RWTH AACHEN UNIVERSITY

Evaluating the impact of an interactive statistical analysis system on learning statistics

Bachelor's Thesis at the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Sarah Theres Völkel

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Ulrik Schroeder

Registration date: 01.07.2014
Submission date: 15.08.2014

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Aachen, August 2014
Sarah Theres Völkel

Contents

List of Figures

List of Tables

Abstract

HCI researchers have difficulties to perform statistical analysis due to inadequate statistical education. To improve statistics learning, we evaluate the impact of an interactive statistical analysis system on learning statistics. This system, VisiStat, illustrates statistics by the use of visualizations and allows users to perform correct statistical analysis as it automatically applies statistical concepts. Addressing common problems in HCI statistical analysis, we replenished VisiStat with an automatically generated report function, which allows the user to create a sufficient reporting text for the results.

In this thesis, we investigate if students can benefit from using VisiStat, complementing a traditional statistics lecture, by exploring it and, thus, construct their own knowledge. Therefore, we use a quantitative as well as a qualitative approach to assess students' improvements and get in-depth feedback about their learning experience and VisiStat's role in this. Our evaluation revealed that (1) VisiStat is suitable for learning statistics as students developed more knowledge from VisiStat than from a lecture and (2) participants who explored VisiStat first and then attended a lecture outperformed students going through a traditional tell-and-practice procedure. Furthermore, we found out that VisiStat can help to prevent students from conducting common mistakes in statistical analysis. Again, the sequence of the two learning experiences appears to be crucial for students' improvements. Based on students' feedback, VisiStat's role in the learning experience is defined as a tool for practicing and constructing knowledge, encouraging to develop own hypotheses about statistical concepts. The strengths and weaknesses of lecture and VisiStat as well as implications of this study to improve the statistical learning experience are discussed.

Überblick

Aufgrund von mangelnder statistischer Lehre haben Wissenschaftler im Bereich HCI Schwierigkeiten, statistische Analysen durchzuführen. Um das Lernen von Statistik zu verbessern, bewerten wir die Auswirkungen eines interaktiven statistischen Analysesystems auf das Lernen von Statistik. Dieses Analyssystem, VisiStat, veranschaulicht Statistik mithilfe von Visualisierungen und erlaubt Nutzern, korrekte Statistik anzuwenden, da statistische Konzepte automatisch eingesetzt werden. Um übliche Probleme in der statistischen Analyse anzugehen, ergänzten wir VisiStat mit einer automatisch generierten Berichtfunktion, die dem Nutzer erlaubt, einen angebrachten Bericht für die Ergebnisse zu erstellen.

In dieser Bachelorarbeit untersuchen wir, ob Studierende von VisiStat, ergänzend zu einer traditionellen Statistikvorlesung, profitieren, indem sie dieses frei entdecken und eigenes Wissen aufbauen. Um die Verbesserungen der Studierenden sowie ihr detailliertes Feedback zu dem Lernerlebnis und VisiStats Rolle erheben, benutzen wir sowohl einen quantitativen als auch qualitativen Ansatz. In unserer Evaluierung konnten wir zeigen, dass (1) VisiStat für das Statistiklernen geeignet ist, da die Studierenden mehr Wissen durch VisiStat als durch die Vorlesung entwickeln konnten und (2), dass Teilnehmer/innen, die VisiStat zuerst erkundet und anschließend die Vorlesung besucht haben, die Leistungen derjenigen Studierenden übertroffen haben, die einen traditionellen zunächst Theorie, dann Praxis Ansatz durchliefen. Darüber hinaus fanden wir heraus, dass VisiStat helfen kann, Studierende davon abzuhalten, bekannte Fehler in statistischer Analyse zu begehen. Erneut hat sich die Reihenfolge der beiden Lernerlebnisse als entscheidend für die Verbesserungen der Studierenden erwiesen. Basierend auf dem qualitativen Feedback kann VisiStat's Rolle als ein System für praktische Übung und das Aufbauen von Wissen angesehen werden, welches es unterstützt, Hypotheses über statistische Konzepte zu entwickeln. Die Stärken und Schwächen der Vorlesung und VisiStat sowie Schlussfolgerungen dieser Nutzerstudie werden diskutiert, um die Lernerfahrung zu verbessern.

Acknowledgements

First of all, I would like to thank Chatchavan Wacharamanotham, M. Sc., for his extensive support and guidance at every point of time during the course of this thesis. I am very grateful for the time and effort he spent in my supervision, his valuable advice, and constant help. This cooperation was a great benefit for me, resulting in an incredible learning experience.

Furthermore, I would like to thank Krishna Subramanian for his help and support when working on improving VisiStat together. My thanks are also due to Marty Pye, who helped me with the boring task of transcribing.

I would like to thank my thesis supervisors Prof. Dr. Jan Borchers and Prof. Dr. Ulrik Schroeder for their time and support.

A special thanks goes to the students of *Current Topics* this summer semester, who participated readily in the user study and supported to gain insight in statistical learning experiences.

Finally, I would like to thank my friends, especially those, who helped in pilot tests and gave advice to improve the thesis. Last, but not least, I would like to thank my family, who are always supporting me throughout my studies and during this Bachelor's thesis.

Thank you!

Sarah Theres Völkel.

Conventions

Throughout this thesis we use the following conventions.

Text conventions

Definitions of technical terms or short excursus are set off in colored boxes.

EXCURSUS:
Excursus are detailed discussions of a particular point in a book, usually in an appendix, or digressions in a written text.

Definition:
Excursus

Students' quotations in the categorization table are written in typewriter-style text. Grammar or spelling errors are reported literally without mentioning [sic!] to enhance the quotations' clarity. The used symbols are loosely based on a simplified version of GAT2 [Selting et al., 2009].

```
"VisiStat allows a bit for experimenting
and yeah going into depth and thinking
about why a specific test is chosen at a
time."
```

The whole thesis is written in American English.

Chapter 1

Introduction

"I mean, it's statistics, no one likes to learn statistics. And if someone does, they are weird people"

—A participant and student of HCI

Insufficient use of statistics among HCI researchers

The HCI researcher has a field, she or he is a specialist in but there are several other skills the researcher has to develop, for example knowledge about research methods or paper writing. A crucial ability of the HCI researcher is to use proper statistical analysis to support and evaluate his or her research findings. However, the majority of HCI researchers struggles to perform appropriate statistical analysis. Cairns [2007] found out that even papers in respected HCI journals and conference proceedings do not meet a sufficient standard of statistical analysis. Based on the discovered mistakes, he named the four central problems (1) insufficient reporting, (2) not taking assumptions for statistical tests into account, (3) over-testing, and (4) inappropriate testing, which are committed most frequently in HCI research. In Chapter 2.1 the current state of statistical use in HCI research as well as Cairns' analysis of the most frequent problems are presented.

The problem of statistical education

The underlying problem is attributed to a lack of statistical education. Instead of practical courses, which show a step-by-step procedure for necessary statistical concepts,

researchers often only attend a theory based class during their studies and try to teach themselves by reading a book. These books often consist of about 1000 pages and time constraints due to deadlines make it difficult to know where to start reading. Furthermore, statistical procedures are often not practiced everyday but only applied when necessary [Cairns, 2007]. But how can the problem of inadequate statistical education be overcome? As the difficulties with statistical education are not limited to the field of HCI, numerous researchers dealt with the reasons for these difficulties and elaborated suggestions for improvements, which are described in Chapter 2.2.

VisiStat - interactive statistical analysis system

A promising possibility to improve learning statistics is the use of technology in statistical education. Demands on these systems as well as previous research are introduced in Chapter 2.3. Subramanian [2014a] developed the interactive statistical analysis system VisiStat, aiming to help researchers performing correct statistical analysis, even beyond their statistical knowledge. In VisiStat, he presented the results of statistical tests directly linked to their corresponding visualizations. A detailed description of VisiStat is given in Chapter 3.2. As he found out that VisiStat enabled the user to conduct interpretations about statistical concepts, VisiStat's potential to be used as a learning tool is investigated in this Bachelor's thesis.

1. Contribution: Reporting function in VisiStat

This Bachelor's thesis aims to make two main contributions. In order to address all of Cairns' four problems, Subramanian's version of VisiStat is completed with an automatically generated and sufficient report of the results. To meet the requirements placed on a scientifically adequate report, different sources are examined and eventually, demands are defined, which are analyzed in Chapter 3.3.1. Based on these demands, a reporting text pattern is developed, illustrated in Chapter 3.3.2.

2. Contribution: Evaluating VisiStat's impact on learning statistics

Secondly, VisiStat, its role in a lecture-based learning experience, and its impact on learning statistics in general are investigated. Therefore a large-scale user study is conducted, which is adapted to Schneider et al.'s methodology [2013]. Instead of replacing a traditional learning experience, like a lecture or book, Schneider et al. intended to complement

the traditional learning treatment with a tangible user interface to improve learning neuroscience. In this thesis, we chose a similar approach, examining whether students benefit more from VisiStat if they explore VisiStat before they attend a lecture. This procedure is called *Preparation for Future Learning* and is introduced in Chapter 2.4. Furthermore, students are asked to evaluate their learning experience to define the role VisiStat can play in a limited exposure, investigating VisiStat's strengths and weaknesses and how they complement a lecture. Eventually, it is studied if VisiStat can prevent students from making the common mistakes in statistical analysis, Cairns found out. Summing up, this Bachelor's thesis aims to find answers to the following research questions:

Preparation for Future Learning Approach

Research questions

- How can VisiStat complement lecture-based learning?
- What are VisiStat's strengths and weaknesses and thus, its role in a statistical learning experience?
- How can VisiStat help to prevent students from making the mistakes Cairns reported?

The user study and the conducted research methods are presented in Chapter 4.1.In Section 4.2, the won results are described for each of the methods. Against the background of the here presented research questions, the results are evaluated and interpreted in Chapter 4.3. The limitations of the applied methods as well as problems which occurred during the user study are mentioned in Chapter 4.4. Finally, the results are summed up and suggestions for future work are given in Chapter 5.

Chapter 2

Related work

In the following chapter, the current state of HCI and statistics is presented and the most alarming problem areas are described (Section 2.1). Afterwards, the underlying problem of statistical education is analyzed and its investigation by different researchers outlined. Furthermore, suggestions for improvements are demonstrated, focusing on learning principles which were developed (Section 2.2). As a solution for this problem the use of *e-learning* is considered, examining other research, dealing with interactive (statistics) tools for learning (Section 2.3). Eventually, in Section 2.4, the *Preparation for Future Learning* approach is explained to introduce an alternative to the traditional tell-and-practice learning, which is applied in this user study.

2.1 Statistical practice in HCI

Statistics is an important method in HCI research [Lazar et al., 2010]. Books describing HCI or behavior research methods usually include a chapter about statistical analysis, especially as questionnaires, which are evaluated statistically in most cases, are widely used in HCI research [Shneiderman and Plaisant, 2010]. However, Gray and Salzman [1998] observed the inappropriate use of research methods in HCI, including the poor practice of statistics.

Nearly 10 years later, Cairns [2007] came to the same conclusion. Investigating 80 papers of high standard from two years (2005 and 2006) of the BSC HCI conference as well as from the 2006 editions of HCIJ and TOCHI, which are prominent HCI journals, he found out that 41 of them made use of statistics. Due to this fact, he also stressed the relevance of statistics in HCI research. Yet, 40 of these 41 papers did not meet the demands of adequate statistical analysis but showed serious misuse of several concepts of null hypothesis significance testing (NHST). Furthermore, the one paper without statistical issues did not use much statistics at all so that it is barely comparable to other papers. Cairns classified these problems into four areas: reporting, checking assumptions, over-testing, and use of inappropriate tests [Cairns, 2007]. The next paragraph introduces these areas and their underlying problems.

Four problem cases in HCI statistics use

1. Reporting

The most often made mistake, detected by Cairns [2007] in 25 cases, is insufficient reporting of results of NHST. The *American Psychological Association* (APA) [2010] published a manual, which described proper research publications, including the complete description of results. Cairns applied their standard to the examined papers, finding several misuse. For example, he named missing test statistics or no information about which values are being compared. In Chapter 3.3, the American Psychological Association's reporting guidelines are presented in detail. Additionally, researchers often failed to check the assumptions of NHST. For parametric tests the data has to be normally distributed (within each group) and at least interval scale. Moreover, homogeneity of variances has to be fulfilled. In ten papers, the assumptions were violated without drawing the correct consequences of using either using robust or non-parametric tests. A third problem is over-testing of data (15 cases), resulting in possible false positive results. Over-testing describes the increased probability of significant values by chance when several tests are conducted on the same data. Another problem, which is related to over-testing, involves four-way (or even more) ANOVAs, which have to compare many values so that there is a chance of 0.54% of at least one accidentally significant result. Over-testing can also mean that several measures are performed on the same data without controlling possible relationships

2. Assumptions

3. Over-testing

between the dimensions. Eventually, 12 papers made use of inappropriate testing, which is related to the other problems. The performance of pairwise *t*-tests instead of an ANOVA or parametric instead of non-parametric tests are examples for this problem. Furthermore, some writers were not aware of the meaning of ANOVA results as they directly drew conclusions which group causes the significant difference without conducting post-hoc tests [Cairns, 2007].

4. Inappropriate testing

But why do researchers fail to apply statistics in an adequate way? Cairns [2007] attributed this problem to insufficient statistical education. Apart from Cairns, several other researchers investigated the misconceptions students have with statistics and developed approaches to overcome these obstacles. Some of these approaches are presented in the following section.

2.2 Problems in Statistical Education

Cairns [2007] assumed that the reason for the misuse of statistics among HCI researchers is insufficient statistical education. He claimed that statistics is difficult to understand and books are unable to convey the complex topic satisfactory. Instead of theoretical learning, he emphasized the importance of practicing statistics for really understanding and using them appropriately. He named psychological researchers as an example as they have proper lectures about statistics [Cairns, 2007]. However, HCI is not the only research area having problems with statistics but this seems to be an interdisciplinary obstacle as researchers reported comparable difficulties in health science [Zhu, 2012], geology [Johnson, 1999], and biology [Zuur et al., 2010]. Even psychologists, for who statistics has been an integral part of research for a long time, complained about common misinterpretations of statistical concepts among students as well as their teachers [Haller and Krauss, 2002]. Furthermore, statistics problems are investigated at all age levels as well as different points of view (teacher vs. student). For example, Schwartz and Martin [2004] investigated the use of different learning approaches for improv-

Statistics problems everywhere

ing 9th grade pupils' understanding of statistics, whereas Leavy et al. [2013] examined secondary teachers, who perceived statistics learning as well as teaching as difficult and consequently, avoided it in class. Moreover, much literature focused on statistics education at college level.

Statistics education at college level

In this thesis, we focus on users in college-level education. Zieffler et al. [2008] carried out a review of literature about teaching and learning of introductory statistics at college level. The following subsection presents their categorization of literature and gives example of studies in each category.

2.2.1 Literature review of statistics education at college level

Reviewing literature on statistics education at college level, Zieffler et al. [2008] organized the studies into four categories:

- Identification and cause of faulty statistical reasoning
- Assessing of cognitive outcomes from statistical education
- Assessing of non-cognitive outcomes from statistical education
- Difficulties in the teaching of statistics

Each of these categories is briefly described in the course of this subsection, completed by example studies.

Identification and cause of faulty statistical reasoning

As a first category, Zieffler et al. [2008] analyzed studies which identified misconceptions and faulty reasoning of statistical concepts. For example, Garfield and Ahlgren [1988] revealed that students at college level have limited knowledge of basic statistics due to their difficulty of abstract reasoning. Even students who enrolled in an introductory statistics course might not have a profound understanding of the presented statistical concepts [Garfield

et al., 2005]. Furthermore, researchers tried to establish the cause of these misconceptions. Konold [1995] found out that students' intuitive perception of concepts is often informal and not correct. Summing up the results from research, Zieffler et al. [2008] suggested to improve statistics learning and teaching by becoming aware of the nature of students' misconceptions and trying to address them.

Assessing of cognitive outcomes

Whereas studies in the first category aimed to show the frequency of statistical misconception, other researchers focused on the development of reasoning [Zieffler et al., 2008]. Regardless of methodological approach, they all discovered that students reasoning about statistical concepts is limited even after learning about it [Zieffler et al., 2008]. Quantitative analyses tried to assess students' reasoning by creating and applying corresponding tests. Therefore, Garfield [1998] developed the *Statistical Reasoning Assessment (SRA)*, consisting of 20 items which assess students' reasoning of different types of errors. The SRA was applied by several researchers, revealing "surprisingly similar (and poor) results despite country or type of course" [Zieffler et al., 2008]. Garfield et al. [2007] developed a second version of the SRA, the *Comprehensive Assessment of Outcomes in a First Statistics course (CAOS)*, which focused on evaluating whether students understood the underlying overview and intentions of statistical concepts. Qualitative approaches, like interviews [Clark et al., 2003], established comparable results, finding out that students' understandings are often naive and limited even when they dealt with basic concepts like mean and standard deviation.

Assessing of non-cognitive outcomes

Apart from the cognitive development of statistical knowledge as a result of a statistics course, non-cognitive outcomes were examined. An important non-cognitive aspect was students' attitude towards statistics [Zieffler et al., 2008]. An attitude described the perceived utility of statistics and students' impression of their own statistical abilities as well as their opinion on the difficulty of learning statistics [Gal and Ginsburg, 1994]. The majority of students attending an introductory statistics class had a negative attitude towards statistics [Autin et al., 2014]. This negative attitude could lead to insufficient studying, resulting in a low score in the exam [Budé et al., 2007]. Finney and

Schraw [2003], too, stressed that the course performance depends on students' self-efficacy. Moreover, anxiety towards statistics can be a factor that has negative effects on achievements in class [Zieffler et al., 2008].

Difficulties in teaching of statistics

Zieffler et al. [2008] presented three categories focusing on the difficulties that occur when learning statistics. However, their last category dealt with opportunities to improve statistics teaching and thus, learning. Providing feedback while students practice and encounter problems was an effective method to enhance statistical knowledge [Lovett, 2001]. Additionally, students' course performance can be improved by cooperative learning [Keeler and Steinhorst, 1995]. The use of technology can help to achieve better results in statistics teaching and was applied several times [Zieffler et al., 2008]. The potential benefits of e-learning are revisited in Section 2.3.

It was shown that the problems and possible approaches to solutions for statistical education have been discussed in full detail in literature. Garfield and Ben-Zvi [2007] summed up all difficulties and formulate arising principles to improve learning. The following subsection introduces these principles.

2.2.2 Improvement: Learning Principles

In an attempt to improve the learning of statistics, Garfield [1995] proposed eight principles for learning statistics, conducting a meta analysis of research dealing with statistics. Twelve years later, Garfield and Ben-Zvi [2007] revisited these principles, which are presented in this section. They reviewed papers focused on how students behave and how they can be taught effectively, using different methods like surveys, observations, and video recording. They came to the conclusion that the principles were still supported by recent literature and could help teachers to improve statistical education. The principles can be grouped by student centered principles, which described learning friendly situations for students, and principles, focusing on advice for teachers and supporting behavior. An eighth principle

stated the use of technology in education. In the following paragraphs these principles are briefly introduced.

Student centered principles

1. Constructing knowledge

The first principle expressed that students have to construct knowledge for efficient learning. Constructing knowledge meant to interpret the taught concepts, integrate them in the current knowledge and form an own meaning. Regardless of the teacher's experience, students will not achieve an understanding, unless they have the chance to interpret the learned theories by themselves. Furthermore, students have to be actively involved in the learning process. By solving problems cooperatively in small groups, students learn to analyze problems and discuss different approaches and own ideas. However, students have to present their ideas to the teacher who evaluates them. Apart from developing solutions, teachers should also encourage their students to practice the knowledge they gained. Practicing includes hands-on activities as well as applying well-known concepts in new situations and learning to analyze and evaluate different circumstances and approaches. In the previous section, the misconceptions of statistical beliefs that students have were discussed. To overcome these misconceptions, students should come aware of and be confronted with their errors. The learning gain is especially improved if students form assumptions of statistical concepts first and then compare their results with the the actual meaning. If this meaning contradicts students' beliefs, teachers should assist them to understand and develop the correct concept. In addition to these advice, teachers are asked to take some general recommendations into account, which are represented in the following paragraph [Garfield and Ben-Zvi, 2007].

