptual change in science. In T. Ward, S. Smith, and S. Vaid, eds., *Conceptual Structures and Process: Emergence, Discovery and Change*, p.p. 461-492. Washington, D.C.: APA Press.
- Dunbar, K., D. Klahr (1989). Developmental differences in scientific discovery strategies. In: D. Klahr und K. Kotovsky, eds., *Complete Information Processing: The impact of Herbert A. Simon*. Hillsdale, N.J. Erlbaum, 109-143.
- Fay, A., Klahr, D. (1996) Knowing about Guessing and guessing about Knowing: Preschoolers' Understanding of Indeterminacy. *Child Development* 67: 689-716.
- Franklin, B. (1931). Depth of water and speed of boats. In N.G. Goodman (Ed.), *Selected scientific letters of Benjamin Franklin* (pp.136-139). Philadelphia: University of Pennsylvania Press.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7: 155-170.

- Hammann, M. (2004). Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung dargestellt anhand von Kompetenzen beim Experimentieren. *MNU* 57/4, 196-203.
- Hammann, M., Phan, T. T. H., Ehmer, M., Bayrhuber, H. (2006). Fehlerfrei Experimentieren. *MNU* 59/5, 292-299.
- Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic books.
- Jegede, O. and Taplin, M. (2000). Trainee teachers' perception of their knowledge about expert teaching. *Educational Research* 42: 287-308.
- Kaplan, C.A., and Simon, H.A. (1990). In search of insight. *Cognitive Psychology* 22: 374-419.
- Kaplan, R., Kaplan, E., Formisano, M., Paulsen, A. (1979). Proportional Reasoning and Control of Variables in seven countries. In: J. Lochhead – J. Clement (Hg.): *Cognitive process instruction: Research on teaching thinking skills*. Philadelphia: Franklin Institute Press.
- Kasanda, C., Lubben, F., Gaoseb, N., Kandjeo-Marenga, U., Kapenda, H., and Cambell, B. (2005). The Role of Everyday Context in Learner-centred Teaching: The practice in Namibian secondary schools. *International Journal of Science Education* 27: 1805-1823.
- Klahr, D. (2000). *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, Mass., London: MIT Press.
- Klahr, D., Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science* 12: 1-48.
- Klahr, D., Fay, A. L., Dunbar, K. (1993). Heuristics for Scientific Experimentation: A Developmental Study. *Cognitive Psychology* 25: 111-146.
- Klahr, D., Simon, H. A. (1999). Studies of Scientific Discovery: complementary approaches and convergent findings. *Psychological Bulletin* 125: 524-543.
- Klayman, J., Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review* 94: 211-228.
- Klayman, J., Ha, Y. (1989). Hypothesis testing in rule discovery: strategy, structure, and content. *Journal of Experimental Psychology.: Learning, Memory and Cognition* 15 (4): 596-604.
- Klieme, E. et al. (2003). Zur Entwicklung nationaler Bildungsstandards. Frankfurt a.M.: *Deutsches Institut für Internationale Pädagogische Forschung* 15.
- Kolmos, A. and Kofoed, L. (2003). Development of process competencies by reflection, experimentation and creativity. *Teaching and Learning in Higher Education: New Trends and Innovations*. University of Aveiro.
- Koslowski, B. (1996). Theory and Evidence: *The Development of Scientific Reasoning* – Cambridge Mass.: MIT Press.