2. Active involvement

3. Encourage practice

4. Be aware and confront with errors

Teacher centered principles

5. Do not underestimate the difficulty

Teachers are advised not to underestimate students' difficulties of even basic statistical concepts. As it was discussed in Section 2.2, statistical concepts often do not cor-

respond to students' intuitive beliefs, resulting in various misconceptions. These misconceptions yield crucial difficulties in learning statistics. Therefore, teachers are encouraged to be aware of these difficulties. On the other hand, teachers should not overestimate students' understanding of statistical concepts. Even students performing well in final exams might simply have understood a specific type of task or calculation and not the underlying concept. Moreover, they often have difficulties to remember this knowledge. To address these failures, teachers are asked to give consistent and helpful feedback. Regarding feedback, the point of time to provide this feedback is most important. On the one hand, students should develop their own hypotheses first and not be disturbed in this process. However, students need time to think about the feedback and incorporate it so that the grade of a final exam is not the appropriate time for a first feedback. In addition, teachers are especially advised to work on their communication skills, being able to give constructive feedback [Garfield and Ben-Zvi, 2007].

6. Do not overestimate the understanding

7. Give consistent and helpful feedback

Use of technology

Garfield and Ben-Zvi [2007] recommended a third approach to successful statistics teaching which is the use of technology to visualize and explore data. However, they stressed to place certain demands on its use. First of all, the used tool should not just be a replacement of the teacher but teachers are encouraged to take advantage of the opportunities a technology tool offers. Thus, technology can be used to visualize data like illustrating boxplots or histograms. Additionally, it is essential that students are able to explore and manipulate the data on their own strengthening their understanding [Garfield and Ben-Zvi, 2007]. In Chapter 3, we investigate in how far traditional learning methods (book and lecture) and an interactive technology tool, like used in this user study, fulfill these eight learning principles. Several technology tools for statistical education have already been developed. The following subsection deals with technology for learning in general and demonstrates some statistical education systems.

8. Technology to visualize and explore data

2.3 E-Learning: a solution for the problem?

Section 2.2 describes one major reason for the alarming state the use of statistics in HCI research is in. But how can this obstacle be overcome? How can learning statistics actually be improved? One of Garfield's and Ben-Zvi's [2007] learning principles proposed to use technology tools which enable students to to visualize and explore data on their own. These technology tools, which make use of "learning conducted via electronic media, esp. on the Internet", were defined as *e-learning* (a combination of electronic and learning) by the Oxford English Dictionary [2014] and revolutionized modern education [Sun et al., 2008]. The research field of e-learning could be described and defined in many ways and still has a lot potential [Friesen, 2009]. However, these definitions shared the use of information and communication technologies (ICTs) with the goal to "facilitate and enhance learning and teaching" [Koper, 2007]. Furthermore, satisfying the requirements on conveying knowledge, e-learning was expected to engage students and help the learning process [Clark and Mayer, 2011].

Seven criteria for successful e-learning

Sun et al. [2008] investigated critical criteria affecting students' satisfaction with an e-learning system. Evaluating nearly 300 questionnaires, they elaborated seven factors, which described 66.1% of the variance of learners' satisfaction. The most crucial aspect was *quality*, achieved by appropriate content presented in adequate time. As they focused on an e-learning course replacing a traditional lecture, they stressed the importance of the *flexibility* of the learning tool. Flexibility was considered to be a key advantage of online learning because it was characterized by the possibility to make use of the system whenever the user wanted to. Furthermore, the *perceived usability* as well as the *perceived usefulness* of the system affected students' actual use of it. On the other hand, the *instructor's attitude* toward e-learning played a critical role as well as their behavior could engage students. The instructor was asked to *assess students' achievements* in different ways, for example by offering the opportunity for self-assessment. The seventh factor was identified as *students' anxiety and self-efficacy* to-

ward the use of the system, which included computer and internet skills [Sun et al., 2008]. However, as in the here presented user study participants with a computer science background are examined, this factor might be ignored.

Interactive visualizations for learning

As stated in the previous section, Garfield and Ben-Zvi [2007] characterized active involvement as well as taking advantage of opportunities for visualizations as the most crucial criteria for the use of technology for learning. Schweitzer and Brown [2007] made use of such visualizations for computer science learners simplifying the understanding of algorithms. To establish these visualizations in the class room, they emphasized the importance of active learning to involve and engage students, resulting in an interactive experience. Perer and Shneiderman [2008] found out that the integrated use of statistics and visualizations can enhance statistical exploratory data analysis.

E-learning tools for math and natural science

Falcão and Price [2009] analyzed the impact of collaboration when exploring a tangible tabletop for learning physics. By collaboratively interacting and interfering with a group, students constructed knowledge in a collective process [Falcão and Price, 2009]. The area of computer science was addressed by Naps et al. [2002], stressing the inevitable condition of an active learning environment for the use of visualization. They claimed that the visualization technology is useless otherwise. Researching into the use of a visualization software for improving geographical education, Pang [2001] observed a positive relationship between the visualization tool and visual understanding as well as creative thinking. Yet, the same results could not be reached for statistical analysis revealing a weakness of the software. However, success has been measured for the effect of e-learning tools on learning statistics, which are focused in the next paragraph.

Studies in e-learning for learning statistics

Tables 2.1 and 2.2 show an overview of researchers investigating statistical analysis tools for learning statistics. Although the use of such tools showed success, several aspects have not been considered yet. Schneider et al. [2013] investigated a tangible user interface to improve learning neuroscience. Examining the interaction between the use of a traditional textbook and the exploration of the tangi-

Who	e-learning tool	Aim	Method	Results
Aberson et al. 2000	Web-based interactive tutorial: Students can explore shapes of sampling distribution for several sample sizes	Compare effectiveness of e-learning tutorial and lecture	111 students used either tutorial or attended a lecture, evaluation: pre/post-test, feedback questionnaire	Tutorial and lecture are comparable in effectiveness: no statistically significant differences, easy and understandable tutorial
Lane and Tang, 2000	Sampling distribution simulation	Compare effectiveness of simulation and textbook as well as the influence of specific vs. non-specific questions when asked about interpreting the results	115 college students used either textbook or tutorial and assigned to specific or non-specific questions or control group, evaluation: transfer questions	Simulation students significantly outperformed textbook students, students asked specific questions achieved better results than those answering non-specific
Schneider (publication pending)	Tangible user interface Combinatorix for small groups to explore probability	Compare the sequence of treatments (Combinatorix → lecture video vs. lecture video → Combinatorix) and its effect on learning probability	24 students, assigned to treatment or control group (AB/BA cross study) Pre- and post-test, coding of students' quality of collaboration, categorization of utterances when exploring the tabletop	Students exploring the tabletop before attending the lecture significantly achieved higher results than the control group

Table 2.1: Use of statistical analysis systems and their contribution to learning statistics in research - Page one

Maxwell 2014	Visual statistics software ViSta (hypothesis testing, calculate confidence intervals, z-scores, transformation, boxplots, descriptive statistics)	Effect of visual statistics software ViSta on students' achievement in elementary statistics course (one semester)	237 undergraduate students randomly assigned to treatment or control group, students used software throughout semester, were given several introduction lessons and three hands-on lessons Evaluation: pre- and post test, assessing attitudes, statistics self-efficacy, perceptions of learning environment, statistical reasoning abilities, informal observations and interviews	No statistically significant difference in achievement, significant difference in students' statistics self-efficacy, statistically significant correlation between overall grade, attitudes towards course, attitudes towards statistics in the field, interpreting and applying statistical procedures, identifying scales of measurement, and the negotiation scale of students' learning environment

Table 2.2: Use of statistical analysis systems and their contribution to learning statistics in research - Page two

ble user interface, they conducted an AB/BA-cross study to detect possible effects due to the sequence. They found out that the hands-own experience with the system resulted in higher scores than reading a textbook extract. Moreover, the sequence of exploring the tangible user interface first and then reading the text book, affected the result positively. This approach, called Preparation for Future Learning, is presented in the following section. An additional influence was represented by the quality of students' verbalizations when exploring the system [Schneider et al., 2013]. Maxwell [2014] was the only one investigating a more complex software which enabled learners to perform inferential statistics like hypotheses tests. The others concentrated on basic statistical knowledge, as probability and the understanding of sampling distributions. In contrast to the study presented in this thesis, Maxwell [2014] introduced the visual statistics software several times during his lecture and let students perform hands-on activities at the end of the entire course. In contrast to this procedure, Schneider et al. claimed the positive effect of the exploration of the system before the lecture, whereas Aberson et al. [2000] as well as Lane and Tang [2000] examined if statistics systems could replace a traditional learning treatment. Their results focused on pre- and post-test measures or feedback questionnaires in order to estimate the system's effectiveness. In addition, Schneider et al. assessed the verbalizations during the exploration and the quality of collaboration, Maxwell [2014] conducted informal interviews when students gave feedback in class. However, they did not investigate learners' reasons for their improvements or the lack of them. Moreover, detailed feedback about the learning experience and the strengths and weaknesses of traditional learning methods as well as e-learning tools has not been investigated yet. This thesis tries to address this gap in research and therefore, investigates advanced inferential statistics learning and a deep analysis of reasons for students' behavior.

2.4 Preparation for Future Learning Approach

The previous section showed different approaches how to use e-learning to improve learning in general and especially statistics learning. In the following user study (Chapter 4) a similar experimental design to Schneider et al. [2013] is chosen. Therefore, the effects of an interactive analysis system complementing a traditional lecture in two different ways are investigated. One group receives a traditional tell-and-practice treatment, attending the lecture first and practicing with the system afterwards. The second group explores the system on their own without previous introduction and learns the theory in the subsequent lecture. This constructive approach is called *Preparation for Future Learning (PFL)* and was developed by Bransford and Schwartz [1999].

PFL approach

Learning new concepts is not a separate process but builds upon prior knowledge and tries to integrate new information. However, students often lack the necessary cognitive structures for this achievement. The PFL approach tried to fill this gap by preparing students for a future learning treatment with a previous learning activity. In this first learning activity, students are encouraged to tackle a problem and then contrast their own solutions and beliefs with the actual solution in the second learning treatment [Schneider et al., 2013], [Bransford and Schwartz, 1999].

Preparation activity

But how is such a preparation designed to set the stage for the future learning? Bransford and Schwartz [1999] stressed the importance of transfer learning in contrast to memorizing facts or procedures, as transfer learning represents a deep understanding of a problem. To gain this deep understanding, they proposed the concept of contrasting cases introduced by the psychologist Gibson [1969].

Contrasting Cases

Contrasting cases describes the advance from a novice to an expert. For example, a gardener or florist knows the differences between lots of roses. She or he knows the different form of the petals, the growth and small changes in colors. In contrast, the novice just admires beautiful roses

and might be able to see the difference between a white and a red rose but not the difference between an English Mary Rose and an English Wife of Bath. Applying this example to statistics, the novice simply knows that there are null hypothesis significance tests. However, the expert knows the difference between test for between- and within-groups design, the dependence on assumptions and number of variables. Thus, the expert is able to distinguish between several phenomena whereas they seem similar to the novice. Contrasting cases encourage the novice to notice these small differences between phenomena they might have not recognized. By contrasting cases, students form their own hypotheses about the reasons for a phenomenon and develop different theories and assumptions [Schneider et al., 2013]. In the interactive statistics analysis system, students might recognize a different test when the number of conditions of the independent variable are increased. They contrast this case to the case before where the independent variable consisted only of two levels. As a consequence, they suspect a reason for this difference and therefore construct an own meaning. Accordingly, they tackled the concept and are prepared for future learning, e.g. in form of a lecture or a book. In this second learning activity they are able to contrast their own theory with the actual meaning [Schneider et al., 2013].

Application of PFL approach

The PFL concept has already been applied successfully. Schneider et al. [2013] showed that students exploring a tangible interface for learning neuroscience first and reading a textbook afterwards outperform students trained with the traditional tell-and-practice approach. Schwartz and Martin [2004] tested the PFL approach on 9th-grade students learning descriptive statistics, finding out that invention activities followed by a lecture strongly improved students' statistics skills. Examining students' knowledge gain in physical phenomena, Schwartz et al. [2011] stated that the tell-and-practice method could undermine the ability to transfer concepts. Students in a tell-and-practice group and PFL students scored the same regrading using formulas but PFL students exceeded them in transferring skills and learning of the ratio structure of physical concepts. Furthermore, they concluded that the PFL approach helped low- as well as high-achieving students [Schwartz

et al., 2011].

Summing up, the PFL approach is promising to enhance learning statistics by using the interactive analysis system VisiStat, which encourages students to contrast cases and therefore, gain a deep understanding of statistical phenomena. The PFL approach includes two learning activities: the inventing part and the theoretical part (e.g. lecture), which contrasts other approaches presented in the previous section. In how far a combination of VisiStat and lecture can improve students' statistical skill is analyzed in Chapter 4. VisiStat as well as the goals for learnable statistics are presented in the following chapter.

Chapter 3

Improvements to VisiStat

In Chapter 2, the gulf between qualitative use of statistical analysis and the current practice in HCI research was shown. It was analyzed that the underlying reason for this problem is the inadequate statistical education. To overcome this obstacle, interactive statistical analysis tools were introduced as a possibility to enhance learning statistics. Such a system was developed by Subramanian [2014a] aiming to support researchers with their statistical analysis.

As a next step, it is investigated, in how far this system can complement traditional statistics learning. An overview of Subramananian's system VisiStat is given in Section 3.2. However, VisiStat only approaches inappropriate testing and checking of assumptions. To address all problems reported by Cairns [2007], the former version of VisiStat has to be replenished with a prevention of over-testing and a sufficient, automatically generated reporting text. This chapter explains the implementation of a reporting functionality. We begin by describing the score of statistical tests of interest. Then, we outline the design and the rationale of the user interface modifications. The last part of this chapter focuses on the design and implementation of the reporting function.

3.1 Scope of statistical tests

The field of possible statistical calculations in scientific research is wide. There is no standard of qualitative statistical analysis in HCI research, mainly taking over psychology's yardstick [Cairns, 2007]. In psychology as well as in HCI research, null hypothesis significance testing (NHST) is widely used just as discussed [Kaptein and Robertson, 2012]. Additionally, effect sizes and confidence Intervals are provided, according to the *American Psychological Association* (APA) [2010]. Furthermore, the APA manual asks the user to provide extensive descriptions of results, which are analyzed in Section 3.3. Below, the use of NHST is discussed.

Score: null hypothesis significance testing, effect size, and confidence intervals

Some researchers even claimed to replace significance testing with effect sizes and confidence intervals, like Cohen [1994]. This discussion is still continued almost twenty years later in HCI research, as Kaptein and Robertson [2012] proposed to substitute NHST by effect size, in particular Cohen's d, and Bayesian analysis. Dunlop and Baillie [2009] intended to introduce the problem of significance testing to HCI research in general, and especially mobile HCI, as well. To compare the different methods, Wetzels et al. [2011] analyzed 855 t-tests applying significance tests, effect size measures, and Bayesian analysis. They found out that significance tests and Bayesian analysis agree on the better hypothesis in general but differ concerning the strength of the effect. As the Bayesian analysis provides more cautious results, it is preferred by the authors. Furthermore, they confirmed the evidence of additional effect size measures for significance tests [Wetzels et al., 2011].

Criticism of NHST

Despite Wetzels et al.'s recommendation [2011] to use Bayesian analysis, the user study in this thesis focuses on NHST because the procedure for Bayesian analysis is complicated and not widely used in HCI. Yet criticized, NHST is still the standard in HCI research due to a simpler mental model and mistakes are prevalent [Cairns, 2007] and [Dragicevic et al., 2014]. Thus, the scope of statistical tests in the user study is limited to NHST complemented by effect size and confidence intervals as recommended by APA. An

Narrowing down statistics for user study

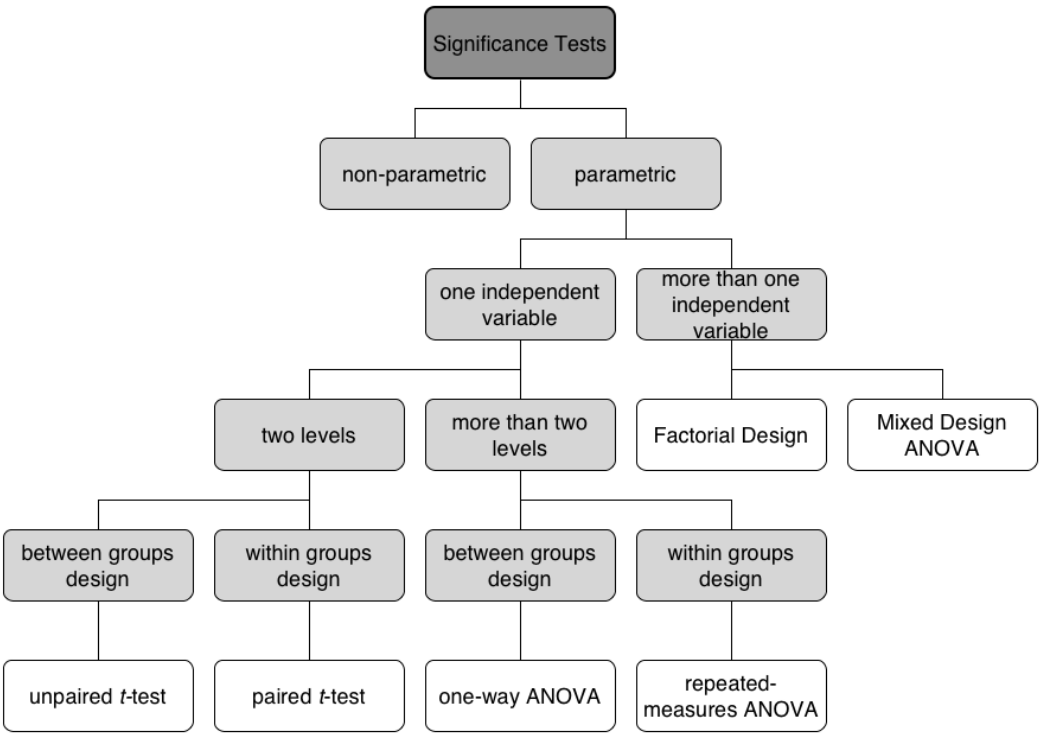


Figure 3.1: Scope of statistical tests in the user study

overview of the used significance tests is given in figure 3.1. Among the significance tests, the selection is narrowed down to parametric tests to prevent overtaxing students as only one lecture (of one and a half hour) in this user study focuses on teaching statistics. However, the difference and principle of parametric in contrast to non-parametric tests is mentioned but students are not asked to specify which non-parametric test is used for which experimental design. They are expected to get an idea of the structure of test selection instead. Additionally, tests with one independent variable are emphasized over factorial ANOVA. For all tests, it is emphasized how the four problems Cairns [2007] defined (cf. Chapter 2.1) can be avoided. The next section describes how a modification of VisiStat addresses these four problems.

3.2 VisiStat - an interactive statistics analysis system

Chapter 2.2 dealt with interactive e-learning tools to improve learning in complex areas students struggle to learn, like statistics. Subramanian [2014a] developed a statistics system, which allows the user to achieve answers for their research questions by interacting with statistical visualizations. Furthermore, the users are able to address statistical tasks they had no previous knowledge of. However, VisiStat was not originally designed as a learning tool but aims to help researchers applying statistics in an appropriate way. Nonetheless, Subramanian [2014a] stated that performing statistical analysis with VisiStat deepens users' statistical knowledge. Thus, VisiStat might improve statistical education.

VisiStat reduces knowledge-in-the-head demands

The user's statistical power is improved by transferring knowledge the user generally has to have in the head to the world [Subramanian, 2014a]. This is a major advantage as Norman [2002] describes the problems people have to "keep knowledge in the head". In case of the performance of a statistical analysis, the user is asked to have knowledge about statistics like the appropriate significance test that has to be chosen for the current situation. Acquiring this amount of knowledge often takes a lot of time to be spent with reading books or attending lectures, and as Chapter 2.2 has shown, even then it is difficult to reach a satisfying amount of knowledge. Additionally, the user has to be able to recall the knowledge during statistical analysis (knowledge in the head) or look it up in references. On the other hand, the researcher needs to be familiar with the current data, for example its distribution. VisiStat addresses these two kinds of knowledge by taking care of the statistical knowledge for the user. It checks the assumptions of a statistical test and based on this, automatically selects the appropriate test. Moreover, it continuously presents visualizations altogether with the statistical results, giving the user constant feedback about the data [Subramanian, 2014a].

In conclusion, it was shown by Subramanian that VisiStat

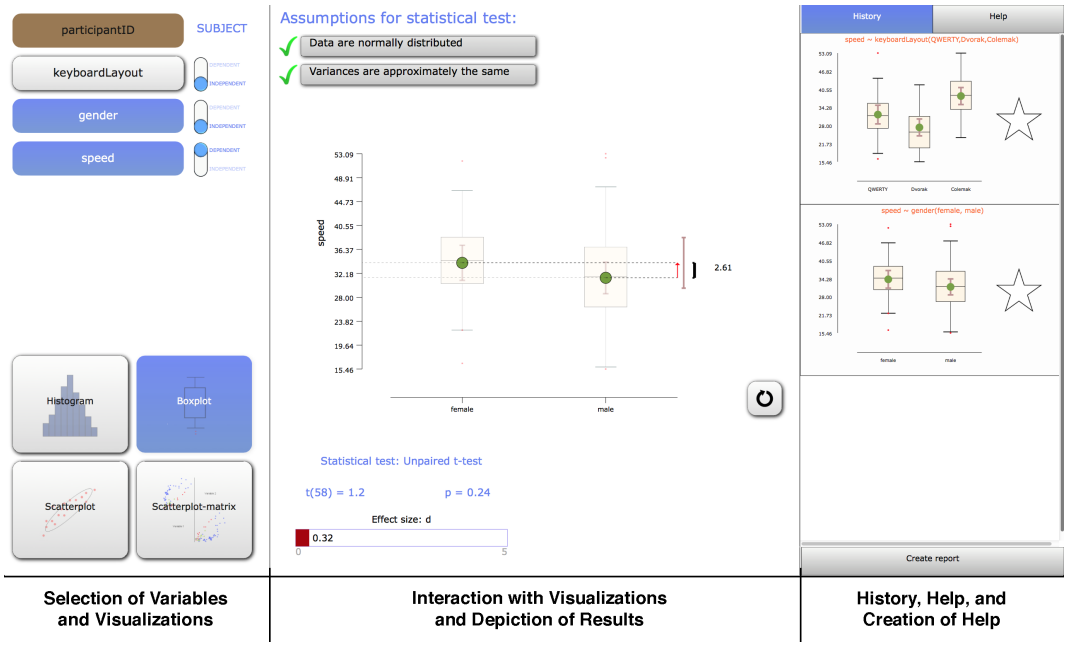


Figure 3.2: Overview of VisiStat subdivided into three parts. Left side: Selection of variables and visualization; middle: Interaction with visualizations and depiction of results; right side: history, help and creation of results

helps researchers to perform complex statistical analysis. But can VisiStat also help students to achieve statistical knowledge by improving the learning experience and therefore preventing them from making mistakes in the future? It is the goal of the present thesis to answer this question by evaluating users' experience which will be presented in Chapter 4. First of all, the next paragraph gives a short introduction to VisiStat. A complete and detailed overview of VisiStat as well as its functionality and components can be found in [Subramanian, 2014b]. Furthermore, some changes that have been made to this original version of VisiStat are outlined.

Overview of VisiStat

The VisiStat view is divided into three sub-parts (figure 3.2). On the left side, the independent and dependent variables are depicted as well as options for different visualizations of the data. The central part of the system is set in the middle, illustrating the actual visualization with which the user can interact, and showing the results of a performed statistical test. A new element is added to this version of

VisiStat which is placed on the right side: A history of the conducted tests represents the possibility to go back to a previous test. Furthermore, it offers the opportunity to create a report for the results of a statistical test. The help functionality is also moved to this right side. In order to perform a statistical test, the user selects independent and dependent variable(s), which are displayed in form of a box plot diagram in the middle part. When he or she now decides to conduct a statistical test, the assumptions for parametric significance tests are examined and the appropriate test is automatically chosen. The user is shown the results of the test and can either choose to investigate a new hypothesis or create a report for the current or a former test. When the user intends to create a report, the view changes to the reporting view and features an appropriate reporting text as well as a figure of the belonging box plot diagram. A detailed outline of the reporting functionality is given in the following Section 3.3.

Addressing Cairns' statistics problems in VisiStat

As explained in the previous paragraph, VisiStat addresses the four problems of statistical analysis in HCI research defined by Cairns. Two of these problems, inappropriate testing and checking assumptions, have already been implemented in Subramanian's [2014b] version of VisiStat. When the user selects to perform a significance test, the data is automatically tested for its normal distribution and homogeneous variances. The user gets visual feedback whether the assumptions are fulfilled or violated. Based on the assumptions, the appropriate test for the given data is chosen by the system. Additionally, a visualization is displayed and shown with the results, like F- or t-value and the corresponding effect size. Cairns' two remaining problems are revised in a second iteration. Whereas the prevention of over-testing was implemented by Subramanian, the implementation of the reporting functionality was added as part of this bachelor's thesis and is described in the following Section 3.3. In case the user conducts pairwise t-tests instead of a one-way ANOVA, she or he is warned that the data might be over-tested and is recommended to use a one-way ANOVA. Figure 3.3 explains the application of the four areas in VisiStat.

Apart from the addition of these two practicalities, some

Figure 3.3: Addressing Cairns' problems of statistical analysis in HCI research in VisiStat: Appropriate Testing, Assumptions, Over-testing, and Reporting

minor changes were made adapting VisiStat to new learners. This includes for example the change of the description "homogeneous variances" to "Variances are approximately the same" or the alteration of the button "Test for differences" instead of "Do Significance Test". After this brief introduction to VisiStat, Garfield and Ben-Zvi's [2007] learning principles are examined again, investigating in how far they are fulfilled by traditional learning methods, such as book and lecture, and by an interactive analysis system like VisiStat.

Usage of learning principles

Table 3.1 outlines that the three different learning methods all have their strengths and weaknesses. Furthermore, the usage of some principles depends on the specific situation,

	Book	Lecture	Interactive system
1. Constructing knowledge	×	×	✓
2. Active involvement	×	✓	×
3. Encourage practice	✓	×	✓
4. Be aware and confront with errors	×	✓	×
5. Do not overestimate the understanding	×	✓	×
6. Do not underestimate the difficulty	✓	✓	✓
7. Give consistent and helpful feedback	×	✓	×
8. Technology to visualize and explore data	×	×	✓

Table 3.1: Usage of learning principles in different learning situations

for example a statistics book can include a practice part but does not have to. Therefore, in this examination the best case for all learning situations is assumed. It can be seen that book and lecture already complement each other, but there are still principles that are not fulfilled, for example the use of technology. A lecture can include technology to visualize data, but it does not support the exploration of this data from the students. In addition, a lecturer as well as a book tell students about statistics but do not construct knowledge, which can be achieved by an interactive analysis system. In conclusion, having a closer look at lecture and system, a combination of these two methods might be promising. In Chapter 4, this table is revisited and evaluated in how far the assumptions can be fulfilled by students' qualitative feedback concerning strengths and weaknesses of lecture and VisiStat.

In a former Version of VisiStat, only two of Cairns' problems (inappropriate testing and assumptions) were addressed. This version was revised and completed by the two remaining problems. In the course of this chapter, the automatic generation of a reporting text is represented. At first, the following section demonstrates the design patterns of the reporting text, based on APA Manual's [2010] recommendations.

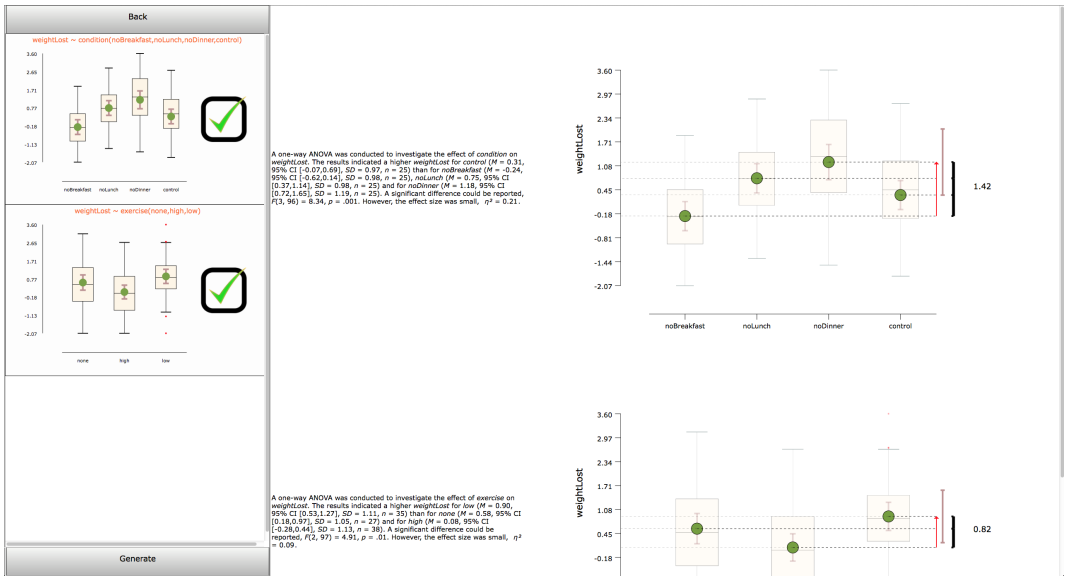


Figure 3.4: The reporting view in VisiStat

3.3 Inserting reporting functionality to VisiStat

Addressing all four statistical problems Cairns detected in HCI research, the possibility to create automatic reports is inserted in VisiStat. These reports use the results VisiStat already provides and complements the system by offering a correct and complete text. Therefore, the user can select to create a report for one or more tests in the history and then reaches to the reporting view, which displays the reporting as well as a corresponding box plot diagram (cf. figure 3.4). The aim of the reporting functionality is to contribute to a scientifically appropriate text, which the user can simply copy to a research paper, ensuring to satisfy the reporting guidelines. As standard for reporting serves the sixth edition of the American Psychological Association's Publication Manual [2010]. In a first step, the demands placed on the reporting functionally by the APA manual as well as textual characteristics are analyzed. Afterwards, the components of the actual reporting text in VisiStat are outlined and described in detail.

3.3.1 Requirements for statistical reports

Cairns [2007] claims insufficient reporting to be the most frequent problem in HCI research. But what is sufficient reporting? To offer researchers guidelines for writing adequate papers, the American Psychological Association [2010] published a publication manual. Their recommendations are presented in this first subsection. Apart from these requirements with regard to content, reporting texts face textual challenges. Therefore, the second part of this subsection deals with the creation of a text type pattern for reporting results. Moreover, evaluation criteria for texts are shortly presented.

The importance of statistical reports

The APA manual [2010] states that reporting serves as justification for the interpretation. Therefore, an overview of the collected data and the performed analysis has to be given. The description has to be detailed enough to enable the reader to understand the process of analysis and empower him or her to conclude alternate interpretations. Nonetheless, writers are admonished to include unexpected and unpleasant outcomes. Summing up, appropriate reporting has to be "accurate, unbiased, complete, and insightful" [APA, 2010].

APA's sufficient set of statistics

For the following development of a pattern for reporting results, APA's description [2010] of an appropriate set of statistics is taken as a yardstick. For inferential statistics tests, it is stressed that null hypothesis significance tests are insufficient but have to be complemented by confidence intervals and effect sizes. Taking this into consideration, the following values are named by the publication manual, establishing a sufficient set of statistics [APA, 2010]:

- exact p-value
- value of statistical test
- degrees of freedom
- effect size
- For each condition of each independent variable

- mean
- 95% confidence interval for mean
- standard deviation
- number of participants (in case of within groups design only once)

APA statistical recommendations

In addition to content related instruction, the APA manual [2010] affords stylistics recommendations. As the analysis was conducted at a particular date in the past, results are described in past tense. Values are rounded at two decimal points except for the *p*-value, which has to be stated exactly. Statistics symbols, such as *p* or *t* are displayed in italic typeface, whereas abbreviations which are not variables appear in standard (e.g. CI). Eventually, an uppercase *N* defines the total number of participants. On the contrary, a subset of participants (e.g. in a between group design) is reported with a lowercase *n*.

Textual requirements

The APA manual defines content and stylistic characteristics of the reporting section in scientific papers. But how can these information be delivered in an appropriate text? To be able to create such a sufficient reporting text, Sandig's model [1997] of text type pattern can be applied. This model explains prototypical characteristics of a particular text type on a grammatical as well as non-grammatical level, aiming at being a standard solution to a textual problem. The textual problem is in this case the reporting text [Sandig, 1997]. Hence, before the actual reporting text is created, the application of Sandig's model [1997] on reporting texts illustrates the important features that have to be taken into consideration. These characteristics are summed up in table 3.2.

Communicative Usability

A web text can be evaluated with the criteria of *Communicative Usability* by Jakobs [2012], which focuses on language as the most important interaction between human and machine. Therefore, the communicative quality of content, interface and further parts like documentations are investigated [Jakobs, 2012]. Furthermore, the Communicative Usability of a text can be described as satisfactory if it supports the reader to solve communicative tasks [Jakobs, 2012]. What is the communicative task of users creating

Type of Action	Prototypical Characteristics
Social aim - Social purpose: justify research results and enable readers to understand analysis - Problem solution: create an appropriate report of results **Situation properties** - Problem situation: presenting research results - Institution: public - Channel: written - Medium: differs (online, journal, ...) **Persons involved in situation** - Writer: author (researcher) - Reader: other researchers (with statistical knowledge) - Relationship: colleagues	**Hierarchy of actions** - main actions: presenting variables and their descriptive statistics, presenting outcome of significance test, presenting effect size - topic: (significant) differences between variables **Pattern of sequence** - which variables were compared - descriptive comparison of variables - outcome of significance test - if necessary: amount of effect size **Pattern of formulation** - scientific language - includes statistical values - precise sentences **Material shape of text** - statistical values in italic - often combined with graphical depiction of results **Average scale** - a few sentences up to one paragraph

Table 3.2: Application of Sandig's model [1997] of text type pattern on reporting results for null hypothesis significance tests

a text for reporting their results? To write a research paper, dhe addressee wants to deliver all necessary information so that readers can understand and retrace the results [APA, 2010]. In addition, the text serves to prove or reject hypotheses and assumptions the author wants to discuss. Consequently, the reporting text has to fulfill these requirements.

Evaluation of content

As this part focuses on the plain text of reporting, only the content principle is considered hereafter. Evaluating the content, the comprehensibility has to be examined, which is addressed by several approaches. A rather simple possibility is Flesch's formula [1948] for reading ease, assuming that a text consisting of short sentences with short words

has a higher legibility. Flesch's calculation results in score between 0 (low legibility) and 100 (high legibility) with a score under 30 can mainly be understood by university graduates whereas a score between 60 and 70 can be easily processed by 15-year-old students [Flesch, 1948]. An analysis von Flesch's reading ease will be given in Section 3.3.2. However, this approach is also criticized because it does not address the reader's previous knowledge. Groeben and Vorderer [1982] and Ballstaedt et al. [1999] propose methods of optimization, like the use of common words in easy and short sentences, summaries, sequence, and advance organizers. Some of these measures and their application in the VisiStat reporting text are discussed in Section 3.3.2.

Use of low coherence sentences for reporting text

Summarizing the demands on reporting texts, it should be stressed that the reporting text aims to develop an automatic conception of the appropriate and scientific depiction of results. It does not intend to provide an explanatory text which helps to understand the results as this contradicts to the demands on reporting results. It was shown in table 3.2 that reporting texts make use of scientific formulations, which contain mostly low-coherence sentences [Best et al., 2005]. Readers with a high previous knowledge benefit from such low-coherence texts as they are forced to form their own conclusion and thereby develop a deeper understanding of the text [McNamara et al., 1996]. In contrast to them, readers who lack the previous knowledge have difficulties to understand scientific texts because they are not able to fill in the conceptual gaps that arouse from low-coherence texts [Best et al., 2005]. Improving their understanding would require high-coherence texts [McNamara et al., 1996]. However, the APA manual [2010] emphasizes to assume readers with necessary statistical knowledge. Therefore, the results are written in a scientific low-coherence style. On the other hand, Cairns uncovered serious statistical problems in HCI research. Consequently, the reporting text should reach a compromise of scientific appropriate description (based on APA's guidelines) and a good comprehensibility.

Definition: *Low coherence texts*

LOW COHERENCE TEXTS:
"Texts are considered to be low cohesion when constructing a coherent representation from the text requires many inferences." [Best et al., 2005]. For example, the following sentence pairs describe two different levels of cohesion:
1. Statistics is regarded as difficult. Students do not like to learn it.
2. Statistics is regarded as difficult. Therefore, students to like to learn statistics.
The second sentence is easier to process as it explicitly states that *statistics* is difficult to learn. Additionally, the connective *therefore* helps to understand the link between the two sentences.

3.3.2 Development of an automatic reporting text of statistical results in VisiStat

The previous section defined the demands placed on the reporting text. Based on these requirements, a pattern for implementation is developed so that results for every text can be inserted appropriately in this pattern. After the development of the general structure is described, the individual components are presented in detail. Moreover, the textual guidelines, discussed in the previous section, are applied and special features demonstrated.

Field's text is used as a basis for a text pattern

The APA manual [2010] defines necessary details for reporting statistical results. However, it does not provide a textual standard or examples. Due to this reason, we use Field's statistics guide [2013], which offered an example for reporting each statistical test, serves as a basis for the reporting text. Nonetheless, some changes have to be made to Field's texts as his sentences often depended on the grammatical type and semantics of variables and he made use of different values than the APA manual. These differences can be recognized in figure 3.5, which shows an example text based Field [2013] and the corresponding VisiStat text for an unpaired t-test . In the course of this subsection, it is dealt with these differences in more detail.

On average, participants given a Colemak keyboard engaged in higher speed (*M* = 5, *SE* = 0.48), than those given a QWERTY keyboard (*M = 3.75, SE = 0.55*). This difference, -1.25, BCa 95% CI [-2.606, 0.043], was not significant *t*(22) = -1.71, *p* =.101; however it did represent a medium-sized effect, *d* = 0.65.	An unpaired *t*-test was conducted to investigate the effect of *keyyboardLayout* on *speed*. The results indicated a higher *speed* for *Colemak* (*M* = 5, 95% CI [4.75,6.3], *SD* = 0.48, *n* = 26) than for *QWERTY* (*M* = 3.75, 95% CI [3.25, 4.08), *SD* = 0.55, *n* = 24). This difference was not significant, *t*(22) = -1.71, *p* = .101. However, the differences constituted a medium effect size, *d* = 0.65.

Figure 3.5: Comparison of Field's [2013] (left) and VisiStat's (right) reporting text for an unpaired *t*-test

Structure of reporting text

As it can be seen in figure 3.5, the reporting text for unpaired *t*-test describes a four sentences structure.

1. Aim and Method
2. Descriptive results of the conditions of the independent variable
3. Significance result
4. Amount of effect size

These four sentences as well as their components are analyzed in the following paragraphs.

1st sentence: aim and method

> An unpaired *t*-test was conducted to investigate the effect of *keyboardLayout* on *speed*.

The first sentence describes which test was used and which variables were compared. This first sentence is not included in Field's text [2013] but was inserted to enhance the clarity of the text. Due to the amount of necessary statistical values, the second sentence is difficult to read. By starting with this first sentence, this obstacle is overcome as the overall information is already given in this first sentence. Furthermore, Ballstaedt [1999] recommends to use

such *advance organizers* to activate previous knowledge. The naming of the statistical test serves to increases the reader's comprehensibility for the statistical analysis as this knowledge is often limited (Chapter 2.2). The main difficulty, regarding the automatic creation of this first sentence, appears to be the variables which may be available in different grammatical form. Thus, a sentence structure which does not depend on the number (singular vs. plural) and part of speech (adjective vs. substantive) was chosen, but problems might occur nonetheless. To address these problems, the variables were set in italic typeset enabling the user to replace them easily by the desired phrase. Highlighting of phrases is also recommended by Groeben and Vorderer [1982] to improve the comprehensibility.

2nd sentence: descriptive results

The results indicated a higher *speed* for *Colemak* (M = 5, 95% CI [4.75,6.3], SD = 0.48, n = 26) than for *QWERTY* (M = 3.75, 95% CI [3.25, 4.08], SD = 0.55, n = 24).

In this case, the second sentence uses a grammatically correct phrase but if the dependent variable was *errors* for example, the variable would be in plural form and the article "a" before would be unnecessary. However, in many cases, the dependent variable consists of a singular form so that the author must decide which form is correct. Therefore, the variables are highlighted in italic typeset once again. As mentioned before, in this part, the conditions of the independent variable and their relationships are presented by providing the descriptive statistics, the APA manual [2010] set in advance. Due to the many statistical values, this sentence is difficult to read. In order to reduce the difficulty, the sentence structure is designed as easy as possible and contains no further information. Additionally, the use of easy, short, and active main clauses with concrete, illustrative, and well-known words can enhance the comprehensibility [Groeben and Vorderer, 1982].

3rd sentence: significance result

This difference was not significant, $t(22) = -1.71$, $p = .101$.