- Koslowski, B. and Okagaki, L. (1986). Non-Human indices of causation in problem-solving situation: causal mechanisms, analogous effects, and the status of rival alternative accounts. *Child Development* 57: 1100-1108.
- Koslowski, B. and Okagaki, L., Lorenz, C., and Umbach, D. (1989). When covariation is not enough: the role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development* 60: 1316-1327.
- Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review* 96: 674-689.
- Kuhn, D., Amsel, E.D. & O'Loughlin, M. (1988). *The Development of Scientific Reasoning Skills*. Orlando, Fla.: Academic Press.
- Kuhn, D., Angelev, J. (1976). An experimental study of the development of formal operational thought. *Child Development* 47: 697 – 706.
- Kuhn, D., Phelps, E. (1982). The development of problem-solving strategies. In: H. Reese (Hg.): *Advances in Child Development and Behavior* 17, 1-44. New York: Academic.
- Larson, A.E., Wollmann, W.T. (1976). Encouraging the transition from concrete to formal cognitive function: an experiment. *Journal of Research in Science Teaching* 13, 413-430.
- Lawson, A.E. & Wollman, W.T. (1976). Encouraging the transition from concrete to formal cognitive functioning: an experiment. *Journal of Research in Science Teaching* 13: 413-430.
- Livingston, K., and McCall, J. (2005). Evaluation: Judgemental of developmental? *European Journal of Teacher Education* 28: 165-178.
- Longeot, F. (1962). Un essai d'application de la psychologie genetique a la psychologie differentielle. *Bulletin de L'Institut National D'Etude* 18(3) : 153-162.
- Longeot, F. (1965). Analyse statistique de trois tests genetiques collectifs. *Bulletin de L'Institut National D'Etude* 20(4) : 219-237.
- Lord, C.G., Ross, L., & Lepper, M.R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequence considered evidence. *Journal of Personality and Social Psychology* 37 (11): 2098-2109.
- Marmaroti, P., & Galanopoulou, D. (2006). Pupils' Understanding of Photosynthesis: A questionnaire for the simultaneous assessment of all aspects. *International Journal of Science Education* 28: 383-403.
- Mayer, R. E. (1999). *The Promise of Educational Psychology: Learning in the Content Areas*. London: Prentice Hall.
- McCloskey, M. (1983). Naïve theories of motion. In D. Gentner & Steven, A.L. (Eds.) *Mental models*, pp. 299-324. Hillsdale, N.J. Erlbaum.
- Metz, K. (1998). Scientific Inquiry Within Reach of Young Children. In: B. J. Fraser, and K. G. Tobin (Hg.): *International Handbook of Science Education*, 81-96. Kluwer Academic Publishers.

- Metz, K.E. (1985). The Development of Children's Problem Solving in a Gears Task: A Problem Space Perspective. *Cognitive Science* 9 : 431-472.
- Metz, K.E. (1995). Re-assessment of Developmental Assumptions in Children's Science Instruction. *Review of Educational Research* 65: 93 - 127
- Mynatt, C.R., Doherty, M.E. & Tweney, R.D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of experimental Psychology* 29: 85-95.
- Mynatt, C.R., Doherty, M.E. & Tweney, R.D. (1978). Consequences of confirmation and disconfirmation in a stimulated research environment. *Quarterly Journal of experimental Psychology* 30: 395-406.
- Peel, E.A. (1972). *The nature of adolescent judgement*. New York: Wiley-Interscience, Piaget, J. *Biology and knowledge*. Chicago: University of Chicago Press, 1967. Piaget, J. Intellectual evolution from adolescence to adulthood. *Human Development* 15: 1-12.
- Phan, T. T. H., Hammann, M., Bayrhuber, H. (2004). Testing Levels of Competencies in Experimentation. Poster for "Fifth conference of European Researchers in Didactics of Biology: ERIDOB 2004" University of Patras.
- Phan, T. T. H., Hammann, M., Bayrhuber, H. (2006). Testing levels of competencies in experimentation. Paper presented at Sixth Conference of European Researchers in Didactics of Biology (ERIDOB), London.
- Pithers, R.T. (2000). Critical thinking in education: a review. *Educational Research* 42: 237-249.
- Puthz, V. (1988). Experiment oder Beobachtung? Überlegungen zum Erkenntnisgewinn in der Biologie. *Unterricht Biologie* 12 (132): 11-13.
- Root, D.E., Kelley, B.P., and Stockwell, B.R. (2002). Global analysis of large-scale chemical and biological experiments. *Current Opinion in Drug Discovery & Development* 5: 355-360.
- Roth, W.M., Roychoudhury, A. (1993). The Development of Science Process Skills in Authentic Contexts. *Journal of research in science teaching* 30: 127-152.
- Ruffman, T., Perner, J., Olson, D.R., and Doherty, M. (1993). Reflecting on scientific thinking: children's understanding of the hypothesis-evidence relation. *Child Development* 64: 1617-1636.
- Samarapungavan, A. (1992). Children's Judgement in Theory Choice Tasks: Scientific Rationality in Childhood. *Cognition* 45: 1-32.
- Schauble, L and Glaser, R. (1990). Scientific thinking in children and adult. In D. Kuhn, ed., *Developmental Perspectives on Teaching and Learning Thinking Skills*, 9-26. Basel and New York: Karger.
- Schauble, L and Glaser, R., Raghavan, K. and Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *Journal of the learning Sciences* 1: 201-238.