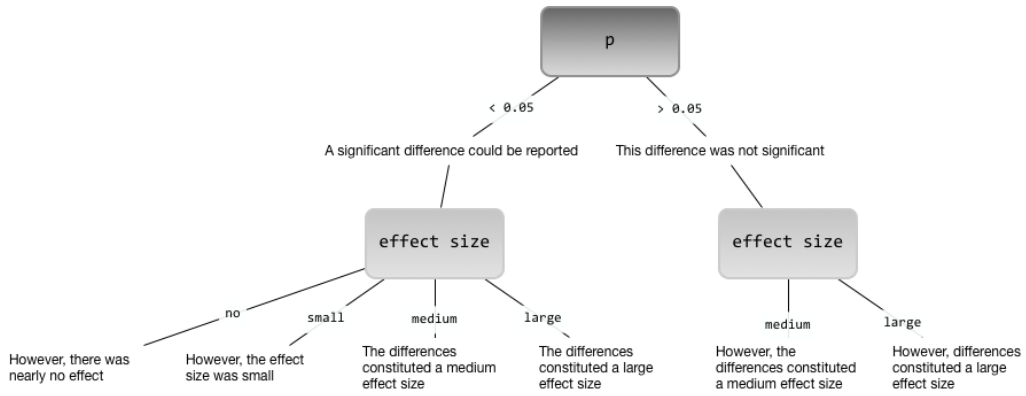


Figure 3.6: Reporting text for significance and effect size result dependent on *p*-value and effect size

The third sentence uses the same choice of words as Field, indicating whether the result of the applied test was significant or not. The display of this sentence depends on the resulting *p*-value so that in case of a significant result, the phrase is adapted accordingly. Anew, the sentence structure is concise. In contrast to Field [2013], the necessary statistical values (test value, here t, and p) are just presented at the end of the clause, increasing the flow of reading. As the APA manual [2010] does not state the necessity of describing the difference in means and the confidence intervals of each mean are already given in the second sentence, this part can be left out. Furthermore, Field's sentence is broken down into two sentences: one for each significance and effect size. On the one hand, the sentences are once again short and therefore easily comprehended. On the other hand, each topic receives one separate sentence establishing a sequence from basis to results [Groeben and Vorderer, 1982].

> However, the differences constituted a medium effect size, $d = 0.65$.

4th sentence: effect size

The final sentence gives information about the effect size and highly depends on the previous sentence concerning significance. Figure 3.6 shows the dependencies of effect size and *p*-value and the resulting displayed sentence. In

A pairwise t-test (Bonf.) revealed that there was a significant difference between *noBreakfast*and *noDinner*, *t*(3, 96) = 4.77, *p* = .001. The differences constituted a large effect size, *d* = 1.28.

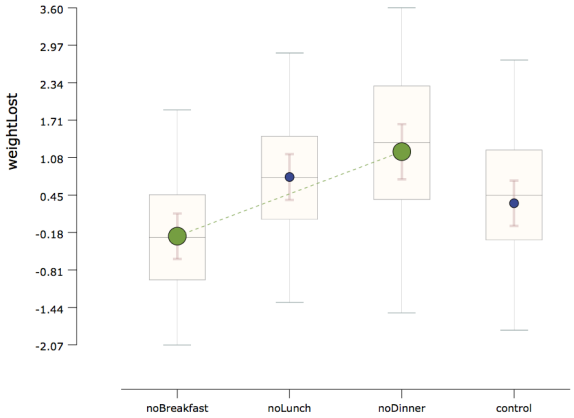


Figure 3.7: Example reporting text for a post-hoc text

case of contradicting results (not significant but high effect size or significant but small effect size) the word "however" is used to demonstrate the contrast of both values. Thus, the contradiction is emphasized even for non-experts by the employment of coherence. Although it was stressed before that coherence does not characterize scientific texts, it was elaborated in Chapter 3.1 that many researchers are still unfamiliar with effect sizes. Therefore, they are supported in developing a meaning of these results.

Generalization about tests

The previous paragraphs introduced the structure of a reporting text for an unpaired *t*-test. But how are the texts for the other tests in figure 3.1 generated? In general, there are three different structures for reporting texts:

- reporting text for tests with one independent variable
- reporting text for tests with more than one independent variable
- reporting text for post-hoc test

Thus, the text remains the same for all tests with one independent variable, regardless of parametric or non-parametric as well as between or within groups design. In order to mention the appropriate test-method and corresponding values (e.g. the effect size for one-way ANOVA is η^2), these values are automatically selected and inserted in

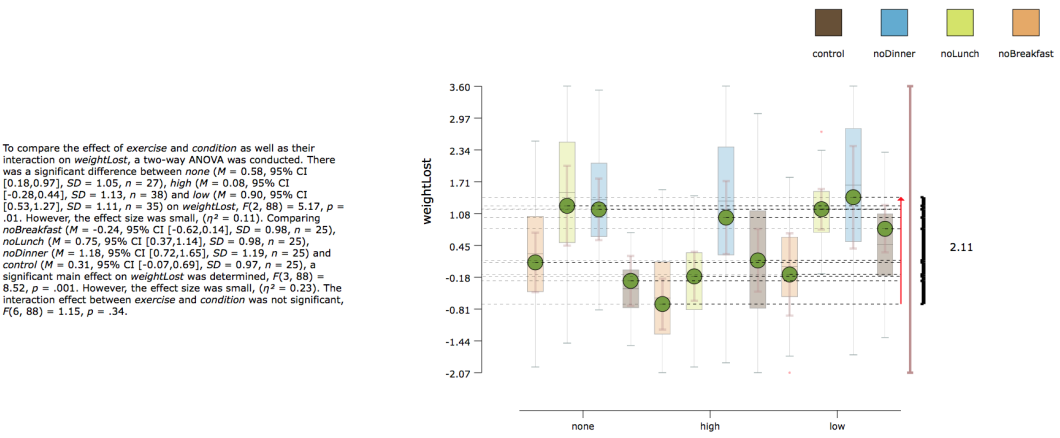


Figure 3.8: Example reporting text for a two-way ANOVA

the text. However, there are some slight changes that have to be made apart from the appropriate statistical values. First of all, for between groups designs, like in the example in figure 3.5 the number of participants is indicated for each condition of the independent variable. When a within groups design is used, all participants cover all treatments, therefore the number of participants should only be given once and would be misleading otherwise. Furthermore, in this case, the number of participants has to be stated with an uppercase N because it describes the entire number of participants (Subsection 3.3.1). Moreover, some grammatical obstacles have to be overcome: The opening article "a" in the first sentence must be transformed to an "an" in case of tests beginning with a spoken vocal (e.g. unpaired t-test). Due to the flexible number of conditions of the independent variable, the enumeration of conditions and their statistical values has to be adjusted to this number so that an "and" instead of a comma is inserted before the last condition.

Post-hoc test and interaction effect

As mentioned before, there are different structures for post-hoc and ANOVA test. The pattern for the reporting text of a post-hoc test is shown in figure 3.7. Always following a one-way ANOVA, the post-hoc test makes use of a different choice of words to create a professional text without repetition. Additionally, the second sentence is unnecessary because the statistical values are mentioned in the reporting text for the one-way ANOVA. The information of the

A one-way ANOVA was conducted to investigate the effect of *keyboardLayout* on *speed*. The results indicated a higher *speed* for *Colemak* (M = 38.21, 95% CI [35.27,41.15], SD = 6.71, n = 20) than for *QWERTY* (M = 31.79, 95% CI [28.56,35.02], SD = 7.37, n = 20) and for *Dvorak* (M = 27.34, 95% CI [24.42,30.27], SD = 6.67, n = 20). A significant difference could be reported, $F(2, 57) = 11.82$, $p = .001$. However, the effect size was small, $\eta^2 = 0.29$.

Figure 3.9: Example reporting text for a one-way ANOVA investigating the effect of different keyboard layouts on typing speed

first (aim and method) and third (significance) sentences are merged instead. The phrase outlining the effect size is still produced in the same way as for the other tests due to the various possibilities of this sentence (cf. figure 3.6). The text for the two-way ANOVA is divided into three paragraphs to characterize the three effects described [Groeben and Vorderer, 1982]. For each independent variable the results are reported in one paragraph, as it is illustrated in figure 3.8. Creating a fluent text, the sentence structure is different for the two effects. Eventually, the interaction effect of both variables and its significant result as well as effect size are stated in the last paragraph.

Comprehensibility of reporting text

In the previous Section 3.3.1, Flesch's Reading Ease [1948] was introduced as a simple possibility to measure the readability or comprehensibility. Calculating this value, the reading ease for the example in figure 3.5 amounts 43.14 when not considering the statistical values. For another example in VisiStat (cf. figure 3.9), the reading ease scores about 52.87 showing the high dependency on the complexity of the variables. However, the insertion of statistical values lowers the readability in addition. Comparing this value in relation to others, the Time Magazine readability index amounts 52 as well, whereas the Harvard Law Review scores about 30. As graduate students can understand a text with a readability index of $0 - 30$ without effort, both values seem to arrive at a compromise of scientific language and good comprehensibility [Grossklags and Good, 2007]. Furthermore, the comprehensibility is increased by a supplementary figure of the corresponding box plot diagram providing a good impression of the results [Groeben and Vorderer, 1982]. The use of a diagram to illustrate the re-

sults is also a characteristic of result sections (cf. table 3.2).

Summary

Summing up, Cairns [2007] found out that insufficient reporting is the most frequent problem in HCI research. The automatic generation of an appropriate reporting text complements VisiStat so that it addresses all four common mistakes mentioned by Cairns and prevents users from committing the same mistakes. How this version of VisiStat can be used to help improving statistics learning in HCI, is investigated in the following chapter.

Chapter 4

Evaluation

Chapter 2 has explained the problem of inappropriate use of statistics among HCI researchers. But how can the use of statistics in HCI research be improved? As already mentioned in Chapter 3, literature suggests to make use of interactive learning systems which facilitate and complement statistics learning. The statistics system VisiStat, which has been introduced in Chapter 3.2, might be such a learning tool. In order to evaluate in how far VisiStat can actually improve and complement traditional statistics learning, a large-scale user study has been conducted as part of the class *Current Topics in Media Computing and HCI*. Investigating how students learn with VisiStat, several research methods have been combined to observe students behavior on the one hand and test their statistical knowledge on the other hand as well as asking how they evaluate their learning experience. Section 4.1 shortly describes the different evaluation methods and gives on overview of the whole experimental design. The gained results from the user study are presented Section 4.2. Based on these results, possible interpretations are discussed and following, alternatives for action are explained (Section 4.3). Completing this chapter, Section 4.4 describes limitations of the different methods and the user study and reflects occurring difficulties.

4.1 Method

To investigate how the interactive statistics system VisiStat complements traditional learning in a lecture, several research methods were mixed, aiming to examine statistics learning from different points of view. The following section has a closer look at these methods and its use in the user study. First of all, an overview of the entire the user study and its experimental design is given. After this, the four methods, knowledge test, user test, feedback questionnaire and interview, are presented in detail, followed by a description of the procedure. Further down the line, the evaluation techniques for each methods are reported. This section concludes with a depiction of the user study's sample concerning their previous statistical knowledge estimation, experience and learning behavior. In the first subsection the user study is introduced.

4.1.1 Experimental Design

AB/BA Cross-Over Study

The entire user study is based on Schneider et al.'s [2013] user study design and therefore, makes use of an AB/BA cross-over study. An overview of the user study is depicted in the image 4.1. In order to find out, how VisiStat complements a traditional statistics lecture, two different orders of treatments were to be tested. The first group (A) follows the preparation for future learning approach (Chapter 2.4) and explores the interactive system first without any prior knowledge, attending the lecture afterwards. In contrast to this, group B is given the traditional tell-and-practice treatment, getting the theoretical knowledge in the lecture and practicing it with VisiStat eventually. It was considered to form a third group which serves as a control group and does not use VisiStat at all but this thought was rejected due to two reasons: On the one hand, as only 36 students participated in the user study, the size of each group would be rather small leading to less comparable results. On the other hand, students should have the same requirements of statistical knowledge for the final exam to exclude unfairness.

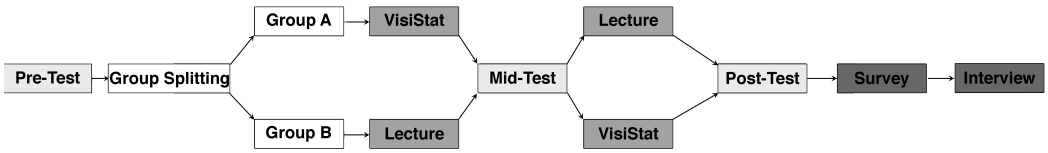


Figure 4.1: User Study Experimental Design

Splitting into Treatment Groups

However, to make sure that the results of both groups are not influenced by the prior statistical knowledge, group splitting was based on the former expertise. Therefore, students were asked to fill out a pre-test, which assesses their statistical knowledge before the user study. Dependent on these results, students were split into four knowledge groups (from low to very high) and were asked to find a team partner within their knowledge group. Afterwards, each team was randomly assigned to one of the two treatment groups taking care that each knowledge group is represented in the same number in each of the two groups A and B. Consequently, our user study followed a between-groups design. A more detailed description of the pre-test results can be found in Chapter 4.2.1 whereas Section 4.1.8 gives a detailed overview of the procedure.

User Test and Lecture

After the groups were announced, each team of the first group was asked to explore the interactive statistics system together by answering some tasks which are similar to HCI research questions (Section 4.1.4). In the following week, students of both groups attended the statistics lecture together. While students of group A have already received two treatments and finished this part of the user study, group B participants were now asked to practice the same tasks as group A with the statistical analysis system VisiStat. During both treatments, students were observed, and in case of the use of VisiStat, recorded for later analysis.

Statistical Knowledge Tests

To be able to reveal whether students' statistical knowledge improves in the course of the user study, a mid-test was set after the first treatment (group A: VisiStat, group B: lecture). Moreover, participants were asked to fill out the post-test after the second treatment (group A: lecture, group B: VisiStat). These tests, which are described in Section 4.1.3, are isomorphic to each other and consist of questions of dif-

ferent levels of knowledge dealing with the four main problem defined by Cairns [2007].

Feedback

In addition to the testing of their statistical knowledge, students were also asked to give feedback about their learning experience during the user study. Thus as the last step, they filled out a feedback questionnaire and were asked several questions in an semi-structured interview, which is described in Section 4.1.7. In order to compare different requirements, like course of study or attendance of statistical lectures before, students gave details about their demographic background. Sections 4.1.6 and 4.1.7 will discuss the questionnaire and interview in detail.

In the following subsections the different applied methods, shortly described in this paragraph, are presented in more detail. At first, the statistical knowledge tests are described.

4.1.2 Hypotheses

Based on Schneider et al.'s work [2013], we derived two hypotheses, which are evaluated. These hypotheses are constructed on the assumptions that students have difficulties to learn statistics and current statistical education lacks to overcome these problems (cf. Chapter 2.2). Taking Garfield and Ben-Zvi's learning principles into account, we consider technology which makes use of visualizations and encourages exploring to be able to address these difficulties in learning statistics. Furthermore, it is assumed that students benefit from exploring this technology tool without previous knowledge and developing their own hypotheses about statistical concepts outperform students who attend a lecture first (telling) and then practice with the system. Summing up, the following two hypotheses for this user study are used:

H1: To learn statistical concepts, students will benefit more from using the interactive statistical analysis system VisiStat, which encourages to construct own knowledge, than from attending a statistics lecture.

H2: Students who explore an interactive statistical analysis first, encouraging them to create contrasting cases, will outperform students who learn in a traditional tell-and-practice procedure.

These hypotheses only cover one part of the research questions, defined in Chapter 1. However, for the other questions, an exploratory approach is chosen as one the one hand, literature does not provide hypotheses about neither the strengths and weaknesses of an interactive statistical analysis system nor how these tools can address Cairns four problems. On the other hand, the exploratory approach allows to investigate students' opinion impartially so that they can determine which strengths and weaknesses they found most important. In Chapter 4.3, these hypotheses are evaluated against the background of the results, presented in Section 4.2. At first, the following sections provide an overview of the used methods, starting with the statistical knowledge tests.

4.1.3 Statistical knowledge tests

In order to be able to compare student's statistical knowledge before and after each treatment, tests which record participants' current level of knowledge, were implemented. As proposed by Schneider et al. [2013], students fill out a pre-test to find out their previous knowledge, followed by a mid-test after the first treatment and finally, a post-test after they completed the user study. To ensure that the results are comparable, they are isomorphic to each other [Schneider et al., 2013] and only differ in case of different examples, order and negation, preventing students from learning effects which might occur otherwise. As mentioned in the previous chapter, Garfield [1998] as well as Delmas et al. [2007] developed questionnaires to assess students reasoning and understanding of statistical concepts. However, these tests mainly deal with basic statistics, for example sampling and distributions. The CAOS does include questions regarding the choice of statistical tests, but does neither focus on Cairns' remaining three

problems nor address effect sizes. Due to these reasons, a separate test was developed for this user study.

The tests consist of five parts: The first part deals with general questions testing student's comprehension of basic statistics concepts and terminology like effect size and significance. This part is followed by knowledge of assumptions, appropriate testing, over-testing, and reporting, adapting to the four problems Cairns identified in the use of statistics in HCI research. This makes it possible to analyze participants' progress in these complicated yet important areas. The order of these areas is determined by the avoidance of learning effects of one part on the other. One example is that assumptions have to be named in the appropriate testing part so that this part has to be after the other. Furthermore, the pre-test contains a demographic questionnaire at the beginning asking for students' age, gender and course of study. Being able to comprehend their previous statistical and learning experience, questions concerning their past contact with statistics and their learning behavior are raised as well. In addition, they estimate their own current statistical knowledge in every test. The pre-test can be found in Appendix A; a description of the sample is given in Subsection 4.1.10.

Moreover, the questions can be categorized in different levels of difficulty. Bloom and Krathwohl [1956] developed a taxonomy of educational objectives, which has been widely used and allows to classify test items in six goals of education. In 2001, Anderson et al. revised this Taxonomy overcoming its criticized limitations [Krathwohl, 2002] [Amer, 2006]. Their improvement results in a "cumulative hierarchy" [Anderson et al., 2001] from the easiest category, *Remember*, to the most complicated, *Create*. To master one category, it is necessary to be skillful in all underlying categories. However, the revised Taxonomy allows categories to overlap [Krathwohl, 2002]. Figure 4.2 shows the scale of the classifications as proposed by Anderson et al. [2001]. Additionally, items are arranged on a second dimension, the *knowledge dimension*, which classifies the kind of knowledge students are expected to provide [Krathwohl, 2002]:

- **Factual Knowledge:** The basic elements that students

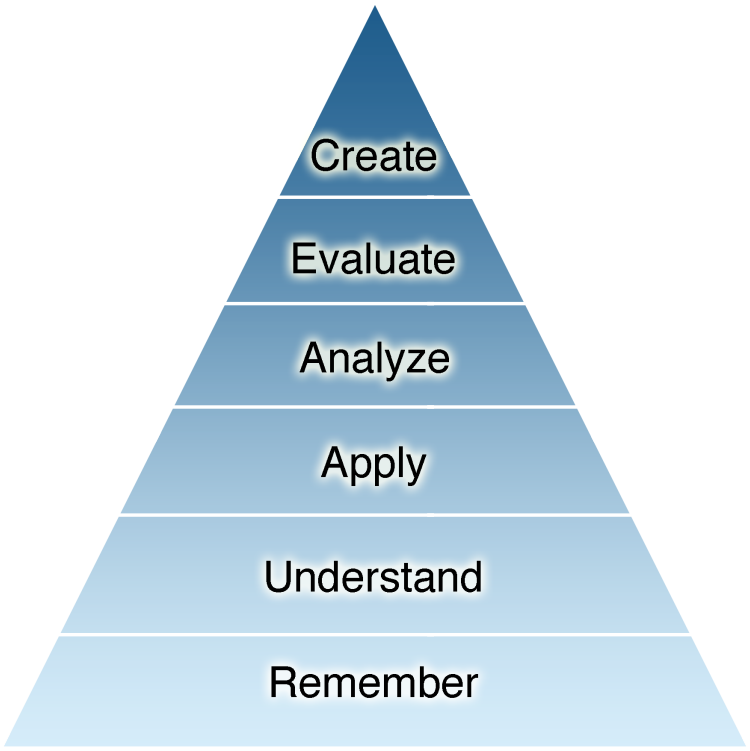


Figure 4.2: Bloom and Krathwohl's Revised Taxonomy - Structure of the Cognitive Process

must know to be acquainted with a discipline or solve problems in it

- **Conceptual Knowledge:** The interrelationships among basic elements within a larger structure that enable them to function together.
- **Procedural Knowledge:** How to do something; methods of inquiry, and criteria for using skills, algorithms, techniques, an methods.
- **Meta-cognitive Knowledge:** Knowledge of cognition in general as well as awareness and knowledge of one's own cognition.

All test items can be organized in Krathwohl's Taxonomy Table as pictured in 4.1. Section 4.2.1 analyzes if students reach different levels of expertise for particular questions

in the taxonomy table. At first, the following subsections presents the other inserted methods besides the statistical tests, starting with user test and lecture.

	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual Knowledge	1G, 1A, 2T, 2R 19.5 points	1G 1 point				
Conceptual Knowledge		2G, 2T, 1O 6 points		1G 1 point		
Procedural Knowledge		1A, 3T 4 points	1R 1 point		3A, 6T, 1O 10 points	1R 10 points
Metacognitive Knowledge						

Table 4.1: Taxonomy Table for Tests (G = General Questions, A = Assumptions, T = Appropriate Testing, O = Over-testing, R = Reporting)

4.1.4 VisiStat condition: Task and observation method

The central part of the user study are the two treatments aiming to improve students' statistical knowledge: On the one hand students attended a traditional lecture, on the other hand they explored statistics individually with an interactive analysis system. The following subsection deals with this user test, whereas the following subsection provides a short overview of the lecture.

Usability for e-learning

Like every software designed for specific users, e-learning systems have to provide a good *Usability* to really be an effective tool for learning [Ardito et al., 2006]. However, the principles defined for an adequate usability for other software have to be extended for e-learning. Zaharias and Poylymenakou [2009] claim that besides principles for web usability, effective engagement as well as motivation to learn have to be addressed. Furthermore, the importance of accessibility and didactic are stressed by Ardito et al. [2006]. Taking these criteria into account, the usability - as proposed for e-learning - of VisiStat complementing of a traditional lecture is investigated by conducting a user test with a retrospective interview and questionnaire.

User Test

Nielsen [1994] describes user tests with actual users as the most crucial method for evaluating usability, naming it even "irreplaceable". For the user study described in this Bachelor's Thesis the user test with the interactive statistics system VisiStat is only one part of the treatment, as the lecture serves as a second learning method. However, the user test allows to gain a direct insight how students actually interact with VisiStat on the one hand, and serves as learning source on the other hand.

Co-discovery Learning

To be able to understand how participants use the system and why they show a specific behavior, they are asked to team up with a partner and solve the given problems together. This method, called *co-discovery learning* or *constructive interaction*, yields in the advantage of natural comments on the system [Nielsen, 1994]. Although Cotton and Gresty [2006] found out the think-aloud method is an appropriate research method for e-learning, a main obstacle is still the uncomfortable situation for many people, especially for a long user test [Nielsen, 1994]. Moreover, thinking aloud might have an impact on the learning experience differing from usual learning behavior. Therefore, co-discovery learning was chosen to eliminate these pitfalls. On the other hand, the main disadvantage of this method is that students might have different learning approaches as well as knowledge levels. However, this could be addressed by letting students choose themselves with whom they want to team up as long as the partner shares the same knowledge group.

Test tasks

To establish a situation as natural as possible, subjects were given several tasks which resemble research questions used in ordinary exercise sheets. They were asked to focus on these tasks and answer them all. As Nielsen [1994] describes the basic rule for test tasks to be that "they should be chosen to be as representative as possible". Jakobs and Lehnen [2005] add that these tasks have to be prototypical for the target group and illustrate prototypical user scenarios. The exercise sheet consists of five main tasks, split into two to five subtasks. Each task deals with one data set, the fourth data set being used twice. Trying not to distract students, the data sets were shortened but still provide the possibility to try different settings than the tasks

Figure 4.3: Room set up for user tests

require. The tasks aim to start with more easy concepts getting more difficult. Additionally, questions which deal with results like significance level and effect size are raised but students were not pushed to focus on particular features. Instead, it is intended that they explore the system on their own. While discovering the first data set, students perform a t-test and a one-way ANOVA test with the same dependent variable and create a report afterwards (in case they do the tasks correctly, as they were only given research questions). The second tasks enables them to contrast paired and unpaired t-test, then create a report for both, and observe the outcome in case one assumption is violated. During these first two tasks they are asked to mark whether the results present a significant difference (first and second data set) and report the effect size (second data set). After this, they practice with the third data set for the first time, conducting a one-way ANOVA and creating a report. This part is followed by the use of a one-way repeated-measures ANOVA, which is complemented by a post-hoc analysis as well as the production of a further report. Returning to the third data set, they can focus on different test types dependent on the assumptions and finally, conclude with a two-way ANOVA. The complete task sheet can be found in Appendix A.

Recording

Apart from users' own impressions about the learning experience, they were also observed and recorded to study their behavior. The setting of the room is shown in figure 4.3. During the user test, a participatory observation took place, enabling the observer to take notes about participants' behavior without interacting with them to prevent unintentional effects. However, in case students had severe problems concerning use of the system or in rare cases statistical concepts, they could be helped by the observer. Furthermore, the user test was screen-recorded complemented by audio recording enabling to retrace their steps for each of the tasks. An over-the-shoulder camera was set in to document hand gestures. How these observations are analyzed, is presented in Section 4.1.9. Beforehand, it is given a short overview of the second learning treatment, the lecture.

4.1.5 Lecture

In contrast to Schneider et al.'s user study, lecture and system do not provide exactly the same information but use their different strengths. Although they have the same content general, students might not have noticed statistical concepts as confidence intervals or over-testing in the system. On the other hand, although creating reports was asked several times in the system tasks it was only shown rather shortly in the lecture. Moreover, the lecture, which lasted for one and a half hour, was longer than students' exploration time of the system, which was about 45 minutes on average (Section 4.1.8). Hereafter, a brief overview presents the lecture, whose slides and recording can be looked up on the Media Computing Group Website[1].

Lecture Overview

The lecture started with a short repetition of basic statistical concepts, focusing on confidence intervals, whose meaning were shown by use of a demo. Afterwards, the importance of effect size was stressed, followed by an example research question. The next part dealt with null hypothesis significance testing, starting with the presentation of t-test. Including students in class, exercises about p-value

[1] http://hci.rwth-aachen.de/cthci

were posed and a further example research question was answered. Subsequently, the assumptions for parametric significance tests were introduced contrasting parametric and non-parametric tests. The next part showed statistical analysis methods for different experimental designs in a decision tree. Starting with between versus within groups design, the tree was extended step by step over number of levels of the independent variable to number of independent variables without going into the details of each test. However, the concept of ANOVA was considered in especially. Furthermore, the problem of type I and II error was demonstrated and use of post-hoc tests shown. Finally, a couple of notes about reporting were displayed, closing the lecture with reading assignments concerning statistics and a brief summary.

After students received both learning treatments, they were asked to evaluate their experience. Therefore, a short feedback questionnaire as well as a retrospective interview were conducted. These two methods are presents in the following two subsections.

4.1.6 Feedback Questionnaire

The statistical knowledge tests show whether and in which parts students improved during the course of the user study. However, they do not explain *why* students improved or did not improve and whether their statistical knowledge increased by the lecture or the exploration of VisiStat. To investigate these questions, the final part of the user study consisted of students' feedback, which was raised in a feedback questionnaire and an interview. This subsection provides an outline of the feedback questionnaire and its advantages for the user study design whereas the next paragraph deals with the interview.

Advantages of Questionnaires

Questionnaires are the most frequently used method in HCI research because they offer many advantages and can be easily conducted. One strong advantage appears to be the ability to "capture the 'big picture' relatively quickly" [Lazar et al., 2010]. Therefore, the results allow to get

an overall impression of users' personal satisfaction concerning the entire learning experience and the exploration of VisiStat in particular. This enables to detect problems as well as benefits subjectively perceived by the students [Nielsen, 1994]. The questionnaire method also holds some drawbacks, which will be discussed in Section 4.4. Due to the widespread distribution of questionnaires, several approved questionnaires have been developed in HCI research. One of these templates is the Technology Acceptance Model by Davis [Davis Jr, 1986], which is presented in the following part.

Technology Acceptance Model

A system, how brilliant it might be in the developer's opinion, is absolutely worthless in case the user does not use it. But when does a user actually use a system and how can this be found out? The question has been addressed by Davis [1986], developing the Technology Acceptance Model (TAM). The original Technology Acceptance Model can predict the actual use of a system by investigating the user's intention to use the system. The intention to use though is determined by the perceived usefulness and the perceived ease of use [Davis, 1989]. Translating this concept to the area of e-learning and the here presented user study, perceived ease of use describes the participant's belief that VisiStat can be used without cognitive effort. On the other hand, the perceived usefulness refers to the student's opinion that his or her statistical knowledge as well as exam grade in the course benefit from using the system [Saadé et al., 2007]. TAM has been approved on the one hand and on the other hand advanced and contributed to by several researchers, as by Saadé and Bahli [2005], Volery and Lord [2000] as well as Yi and Hwang [2003], adapting it to the area of e-learning and information systems. Focusing on general information technology usage, Agarwal and Karahanna [2000] investigated the influence of enjoyment and fun on technology acceptance. Moreover, investigating learners' satisfaction with an e-learning system, Sun et al. [2008] found out that among others perceived usefulness as well as perceived ease of use form critical factors for successful e-learning. The feedback questionnaire is based on TAM and its e-learning variations, which are demonstrated in detail in the following paragraphs.

Questionnaire structure

As the questionnaire is followed by an interview, it exclusively consists of standardized closed-ended questions using a seven point likert scale like the original TAM model. It is divided into seven sub-parts, of which five deal with VisiStat making use of TAM and one for each lecture and overall learning experience. The entire questionnaire can be found in Appendix A. Starting with one of TAM's core concept, the first part approaches the usefulness of VisiStat to learn statistics. Therefore, the first four items are taken from the original TAM-Model [Davis, 1989], slightly adopted to a learning-tool as suggested by Volery and Lord [2000] as well as Yi and Hwang [2003]. The fifth and sixth question are derived from original TAM model but separated into two single items. These questions are complemented by a seventh item proposed by Saadé and Bahli [2005]. The second part collects participants' perception of the ease of use, which operates with a subset of the original TAM questions [Davis, 1989]. This reduction seems to be reasonable as other papers reduce the number of items as well and this selection (or parts) of items also appears in Saadé and Bahli [2005], Yi and Hwang [2003], Agarwal and Karahanna [2000] as well as Volery and Lord [2000]. Additionally, as VisiStat has already been tested for its usability, this part is not the main focus of the user study.

TAM model questions

Extension of TAM model core components

The original main parts of TAM model are completed with questions concerning enjoyment, temporal dissociation and focused immersion. Saadé and Bahli [2005] as well as Agarwal and Karahanna [2000] stress the importance of enjoyment for the attitude toward a system. Yi and Hwang add that enjoyment has a positive effect on usefulness. As enjoyment encourages students to learn [Agarwal and Karahanna, 2000], it can be assumed that students would spend more time on learning statistics when they have fun using VisiStat. The four items, established in the questionnaires, are developed by Davis [1986], Saadé and Bahli [2005], Yi and Hwang [2003], and Agarwal and Karahanna [2000]. The concept of cognitive absorption, presented by Agarwal and Karahanna [2000], includes the dimension of enjoyment and replenishes it with temporal dissociation and focused immersion. This approach is described to be holistic indicating the influence of enjoyment and the perception of time for the use of a system [Saadé

and Bahli, 2005]. Two questions of each dimension, as used by Saadé and Bahli [2005], form the forth part of the questionnaire.

Overall evaluation

Davis' [1986] original questionnaire includes an overall evaluation, which is taken for the last three sections of the questionnaire, giving the participant the possibility to evaluate the lecture and VisiStat as well as their interaction. These items focus on the transfer of statistical knowledge of both learning periods. Furthermore, subjects are asked if they used VisiStat for exam preparation (modified from [Saadé and Bahli, 2005] and [Volery and Lord, 2000]) and recommended the system to others (modified from [Saadé and Bahli, 2005]). In the final part, students are demanded to rate the whole learning experience and how lecture and system complement each other.

As mentioned before, the questionnaire is only one component of the collection of feedback. The other method, an interview, is outlined in the following subsection, closing the method presentation.

4.1.7 Interview

> *"Direct feedback from interested individuals is fundamental to human-computer interaction (HCI) research"*
>
> *—Jonathan Lazar et al. [2010]*

Advantages of interview

The feedback questionnaire described in the previous section offers a quick and wide overview over students' opinion. However, the disadvantage is that it does not give any information why participants obtained this opinion or which parts they liked and which they did not. Addressing these limits of questionnaires, the last step of the user study involved a semi-structured interview. This retrospective interview reveals the opportunity to go into more depth and detect users' reasons for their behavior in the user test, their test results and opinion in the questionnaire. Therefore, the interviewer can ask further questions to fully under-

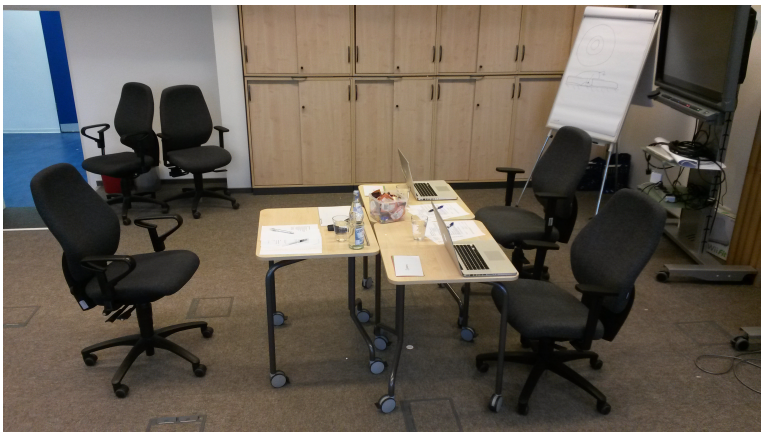


Figure 4.4: Room set up for interviews of group A

stand the participant and determine, what he or she expects of an interactive statistics system as VisiStat [Lazar et al., 2010]. Due to organizational reasons, students are interviewed with their team partner together. However, an advantage is that students can discuss about their opinion or improvement suggestions as well as think about more issues showing their reasons. To be able to analyze the interview afterwards, the session was audio-recorded. The setting of the interviews for group A can be found in figure 4.4, for group B, the interview was conducted in the user test room with a similar structure. In the following paragraph, the interview structure and questions are briefly introduced.

Interview Structure

Interview questions vaguely base on Schneider et al.'s [2013] as well as Naps et al.'s [2002] retrospective questioning. However, they were elaborated and items concerning Cairns' [2007] problems in statistics were added. The interview roughly contains five parts of which the second part is marked as optional and was not asked due to time constraints. In the first part, students are asked to evaluate lecture and statistics system. At first, questions are be raised open-ended so that participants can freely tell about their experience. If necessary or interesting, more questions to encourage students or clarify certain aspects are asked. Afterwards, interviewees evaluate the interaction and roles of lecture and VisiStat. In the third part, students

are presented a research scenario, investigating whether they know a statistics concept and where (lecture or system) they have developed this knowledge. This scenario is extended in every step to cover all four statistics problems mentioned by Cairns [2007]. In part four, participants imagine themselves as a teacher for statistics, encouraging them to suggest possible improvements to VisiStat and lecture. The final part consists of yet another open-ended question enabling interviewees to add any other thoughts. The interview session represents the last part of the user study.

Summing up, all research methods have been described and their appropriateness for the user study has been shown. The following subsection deals with the procedure of the user study. Afterwards, the methods of evaluation will be presented shortly describing how the results in the next section are analyzed.

4.1.8 Procedure

After the four different methods for the user study have been introduced, the following section briefly presents the procedure how the different methods were combined and how the data was actually raised. The entire user study took place from May 7 to June 12, starting with a short presentation about the problem of learning statistics in the lecture Current Topics of HCI. Students were given the possibility to choose between participating in the user study or completing an exercise about statistics as part of passing the course. An overview of the user study and the current task to do was given constantly on the Media Computing Group Website[2].

Pre-test and group splitting

In a first step, participants were asked to fill out the pretest online at home within one week. The results were directly calculated and students split up into four knowledge groups dependent on their score in the test. In the next lecture, the four groups were shuffled and presented without telling students the meaning of the groups so that they did not know how well they performed. Ensuring that students

[2] http://hci.rwth-aachen.de/statstudy

only pair with a partner of comparable statistical knowledge, each participant was asked to team up with someone in her or his knowledge group. After teams were matched, they were randomly assigned to group A or group B.

Group A User Test

From May 19 to 26, group A students participated in the first part of the user study. In a test room, they were given a short explanation and overview of the user study by the experimenter, followed by a video introducing the system. This introduction video briefly described VisiStat's components and showed a typical usage of a significant test and following report. Afterwards, students were asked to explore the system on their own, solving all exercises described in Section 4.1.4. Users were allowed to use the system for fifty minutes at most and were asked to proceed in case they stuck with an exercise and ran out of time. In case students finished earlier, they were offered the possibility to practice with the system on their own. During the VisiStat session the leader of investigation was continuously present in the room so that students could be helped if necessary. However, students were asked not to pose questions concerning statistical content. Furthermore, the leader of investigation informally observed the exploration. To be able to analyze the experiment afterwards, the students' interaction as well as the screen were recorded. After students finished exploring the system, they filled out the mid-test separately and thanked for their participation.

Lecture

The statistics lecture took place on May 27, 2014 and lasted for one and a half hour. Three students, who could not attend the lecture, were given the possibility to watch a video of the lecture the following day. The leader of investigation informally observed the lecture as well so that students' feedback in the interview session could be understood easily. After the lecture, students were asked to fill out either the post-test (group A) or the interim test (group B) at home until the next day preventing influences due to the different times of answering.

Group A Interview

Group A students were asked to return for a retrospective interview after they attended the lecture to evaluate their overall learning experience. First of all, they answered the feedback questionnaire on their own, which did not took

longer than five minutes. As a second step they were interviewed together for approximately twenty to thirty minutes. However, many students wanted to give more feedback and discussed even longer. After the interview, students from group A completed the user study.

Group B User Test and Interview

Participants from group B performed the same steps as group A students. One difference ist of course that they filled out the post-test after the exploration of VisiStat. In contrast to group A participants, they did the interview directly after the post-test because they already covered both treatments. Eventually, they were thanked for their participation ending the user study.

After the user study was successfully executed, the results have to analyzed. The following section presents how the different data is evaluated describing the different methods for statistical test, observation, questionnaire, and interview.

4.1.9 Methods of Evaluation

Inferential statistics for test evaluation

In this section, the different methods for evaluating the collected data are briefly described. The correctness of students' answers in the three statistical knowledge tests was checked on the basis of the grading sheet, which can be found on the attached DVD. Afterwards, the data was processed and analyzed with the help of *IBM SPSS Statistics*. To compare the two independent groups, unpaired *t*-tests were used in case of homogeneous variances and a normal distribution. In case of heterogeneous variances, which were determined with Levene's test, an unpaired *t*-test as well as a Mann-Whitney-U-test were performed. If the data was not normally distributed, Mann-Whitney-U-tests were conducted. We assume significant differences between the two populations from a level of $p < .05$. Regarding the effect size, Cohen's d was calculated for normally distributed data and the coefficient for determination for non-parametric tests. If the variances were approximately the same, we used the standard deviation of the control group (group B) to calculate the effect size as recommended

by Field [2013]. If this was not the case, the pooled within standard deviation was computed [Gravetter and Forzano, 2011]. As two participants had a significant higher previous knowledge, their results were left out in the evaluation to prevent distortion in group B due to their scores. A similar evaluation approach was chosen for the feedback questionnaire. Therefore, the average mean of all items of one dimension was calculated, treating the data as interval scale similar to Saadé et al. [2007].

Feedback questionnaire evaluation

Coding of utterances

Furthermore, we asked students to discuss the procedure and results during their exploration of VisiStat as if they were doing an assignment sheet cooperatively. In a first step, two videos of group A and one video of group B were analyzed regarding their utterances to get a first impression. Based on Schneider et al.'s [2013] categorization scheme of students' statements while exploring a tabletop for learning neuroscience, we investigated if a utterance was an observation, prediction, confrontation, or general rule of statistical concepts. Investigating differences between group A and B and their use of the help function, the ratio of using help for each team was calculated. As a result, the utterances were categorized according to the coding scheme described in table 4.2.

Category	**Explanation**
Observation	"And now we have better effect size than we have earlier"
Prediction	"I took [the data] without [transformation]. So [that] when they used Welch's ANOVA"
Confrontation	"Let me [check the help on the ANOVA]", "So we go back and then we check? we transform it"
Rule	"Yeah, we have ANOVA here, so that the effect size was Eta-squared."