- Schauble, L. (1990). Belief revision in Children: The Role of Prior Knowledge and Strategies for Generating Evidence. *Journal of experiment child psychology* 49: 31-57.
- Schauble, L. (1996). The Development of Scientific Reasoning in Knowledge-Rich Contexts. *Developmental Psychology* 32: 102-119.
- Schauble, L., Klopfer, L.E., Raghavan, K. (1991). Students' Transition from an Engineering Model to a Science Model of Experimentation. *Journal of Research in Science Teaching* 28: 859-882.
- Science Curriculum Improvement Study (1970). *Subsystems and variables teacher's guide*. Chicago: Rand McNally.
- Seifert, T.L. (2004). Understanding student motivation. *Educational Research* 46: 137-149.
- Shaklee, H & Paszek, D. (1985). Covariation judgement: Systemetic Rule Use in Middle Childhood. *Child Development* 56: 1229-1240.
- Siegler, R., Liebert, R. (1974). Effects of contiguity, regularity and age on children's causal inferences. *Developmental Psychology* 10: 574-579.
- Siegler, R.S., Liebert, R.M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: designing a factorial experiment. *Developmental Psychology* 10: 401-402.
- Simon, H. A. & Lea, G. (1974). Problem solving and rule induction: A unified view. In L.W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Erlbaum.
- Simon, H. A. (1999). Problem Solving. In: R. A. Wilson und F. C. Keil (Hrsg.). *The MIT Encyclopedia of the Cognitive Sciences*, 674-676. Cambridge, MA: MIT Press.
- Simon, H. A. (1981). Wissenschaftliche Entdeckung und die Psychologie des Problemlösens. In: H. Neber (Hrsg.) *Entdeckendes Lernen*. 3. Auflage. Weinheim: Beltz, 104-125.
- Sloutsky, V.M., Rader, A.W, Morris, B.J. (1998). Increasing informativeness and reducing ambiguities: adaptive strategies in human information processing. In M.A. Gernsbacher and S.J. Derry, eds., *Proceeding of the Twentieth Annual Conference of the Cognitive Science Society*, 997-1992. Mahwah, N.J.: Erlbaum.
- Sodian, B., Zaitchick. D, Carey, S. (1991). Young children's transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching* 28: 859-882.
- Sturman, L. (2003). Teaching to the test: Science or intuition? *Educational Research* 45: 261-273.
- Tamir, P. (1989). Training teachers to teach effectively in the laboratory. *Science Education* 73(1): 59-69.
- Tamir, P. and Lunetta, V.N. (1978). An analysis of laboratory activities in BSCS Yellow version. *The American Biology Teacher*, 40, 353-357.
- The PISA 2000: Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy.

- The PISA 2003 Assessment Framework- Mathematics, Reading, Science and Problem Solving Knowledge and Skills (OECD 2003).
- Tsai, C.C. (2005). Research and trends in science education from 1998 to 2002: a content analysis of publication in selected journals. *International Journal of Science Education* 2: 3-14.
- Tschirgi, J.E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development* 51: 1-10.
- Vosnisdou, S., and Brewer. W.F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology* 24: 535-585.
- Vosnisdou, S., and Brewer. W.F. (1994). Mental models of the day/night cycle. *Cognitive Science* 18: 123-183.
- Wagner, U., Wurr, A. (1998). Zitroneneis mit Erdbeerschaum und Pfefferminzdekor. Köln: Aulis.
- Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* 12: 129-140.
- Weinert, F.E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: F. E. Weinert (Hg.): Leistungsmessung in Schulen. Weinheim und Basel: Beltz Verlag 17 – 31, 27 f.
- Weisel, M., Haller, K. et al. (1998). Ziel, die Lehrende mit dem Experimentieren in der naturwissenschaftlichen Ausbildung verbinden. *Zeitschrift für die Didaktik der Naturwissenschaften* 4: 29-44.
- Wollman, W. (1975). Intellectual development beyond elementary school VI: Controlling variables: A survey. *School Science and Mathematics*.
- Wollman, W.T. & Lawson, A.E. (1978). The influence of instruction on proportional reasoning in seventh graders. *Journal of research in science teaching* 15: 227-232.