Table 4.2: Coding of students' utterances during the exploration of VisiStat

Regarding the interviews, we used the Grounded Theory procedure, developed by Glaser and Strauss [2009], to eval-

uate students' feedback. Therefore, students' utterances were analyzed and taken to pieces. These pieces are examined in a second step and assigned to categories in several iterative phases. In contrast to deductive approaches, these categories are developed recursively and build upon the underlying data. Thus, the strength of this procedure is to develop theories about participants opinion based on the data and without prior bias [Breuer, 2009]. As closing part of this method section, the participants of the user study are presented in the following subsection.

4.1.10 Participants

The user study's population consists of all students attending the class *Current Topics of HCI* offered by The Media Computing Group of the RWTH Aachen in the summer semester 2014. Of those 39 students 37 agreed to participate and thereby form the elements of the sample. As one student's previous statistical knowledge was remarkably higher than the others', a student assistant with a comparable knowledge was asked to participate in the user study as well to team up with this participant. Having performed every step of the user study similar to the other students, he will be treated as any other respondent in the following. Furthermore, two students dropped out during the course of the user study so that the final sample is composed of 36 participants, who completed the whole user study. These 36 participants include the two participants with significant higher previous knowledge, which are not regarded in the results.

Previous knowledge level

As explained in Chapter 4.1.1, students were split into two groups with different order of statistical learning treatment. To ensure that potential divergent results of both groups are not influenced by different levels of previous knowledge, four groups of students were set up dependent on their present statistical expertise, calculated on basis of their pre-test results. In the next step, participants of the four knowledge groups were divided almost equally into the two treatment groups A and B. Group A consists of 16 students from which four students belong to knowledge

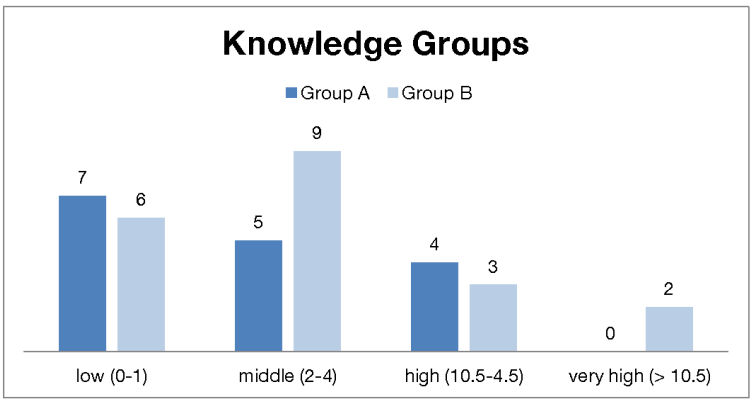


Figure 4.5: Distribution of participants across knowledge groups in treatment groups

group *high*, five to knowledge group *middle*, and seven participants with *low* previous statistical knowledge. There are 20 participants who are part of group B. This group can be split into six students with *low* previous knowledge and nine whose expertise amounts to *middle*, whereas three students have a *high* and two students even a *very high* knowledge at their disposal (cf. figure 4.5). Further information about the division of the different knowledge groups are presented in Chapter 4.2.1. The different numbers of students in the two groups A and B is attributed to organizational reasons (group B's test dates were later than group A's) and the fact that only one team has a *very high* previous statistical knowledge. Chapter 4.4 gives an overview about limitations of the user study.

Demographic background

Among the 16 participants in group A four are female and 12 are male. They are between 23 and 29 years old with an average age of 25.25 ($SD = 1.65$). Furthermore, students come from four different courses of study apart from one additional Erasmus student (cf. figure 4.6). Their current semester of studying amounts to 3.13 on average ($SD = 1.26$) being in their second to fifth semester (the typical duration of a Master's program amounts to four semesters, for a Bachelor's program it is six semesters). In group B six female and 13 male respondents can be found, an additional test person states to be of an other gender. For this group, the average age comes to 24.45 ($SD = 2.74$) whereas

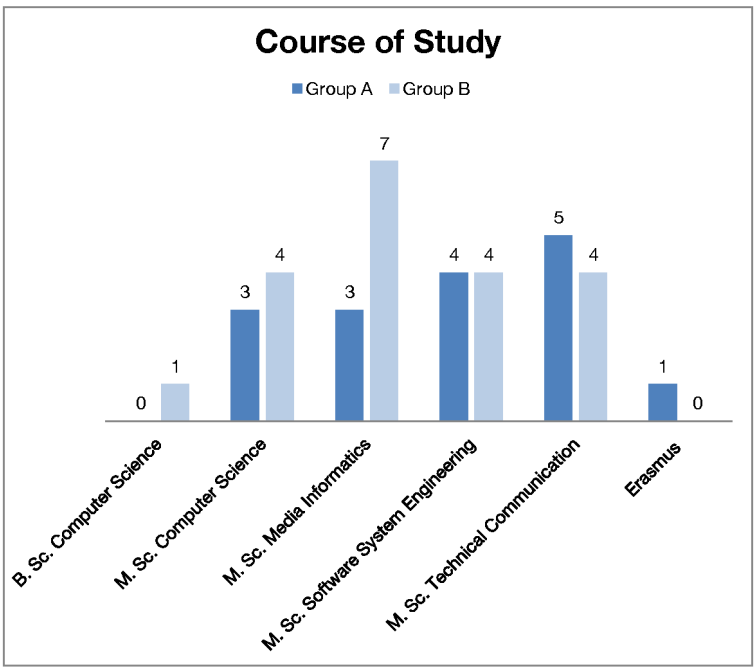


Figure 4.6: Courses of study in both groups A and B

students' age runs between 20 and 29 years. It has to be noted that one participant claims to be "> 28". In order to make following calculations easier, this notation is saved as 28 years. Most participants study Media Informatics, the master studies in Computer Science, Technical Communication, and Software System Engineering are represented by four students each. In contrast to group A the current semester of studying runs between the first and sixth semester with an average semester of 2.75 ($SD = 1.29$).

Previous statistical experience

Apart from the demographic background, students were also asked about their previous statistical experience and learning behavior. Students from both groups developed statistical knowledge before, most of them at more than one opportunity. Two thirds of participants learned about statistics in school, 32 students attended at least one university lecture, teaching statistics. Only three persons read books about statistics. Furthermore, nine students have already used statistics in a seminar work, thesis or paper whereas twice as many students are in group B. One stu-

dent alone used an interactive statistics learning system before. An overview of participants' statistical learning methods can be found in figure 4.7. Although students experienced statistical knowledge before, they estimate their statistical knowledge rather low as half of group A's test persons described it as low. Another 31.3% even state to have a very low expertise and only 18.8% think they have a middle statistical know-how. In group B students are slightly more confident with 45% describing their knowledge as middle. In this group, 40% estimate their expertise low and 15% as very low.

Learning behavior

When asked about their preference of learning in their last learning situation, nine group A students prefer understanding the theory first and then practicing whereas six state to do practical application first followed by learning the theory. Additionally, one student indicates a combination of both learning basic theory at first, then apply this knowledge and understanding in a third step. In group B, understanding the theory first followed by practice is favored by 12 participants. Similar to group A less students (for group B seven) learn by practicing first and then have a look at the theory behind it. Doing both is claimed by one further student.

In what way these characteristics of the sample, especially the actual previous statistical knowledge, influence the students' course of learning experience will be dealt with in the following chapter. This chapter presents the results from the different research methods described before showing the possible improvements in the three different tests. Furthermore, it will be analyzed students' interaction with the interactive statistics system VisiStat as well as a summary of their opinion about the whole learning experience will be given.

4.2 Results

This section presents the results of the in Section 4.1 described empirical user study. As in the previous chapter, the results of the statistical knowledge tests are elaborated

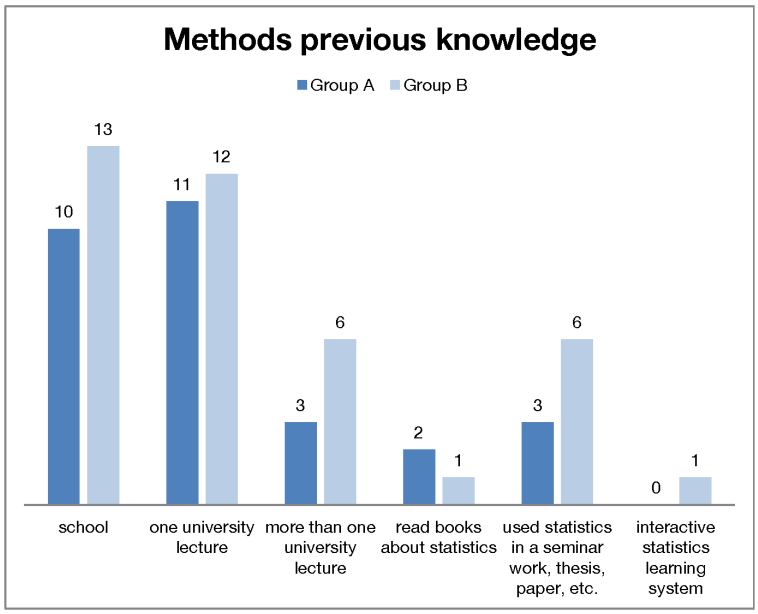


Figure 4.7: Methods students developed statistical knowledge with in groups A and B

at first, illustrating students' progress of knowledge. Afterwards, the observations from the user test are categorized and described to be able to analyze participants behavior towards and attitude with VisiStat. The third part reveals students' opinion of the learning experience, evaluating their answers from the feedback questionnaire. These results are completed by the conclusions derived from the interviews. Answering the research question, it is focused on the differences between the two groups of treatments in all parts.

4.2.1 Statistical Knowledge Tests

The results of the statistical knowledge tests are used to determine students' progress of knowledge in the course of the user study. Special attention is paid to differences between the two treatment groups, investigating if one order surpasses the other. Therefore, three areas are examined: overall test results, test results in five different statis-

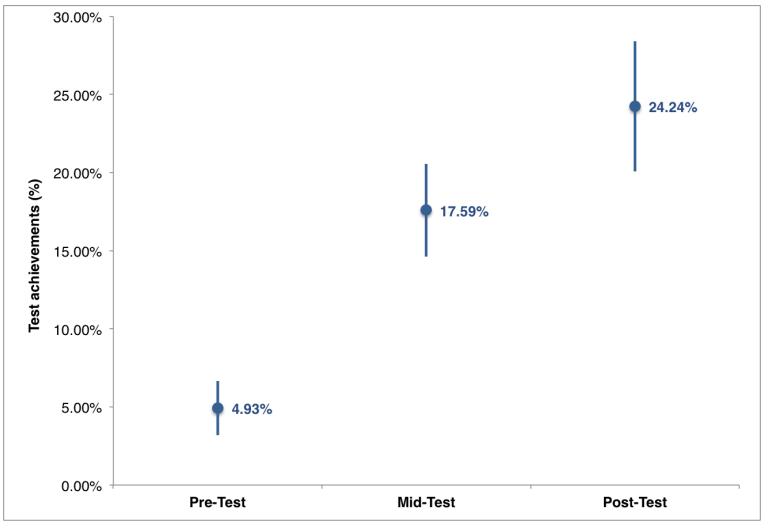


Figure 4.8: Students' development of overall test results from pre-test to post-test

tical concepts (General Questions, Assumptions, Appropriate Testing, Over-testing, and Reporting), and test results for the different learning questions (cf. matrix Section 4.1.4. The section begins with demonstration of the overall results in the three tests.

Overall Results

Comparison pre- to post-test

The overall results of students' tests showed a constant improvement from pre-test over mid-test to post-test on average, as can be seen in figure 4.9. Whereas students in the pre-test scored not even five percent ($M = 4.93$, 95% CI $[3.19, 6.67]$, $SD = 4.99$, $n = 34$), these results considerably improved to 17.59% on average in the mid-test (95% CI $[14.62, 20.57]$, $SD = 8.4$, $n = 33$). However, although students achieved again higher outcomes for the post-test, the results did not exceed a mean of 25% ($M = 24.24$, 95% CI $[20.06, 28.42]$, $SD = 11.79$, $n = 33$).

Differences between groups in mid test results

Evaluating which order of treatments positively influences statistics learning, the average results from group A (VisiStat -⟶ lecture) and group B (lecture -⟶ VisiStat) were

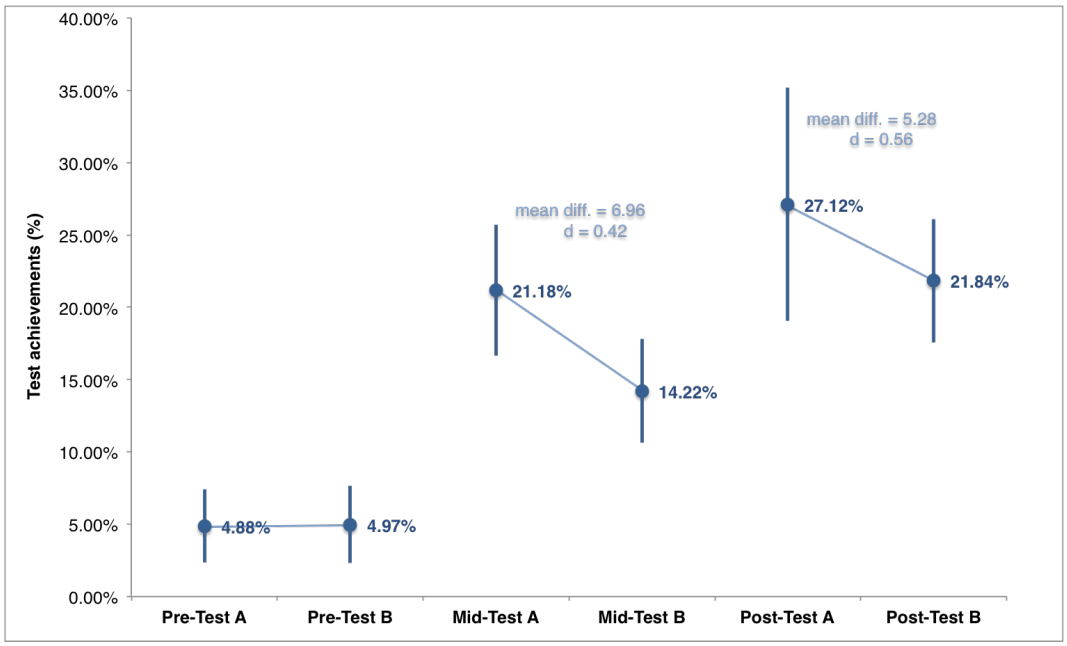


Figure 4.9: Overall test results for group A (VisiStat → lecture) and B (lecture → VisiStat) in pre-, mid- and post-test

compared. As students were split equally after the pre-test into groups dependent on their test results, no differences occurred between group A and B before the first learning treatment. However, after the interim-test higher test results for group A could be observed, revealing a mean difference of 6.96% between the two groups. Participants exploring VisiStat first scored 21.18% on average (95% CI $[16.66, 25.7], SD = 8.49, n = 16$), the students, who attended the lecture first, achieved 14.22% on average (95% CI $[10.63, 17.8], SD = 6.97, n = 17$). An unpaired t-test was conducted, comparing the two groups. A significant difference could be reported, $t(31) = 2.58, p = .015$. However, the effect size was small ($d = .42$).

Differences between groups in post test results

On the post-test, group A outperformed group B again. The difference of means was 5.28 in favor of "VisiStat → lecture" students, which showed an average result of 27.12% (95% CI $[19.05, 35.19], SD = 14.58, n = 15$). Contrary to this, "lecture → VisiStat" participants achieved a score of 21.84% (95% CI $[17.59, 26.1], SD = 8.56, n = 18$). As the assumption of homogeneous variances is violated, an un-

paired *t*-test under the condition of unequal variances (Levene test: $F = 9.653, p = .004$) as well as a Mann Whitney U test were conducted, comparing the results of the two groups. However, both significance tests revealed no significant differences between group A and B for the post-test, $t(21.73) = 1.235, p = .23$ for *t*-test and 0.464 for Mann Whitney U-test ($U = 114.5$). Furthermore, Cohen's d was calculated in two ways, as discussed before in Section 4.1.9. Using control group's standard deviation (traditional learning) for the calculation, amounted in a medium-sized effect ($d = .56$). Yet, it has to be stressed that the control group's SD remarkably differs from group A's standard deviation. The application of a pooled within-groups standard deviation revealed only a small effect ($d = .41$).

Results for each topic

It was described in Section 4.1.3 that the tests consist of five sub-parts, which are adapted to Cairns' four main problems in statistics application in HCI and one part relating to general statistical basic knowledge. Therefore, for each of these five dimensions the results after the first treatment and the second treatment are compared for both groups A and B. These results are summed up in figure 4.10 and are presented in the following paragraphs. Concerning the interpretation of the results, it has to be stressed that each of the five sections consists of a different amount of questions and reachable points. Whereas students can achieve about seven points for the sections general questions and assumptions, the record for appropriate testing amounts to about 13 points and even 23 for reporting. In contrast, in the over-testing part only two points can be scored. To enhance the comprehensibility, the corresponding statistical values for each dimension are stated in tables, which can be found in this subsubsection as well.

General Questions

Analyzing the scores for general questions, a descriptive difference could be identified in the mid-test results as students, attending only the lecture, scored on average nearly 10% more than those participants exploring VisiStat. However, this difference was not significant. After the second

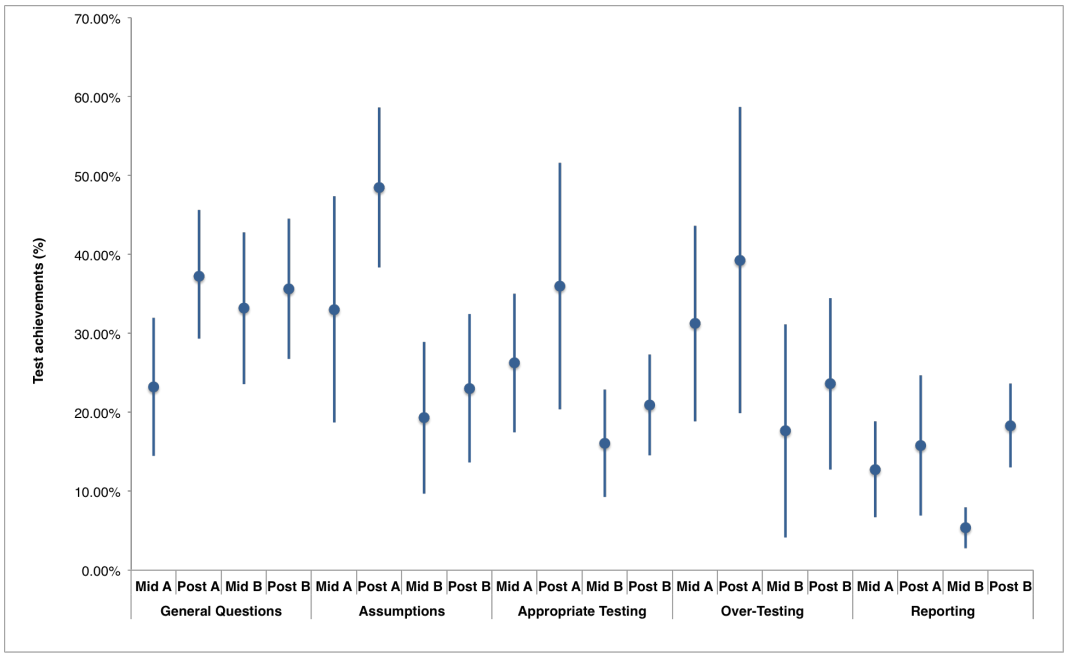


Figure 4.10: Students' test achievements in mid- and post-test for each statistical topic

treatment, group A caught up with group B, achieving even a slightly higher result. The corresponding statistical values can be examined in table 4.3.

	M	95% CI	SD	n	Statistical Test	Test result	p	Effect size
Mid A	23.21	[14.48, 31.95]	16.39	16	Mann-Whitney-U-test	$U = 176$	.157	$r = 0.25$
Mid B	33.19	[23.57, 42.82]	18.72	17				
Post A	37.22	[29.33, 45.67]	13.68	15	unpaired t-test	$t(31) = 0.28$	.781	$d = 0.09$
Post B	35.65	[26.79, 44.5]	17.81	18				

Table 4.3: Statistical results of general questions part in mid- and post-test for group A and B students

Assumptions

After the first treatment, students attending the lecture answered about 20% on average of questions regarding assumptions for statistical tests correctly. In contrast, participants exploring the interactive system could achieve about 33%. Whereas group A students could improve further after the lecture gaining nearly 50%, learners in group B advanced about 3%. A Mann-Whitney U-Test revealed a significant difference $U = 41$, $p = .001$). The differences constituted a large effect size, $r = -0.5775$.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	33.04	[18.74, 47.34]	26.84	16	Mann-Whitney-U-Test	$U = 95.5$	.146	r = -0.26
Mid B	19.33	[9.72, 28.94]	18.69	17				
Post A	48.47	[38.34, 58.59]	17.54	15	Mann-Whitney-U-test	$U = 41$	.001	r = -0.58
Post B	23.02	[13.61, 32.42]	18.91	18				

Table 4.4: Statistical results of assumptions part in mid- and post-test for group A and B students

Appropriate Testing

Concerning participants' test results for appropriate testing questions, the PFL-group outperformed the traditional learning group after the first as well as after both treatments. Whereas a Mann-Whitney U test revealed significant differences for the post-test (cf. table 4.5, an unpaired *t*-test did not detect significant differences. However, large effect sizes could be measured for mid- ($d = 0.771$) and post-test, which is $d = 1.17$ when taking the control group standard deviation and amounts to $d = 0.74$ when calculating with the pooled-within group standard deviation as the variances are not homogeneous.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	26.25	[17.48, 35.02]	16.46	16	unpaired *t*-test	$t(31) = 1.97$	.058	$d = 0.77$
Mid B	16.08	[9.3, 22.86]	13.19	17				
Post A	35.99	[20.39, 51.59]	27.01	14	unpaired *t*-test	$t(17.53) = 1.92$	.071	dc= 1.17
Post B	20.94	[14.57, 27.31]	12.81	18	Mann Whitney U-test	$U = 73$	.045	dp= 0.74

Table 4.5: Statistical results of appropriate testing part in mid- and post-test for group A and B students

Over-Testing

The results for the over-testing section were comparable to the results for appropriate testing, showing better descriptive scores for group A students after mid- and post-test. Participants in group A performed about 33% after the post test, in group B an average score of 23% was achieved. However, neither significant differences nor remarkable effect sizes could be identified.

Reporting

The reporting were generally low, not surpassing 20%. An interesting development could be observed as after only exploring VisiStat group A students significantly outperformed group B students as shown in an unpaired *t*-test

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	31.25	[18.85, 43.65]	23.27	16	Mann Whitney U-test	$U = 95.5$	.146	r = -0.28
Mid B	17.65	[4.12, 31.1]	26.17	17				
Post A	39.29	[19.91, 58.66]	33.56	14	Mann Whitney U-test	$U = 93$	.22	r = -0.23
Post B	23.61	[12.76, 34.46]	21.82	18				

Table 4.6: Statistical results of over-testing part in mid- and post-test for group A and B students

($t(20.4) = 2.388$, $p = .027$). Furthermore, a Mann-Whitney U test was conducted because the variances were not homogeneous, revealing no significant difference ($U = 147.5$, $p = .063$). However, regardless of calculating the effect size with the control group standard deviation ($d = 1.4653$) or with the pooled-within standard deviation ($d = 0.8492$), the differences constituted a large effect size. Despite this difference in the mid-test, after students in group A attended the lecture as a second treatment, and group B participants used VisiStat, a slight difference could be recognized in favor of group B.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	12.77	[6.7, 18.84]	11.39	16	unpaired t-test	t(20.04) = 2.39	.027	dc= 1.47
Mid B	5.37	[2.78, 7.97]	5.05	17	Mann Whitney U-test	$U = 84.5$	.063	dp= 0.85
Post A	15.81	[6.94, 24.67]	15.36	14	Mann Whitney U-test	$U = 147.5$	.419	$r = 0.14$
Post B	18.32	[13.04, 23.61]	10.63	18				

Table 4.7: Statistical results of reporting part in mid- and post-test for group A and B students

Results for the different learning tasks

Apart from different topic, the tasks in the three tests addressed different learning levels as discussed in Section 4.1.3. To examine how students improved in each of these levels, the results for each dimension are represented in the following section. Therefore, the outcomes for group A and group B students are compared after the first and second treatment. In the following, the results are analyzed regarding their knowledge level, beginning with fac-

tual knowledge, then moving on to conceptual knowledge and eventually, the procedural knowledge is considered. Within in each knowledge level, the different degrees of learners' cognitive processes (like remember, understand, evaluate, etc.) are investigated (cf. table 4.1).

Factual Knowledge - Remember

Regarding the factual knowledge, two cognitive process dimensions, remember and understand, are asked for in the questionnaire. Figure 4.12 provides an overview of the results, the statistical values can be comprehended in table 4.8. Whereas students in group A could improve their score in the remember category from mid- to post-test about more than 15%, students in group B scored only slightly better after the second treatment. In the mid-test, students in group B outperformed group A learners by about 3%. Yet, after the post-test, a better result could be observed in favor of the PFL group. Both differences were not significant (table 4.8). However, in the post-test the differences constituted a medium effect size, $d = 0.69$.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	17.88	[12.61, 23.15]	9.89	16	unpaired *t*-test	$t(31) = -0.56$	.58	$d = 0.18$
Mid B	20.1	[13.62, 26.57]	12.6	17				
Post A	31.37	[21.83, 40.91]	16.53	14	unpaired *t*-test	$t(30) = 1.78$	.085	$d = 0.69$
Post B	21.76	[14.79, 28.73]	14.02	18				

Table 4.8: Statistical results of factual knowledge remembering questions in mid- and post-test for group A and B students

Factual Knowledge - Understand

Examining students' understanding of factual knowledge closer, a Mann-Whitney-U-Test revealed a significant difference between group A and group B participants after the first treatment, $U = 198$, $p = .025$. Students in group B could reach a mean of 70.59% (95% CI $[46.44, 94.74]$, $SD = 46.97$, $n = 17$) in contrast to the average score of 25% in group A (95% CI $[1.17, 48.83]$, $SD = 44.72$, $n = 16$). A medium effect size could be measured, $r = 0.45$. While students in group A could improve to over 70%, the results for group B deteriorated about 15% so that the PFL group outperformed the traditional group after the second treatment. This difference was not significant (table 4.9). Furthermore, it has to be stressed that only one question in the statistical knowledge tests could be categorized as a factual knowledge understand question, allowing different inter-

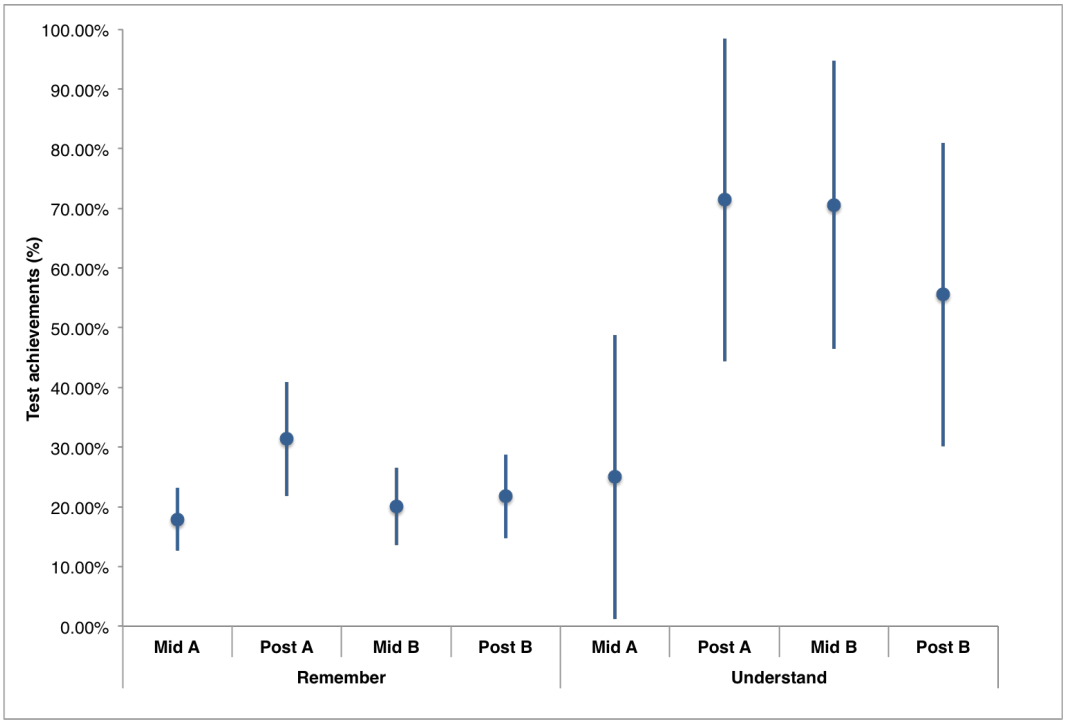


Figure 4.11: Students' test achievements in mid- and post test for factual knowledge

pretations of the results.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	25	[1.17, 48.83]	44.72	16	Mann-Whitney-U-test	$U = 198$	.025	$r = 0.45$
Mid B	70.59	[46.44, 94.74]	46.97	17				
Post A	71.43	[44.36, 98.5]	46.88	14	Mann-Whitney-U-test	$U = 106$	.464	$r = -0.16$
Post B	55.56	[30.13, 80.98]	51.13	18				

Table 4.9: Statistical results of factual knowledge understanding questions in mid- and post-test for group A and B students

Conceptual Knowledge - Understand

On the next level of knowledge, the conceptual knowledge on the cognitive dimensions understand and analyze is studied (cf. figure 4.12. Concerning understanding, the results for both groups were comparably low around 20%. After the first treatment, the results for group A and B were approximately the same (cf. table 4.10). In contrast to the traditional learners, who scored slightly worse in the post test, group A participants could improve around ten percent points. A significant difference could be reported with

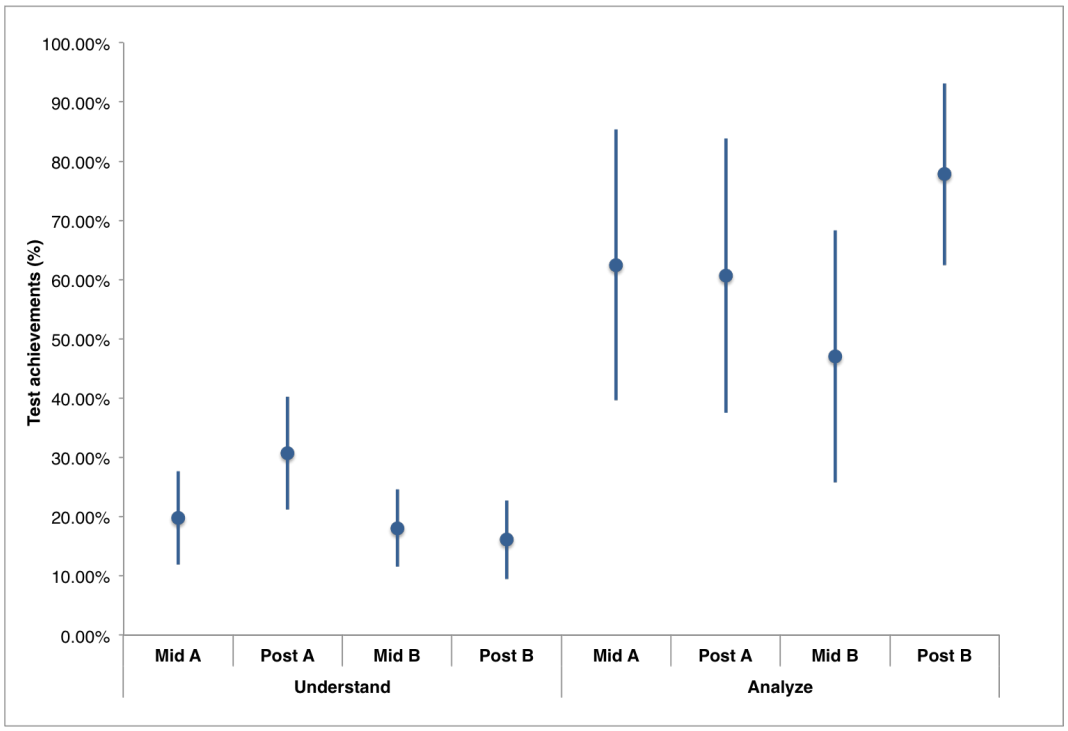


Figure 4.12: Students' test achievements in mid- and post test for conceptual knowledge

the help of a Mann-Whitney-U-Test, $U = 61$, $p = .013$, showing a medium effect size, $r = -0.45$.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	19.79	[11.89, 27.69]	14.83	16	unpaired t-test	$t(31) = 0.37$	.717	$d = 0.14$
Mid B	18.04	[11.51, 24.57]	12.7	17				
Post A	30.71	[21.25, 40.18]	16.39	14	Mann-Whitney-U-test	$U = 61$	.013	r = -0.45
Post B	16.11	[9.47, 22.75]	13.35	18				

Table 4.10: Statistical results of conceptual knowledge understanding questions in mid- and post-test for group A and B students

Conceptual Knowledge - Analyze

For the level of conceptual knowledge, analyzing ability resulted in higher scores. Whereas group A students performed about 60% in both tests with slight deterioration, group B participants enhanced their knowledge about 30%, catching up and even surpassing group A after the second treatment. No significant differences or noteworthy effect sizes were detected. However, this entry again only con-

sisted of one question and therefore, has to be interpreted cautiously.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	62.5	[39.68, 85.32]	42.82	16	Mann-Whitney-U-test	$U = 108$	.326	r = -0.19
Mid B	47.06	[25.8, 68.32]	41.35	17				
Post A	60.71	[37.57, 83.86]	40.09	14	Mann-Whitney-U-test	$U = 155.5$	.267	$r = 0.22$
Post B	77.78	[62.47, 93.09]	30.79	18				

Table 4.11: Statistical results of conceptual knowledge analyzing questions in mid- and post-test for group A and B students

Procedural Knowledge - Understand

Most question are based on the procedural knowledge dimension, whose results are outlined in figure 4.13. At first, students' understanding is examined (cf. table 4.12. After the first treatment, PFL participants outperformed the traditional learners about nearly 10%. This descriptive difference resulted in a statistically significant difference after the post-test ($U = 73$, $p = .045$) as group A students could improve further to a mean of 41.07% (95% CI $[24.46, 57.68]$, $SD = 28.77, n = 14$). Contrary to this, participants in group B achieved an only slightly higher average score of 22.22% (95% CI $[13.81, 30.63]$, $SD = 16.91$, $n = 18$). The differences between post A and post B constituted a medium effect size, $r = -0.37$.

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	28.13	[15.37, 40.88]	23.94	16	Mann-Whitney-U-test	$U = 110$	.363	r = -0.17
Mid B	19.85	[9.43, 30.28]	20.28	17				
Post A	41.07	[24.46, 57.68]	28.77	14	Mann-Whitney-U-test	$U = 73$	.045	r = -0.37
Post B	22.22	[13.81, 30.63]	16.91	18				

Table 4.12: Statistical results of procedural knowledge understanding questions in mid- and post-test for group A and B students

Procedural Knowledge - Apply

Application of procedural knowledge resulted in low scores, not surpassing 10% (cf. table 4.13). It has to be emphasized that again only one question could be assigned to this category. However, it is interesting to note that the development of results differed among group A and group B. Whereas PFL participants deteriorated from mid- to post-test, traditional learners had no knowledge at all in the mid-test but could improve to about 8%, which is the highest average score in this category. Neither significant differences nor remarkably effect sizes could be identified.

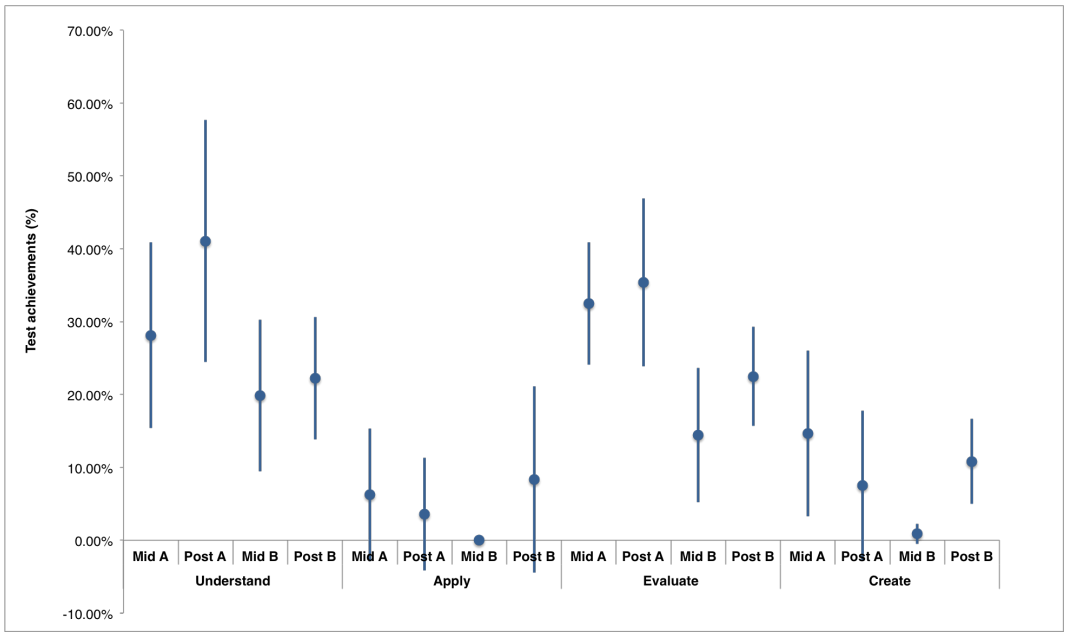


Figure 4.13: Students' test achievements in mid- and post test for procedural knowledge

	M	95% CI	*SD*	*n*	Statistical Test	Test result	*p*	Effect size
Mid A	6.25	[-2.85, 15.35]	17.08	16	Mann-Whitney-U-test	$U = 119$	.557	r = -0.26
Mid B	0	[0, 0]	0	17				
Post A	3.57	[-4.14, 11.29]	13.36	14	Mann-Whitney-U-test	$U = 131.5$	.837	$r = 0.07$
Post B	8.33	[-4.46, 21.13]	25.73	18				

Table 4.13: Statistical results of procedural knowledge applying questions in mid- and post-test for group A and B students

Procedural Knowledge - Evaluate

Regarding evaluation on the procedural knowledge dimension, group A students significantly outperformed group B students after the first as well as the second treatment (cf. table 4.14). Participants exploring VisiStat first achieved an average score of 32.5% in the mid-test, surpassing the lecture participants, which scored about 14%. A Mann-Whitney-U-Test revealed a significant difference in the mid-test, $U = 95$, $p = .003$. Furthermore, a large effect size could be measured, $r = -0.51$. Although students improved after exploring VisiStat, gaining an average result of 22.5%, group A students exceeded them again with an average score of 35.56. A significant difference could be reported with an unpaired t-test, $t(30) = 2.165$, $p = .039$.

For this difference, a large effect size could be detected, $d = 0.94$.

	M	95% CI	SD	n	Statistical Test	Test result	p	Effect size
Mid A	32.5	[24.07, 40.93]	15.81	16	Mann-Whitney-U-test	$U = 55.5$	.003	r = -0.51
Mid B	14.41	[5.19, 23.63]	41.35	17				
Post A	35.36	[23.84, 46.88]	19.95	14	unpaired t-test	$t(30) = 2.17$	.039	$d = 0.94$
Post B	22.5	[15.72, 29.28]	13.64	18				

Table 4.14: Statistical results of procedural knowledge evaluating questions in mid- and post-test for group A and B students

Procedural Knowledge - Create

Only one question dealt with students abilities to create procedural knowledge. However, ten points could be achieved for this question asking students to write a report. Again, different directions of development could be observed as students in group A deteriorated from mid to post test in contrast to group B participants, whose skills improved (cf. table 4.15). After the first treatment, the explorers of VisiStat scored about 15%, surpassing the lecture participants, who did not gain 1% on average. In the post-test group A learners achieved 7.5% on average, showing worse results than group B students with an average score of about 10%. However, the group B was not able to surpass the results group A obtained in the mid-test. Although no significant differences could be identified, medium effect sizes could be measured for both results, $r_{mid} = -0.34$, $r_{post} = 0.31$.

	M	95% CI	SD	n	Statistical Test	Test result	p	Effect size
Mid A	14.69	[3.32, 26.05]	21.33	16	Mann-Whitney-U-test	$U = 95$	.146	r = -0.34
Mid B	0.88	[-0.48, 2.24]	2.64	17				
Post A	7.5	[-2.8, 17.8]	17.84	14	Mann-Whitney-U-test	$U = 168$	.116	$r = 0.31$
Post B	10.83	[4.97, 16.7]	11.79	18				

Table 4.15: Statistical results of procedural knowledge creating questions in mid- and post-test for group A and B students

The results presented in this section are summed up in table 4.16. After the test results have been represented in this section, the following sections deal with the evaluation of the other methods. The next part demonstrates the results of the observations, which are followed by the quantitative analysis of students' feedback in the questionnaire. Eventually, students' opinion is evaluated qualitatively in the last part, which presents the results of the interviews.

	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual Knowledge	Marginally PFL	Lecture*				
Conceptual Knowledge		PFL		VisiStat, traditional*		
Procedural Knowledge		PFL	*		VisiStat, PFL	*
Metacognitive Knowledge						

Table 4.16: Evidence suggests which approach was crucial for students' improvements (VisiStat vs. lecture, PFL vs. traditional tell-and-practice) for each learning dimension

4.2.2 Observation

In a first step, three teams' explorations with VisiStat were recorded regarding the categorization scheme in Section 4.1.9. The results are shown in figure 4.14. It can be noticed that students form predictions about VisiStat's behavior but are not able to interpret all of them. To confirm or reject their predictions, students used either the help text, the result section, or the reporting text. Furthermore, they established interpretations based on graphs. However, it has to be stressed that these results cannot be generalized but a full analysis of the video recordings is necessary.

Furthermore, students' use of the help function in VisiStat was analyzed, revealing that the teams in group A used the help function more often in 15.98% on average of their exploration time (95%CI $[5.84, 26.12]$, $SD = 12.13$, $n = 8$). Group B teams made use of the help function in 10.63% of their exploration time on average (95%CI $[3.43, 17.83]$, $SD = 9.37$, $n = 9$). However, these differences were not significant ($t(15) = 1.025$, $p = .321$) but constituted a medium effect size ($d = 0.57$). The results are summed in figure 4.15.

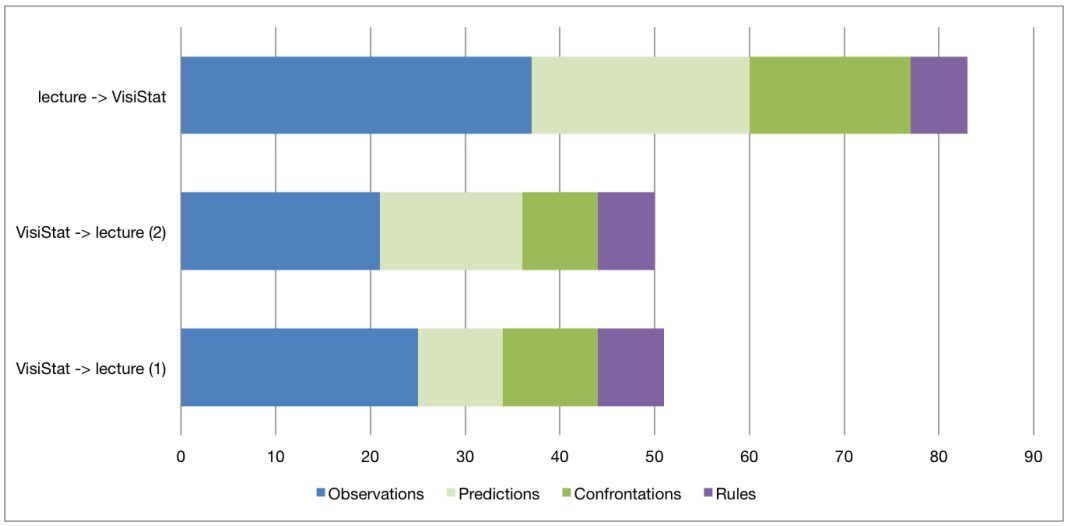


Figure 4.14: Preliminary coding of three observations. Participants' utterances were analyzed and categorized into observations, predictions, confrontations, and rules.

4.2.3 Feedback Questionnaire

Before analyzing the results of the individual dimensions, a brief outline of the results is given of the results, which are illustrated in figure 4.16. Except for the evaluation of focused immersion, the results are generally around a value of five on the seven point likert scale. Focused immersion is the only dimension as well, which is assessed better by group A than by group B. For all other categories, group B students rated better than group A. However, most of these differences were only small so that no significant differences could be reported between the two groups. In the course of this section, each dimension and their results are presented in the order of the questionnaire, beginning with perceived usefulness.

Perceived usefulness

The two core components of TAM, perceived usefulness and perceived ease of use, are evaluated around a mean of five by the participants. Students in group A rated the usefulness of VisiStat with 5.04 (95% CI $[4.49, 5.58]$, $SD = 1.03$, $n = 16$) on average. In group B, it was graded slightly better with a mean of 5.33 ((95% CI $[4.87, 5.79]$, $SD = 0.93$, $n = 18$). Regarding the perceived ease of use, a difference of

Perceived ease of use

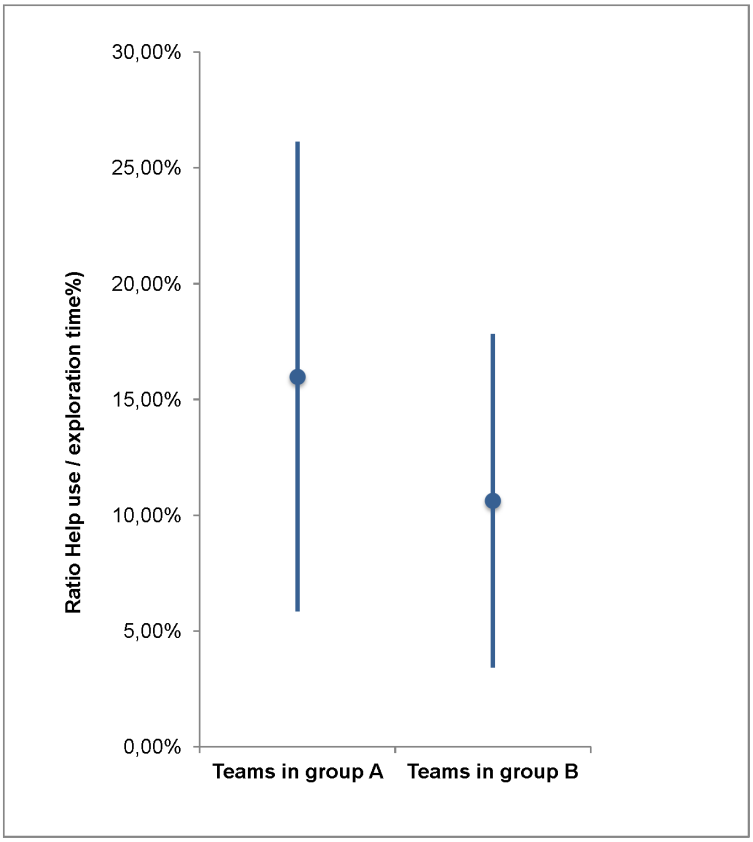


Figure 4.15: Students teams' part of help function usage in overall exploration time of VisiStat

a half interval could be observed between the two groups, as group B participants perceived the ease of use about an average value of 5.18 (95% CI $[4.8, 5.56]$, $SD = 0.76$, $n = 18$) in contrast to group A, who agreed only with 4.58 (95% CI $[3.98, 5.17]$, $SD = 1.11$, $n = 16$) to a satisfactory usability. Although no significant differences could be detected ($t(32) = -1.86$, $p = .072$), the differences constituted a medium to large effect size, $d = 0.79$.

Enjoyment

The traditional learners enjoyed the use of VisiStat with a mean of 5.21 (95% CI $[4.92, 5.5]$, $SD = 0.58$, $n = 18$), the students treated with the PFL approach described the experience's fun with 4.88 (95% CI $[4.62, 5.13]$, $SD = 0.47$, $n = 16$). Participants in group A stated with an average of 5 (95% CI $[4.27, 5.73]$, $SD = 1.38$, $n = 16$) that the flew when

Temporal dissociation

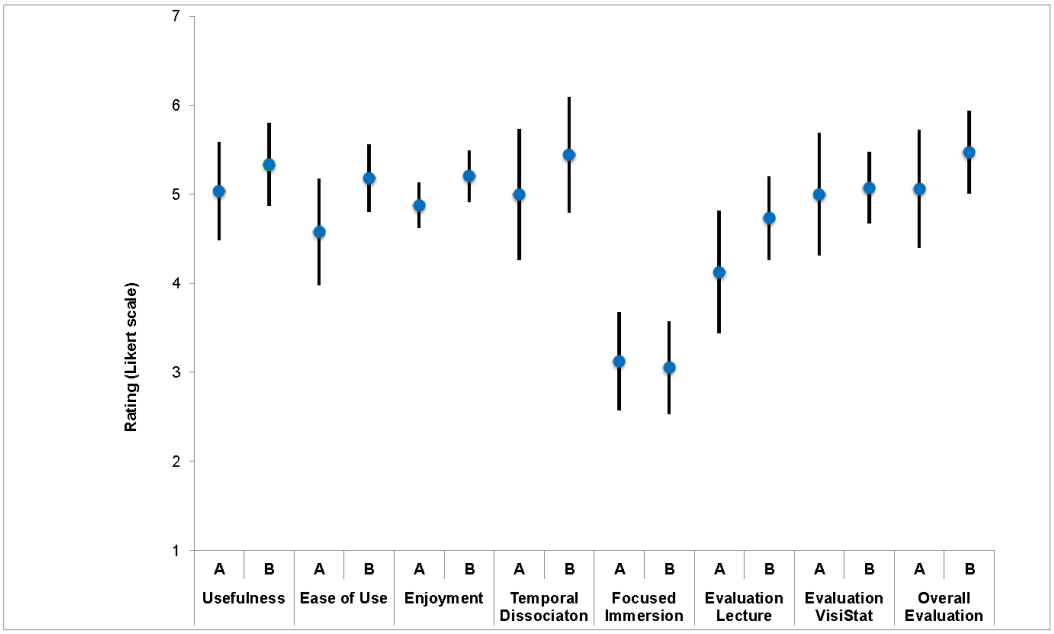


Figure 4.16: Students' evaluation of VisiStat and lecture in the feedback questionnaire

they used VisiStat, evaluating slightly worse than learners in group B ($M = 5.44$, 95% CI $[4.8, 6.09]$, $SD = 1.29$, $n = 18$). On average, the users did not totally absorb while exploring VisiStat with nearly no difference between group A ($M = 3.13$, 95% CI $[2.58, 3.67]$, $SD = 1.02$, $n = 16$) and B (95% CI $[2.54, 3.57]$, $SD = 1.04$, $n = 18$).

Focused immersion

The lecture was evaluated with a 4.13 out of 7 on average (95% CI $[3.44, 4.81]$, $SD = 1.29$, $n = 16$) by students of group A, whereas participants in group B appreciated the lecture more, assessing it with a mean of 4.74 (95% CI $[4.29, 5.2]$, $SD = 0.94$, $n = 18$). Even though no significant differences could be reported, the calculation of Cohen's d revealed a medium effect size, $d = 0.65$. However, when considering Kolmogorov's test, the data for group A was not normally distributed so that Pearson's r was considered as well, which only showed a small effect, $r = 0.22$. Examining the overall evaluation of VisiStat, a better assessment could be identified with only slight differences between the group A ($M = 5$, 95% CI $[4.32, 5.68]$, $SD = 1.29$, $n = 16$) and group B ($M = 5.07$, 95% CI

Evaluation of Lecture

Evaluation of VisiStat

Overall evaluation

$[4.68, 5.47]$, $SD = 0.8$, $n = 18$). The overall learning experience is evaluated with a mean of 5.06 (95% CI $[4.4, 5.72]$, $SD = 1.24$, $n = 16$) by group A and a 5.47 (95% CI $[5.01, 5.94]$, $SD = 0.93$, $n = 18$) by group B. After the quantitative data has been analyzed and presented, the next section deals with the qualitative results the interviews produced.

4.2.4 Interview

The interviews with students were conducted to gain full insight into students' opinions and behavior. The following tables show the results of students' responses in form of categories, developed with the *Grounded Theory Approach* [Glaser and Strauss, 2009]. Apart from the category's name, an explanatory description of each category is given, being supported by a prototypical direct quotation from the participants. Eventually, the last two columns outline how many students of the 16 members of the PFL (A) and the 18 members of the traditional learning (B) group agree with this statement. The categories are arranged by their occurrence in VisiStat and lecture. In both dimension, it is differed between strengths and weaknesses. Furthermore, the fifth table illustrates students' comments regarding how lecture and system complement each other. Students' attitude towards learning statistics is summed up in the last table.

Category name	Explanation	Quotation	A	B
In VisiStat, learnt statistical concepts can be applied and practiced	VisiStat can be used to practice statistical concepts after they have been leant before (e.g. in the lecture) and show how concepts are visualized and how to apply them.	"The lecture explained ehm already about this ANOVA, the paired test, unpaired test, we got introduced first before we did this [VisiStat] [...] yeah these ones like he said (?) independent variables and then dependent variables and then you can practice some things this is the good thing"	8	18
VisiStat provides a first overview of statistical concepts	VisiStat can be a starting point to explore statistical concepts and provides a first overview.	"So it's good to see that [VisiStat] as a starting point but to really understand it and to see more examples and have all the background information the lecture I think is therefore advanced people a bit better to just learn the stuff but as a starting point I'd recommend the system just to play around a bit and see what you can do with statistics"	12	12
VisiStat supports to build knowledge of reporting	Knowledge in form of important values (incomplete) of how to report results appropriately can be developed in VisiStat.	"Uhm: I I actually did get the standard of reporting from the system and they were/ whenever we generated the report and then we got like significant or unsignificant."	11	13
VisiStat supports to develop hypotheses about statistical concepts	VisiStat encourages the user to develop hypothesis about statistical concepts, like the characteristics of the data determining VisiStat's choice of a statistical test.	"VisiStat allows a bit for experimenting and yeah going into depth and thinking about why a specific test is chosen at a time."	10	13
Two Assumptions are easily remembered from VisiStat	The visible checking of the two assumptions (normally distributed and homogenous variances) allows to easily remember and learn these assumptions in VisiStat.	"I liked ehhm=that one could see the assumptions for the statistical test ehm what is shown here with data normally distributed (-) ehm its kind of a check list (-)"	13	9
VisiStat visualizes statistical concepts	Visualizations in VisiStat support to understand statistical concepts and visualize the theory, helping to remember them.	"I would think so for example the different types of diagrams box plot or whatever ehm you have seen it before and you will have a basic understanding of what it tells you where the median	11	11

		is or.”		
VisiStat allows high data visualization interactivity	VisiStat allows high interactivity by enabling to observe the results (e.g. a graph) after changing the input.	“Maybe ehm that you can explore on your own what changes when you connect something or when you do not connect […] and see the result”	10	11
VisiStat provides knowledge to differ between tests	VisiStat provides a basis from which to learn about different tests and know some characteristics how to contrast between them.	“Yeah I think we had two tasks, one was one way ANOVA and one was two way ANOVA and eh there I got introduced to these two tests [in VisiStat]”	8	12
VisiStat provides a hands-on experiment	In VisiStat users can do something on their own, experiment, and go into depth.	“The advantage is that you can try a bit on your own and experiment with the data set which obviously you can't do in the lecture because eh=ja [the lecturer] is doing everything and presenting a lot of stuff that eh (--) comes with a=a lot of speed and it can't remember everything and VisiStat allows a bit for experimenting”	5	14
Reporting text helps to understand results	The automatically generated reporting text in VisiStat helps to understand the results.	“When we generated the report I think its I don't know if it's correct but in my opinion it's standard for for the report and it helped us to understand the results yeah”	7	11
VisiStat prepares for the lecture	VisiStat prepares for the lecture so that users have already familiarized with concepts, developed a basic understanding or heard terms so the lecture can be followed and understood better. Furthermore, they focus on problems in the lecture they discovered when using VisiStat.	“I think it's quite suitable if you would do the study first and then the lecture because eh yeah you get so much information about statistical test and I think it's better to ehm do the study because ehm you have better visualization and (-) you have a little bit more knowledge if you go into the lecture so that you don't sit there and think oh god I don't know anything but yeah I think yeah it's quite hard to follow [the lecturer] the whole time in the lecture and if you have this knowledge then maybe you can like fell asleep for a few minutes and then get	14	2

		back and eh ehm ehm and have all the ehm known yeah.”		
Use of VisiStat is fun	The use of VisiStat is a satisfying experience and makes fun.	“I think using this [VisiStat] I think it was much more fun than the lecture and you do things, click things and yeah I like I personally I like to learn by the hands-on-experience”	5	8
VisiStat is easy to use	VisiStat is easy and intuitive to use.	“It [VisiStat] also has the advantage that (-) it's easier to use than other tools. If I like if I just want to sit down and have a data set and I want to try statistics on it using excel. That’s not fun because I not only need to learn statistics I also need to learn excel which is almost as hard as learning statistics on itself. so it's an easy entry point for playing around with data and seeing different outcomes (-)”	5	6
VisiStat’s help function offers explanations	The help function in VisiStat offers explanations of statistical terms.	“I think it's suitable for learning because we have always this help category where you can hover over the parts you don't know. I think it's very good explained so that you understand the unknown parts of kind of some test that you don't know before.”	5	6
Report function in VisiStat is useful	The possibility of generating a report in VisiStat is considered as useful, e.g. for writing a paper.	“The report function was nice for if I wanted to write a paper really because everything was in there and was really easy”	4	7
VisiStat provides practical knowledge to apply reporting	VisiStat teaches how to actually write an entire report for a paper.	“There are few things I'd say I learnt from the system especially when it comes down to how I'd actually write a report and the results down and which parts are important for a result”	4	6
VisiStat can be explored at individual speed	VisiStat can be explored at the users' own speed so that not understood concepts can be revisited and addressed longer whereas already	“In the lecture sometimes what happens is you won't be able to concentrate the kind of things he shows in class. Here [in the system] nothing like that	6	3

	understood concepts do not have to be paid further attention to.	can happen, there is always an option of going back and seeing what it is and then coming back what you're doing"		
Low background knowledge is required to use VisiStat	To use VisiStat, low or even no background knowledge is required because it automatically applies the appropriate statistical concept.	"We didn't know at least I didn't know more stuff the (--) tests and what=what they actually mean so we when it just selects (base) done how the/ of course after the lecture we found out a lot of stuff […] but there when VisiStat was doing it for us it was quite easy I mean we didn't have to know anything"	7	1
VisiStat is useful for research	VisiStat is useful for applying statistics and writing the results in a thesis or paper.	"I really liked it for if you really have to do a study and you need to have like something to put in your paper"	5	3
VisiStat provides information to learn	VisiStat provides information about statistical concepts, which can be learnt during exploration.	"It was very user-friendly very interesting also the outcome that it is already written all the informations important ehm (-) that was so you know this test normally you need these values this"	6	1
VisiStat has a user-friendly interface	VisiStat provides a well designed interface which provides a good interaction with the use.	"In general it was I think very good the interface […] it was very user-friendly"	1	6
VisiStat provides to learn new concepts	New statistical concepts can be revealed by the user, like the correct choice of a statistical test or the different amounts of the effect size.	"When you don't know anything about statistics and you first time using it, it is really to understand the concepts the new concepts in a way you wouldn't get to know the new concepts"	4	0
VisiStat compensates for lack of statistical knowledge	Due to the automatic correct application of statistical concepts, users do not have to worry about making mistakes and are prevented from making mistakes because their lack of statistical knowledge is compensated for.	"So sometimes is it that you know even you might have lost track of the lecture that you know what is what was what is a significant (?) like I mean like for a moment I just forgot that there is something called a post-hoc test when it came and showed now you need to do a post-hoc test if that if it required then it got something back"	2	2
VisiStat supports to become aware of over-testing	Users become aware of the risk of over-testing data, especially as the	"Yeah after using the system I tend more to eh use post hoc tests because	2	2

	system recommends to use post-hoc tests.	it was in the system but ehm why it is I do not know it was a similar situation question from there and the system recommended a post test and then ok similar situation and I thought why it is because of over-testing I did not know it.”		
VisiStat is useful for exam preparation	VisiStat would be used again by students to learn statistics in order to prepare for the final exam.	“I would use it [the system]" " in what context?" "also for the exam preparation" "so you like for practicing" "mhm" "or for learning the theory" "for both ehm also because yeah you have so many help other options and can look up”	2	2
VisiStat can be used for approval if concept was understood correctly	To check if a statistical concept was understood correctly, the user can test his/her understanding with VisiStat.	“Or I will use it but for preparation for the exam I think I will use it really shortly or not because in an exam it normally more about the theory and you do not can understand the theory using the/ it's only for approval if you have understand it or not if you have understand the the concept or if you do not have understand the concept”	0	4
Three assumptions can be learned from VisiStat	VisiStat supports to learn all three assumptions for the use of parametric significance tests.	“There are more than two conditions we thought one-way ANOVA was conducted because there was difference in the tests, all of the questions were the same the tests conducted were different for different kinds of datasets because of the conditions”	0	2

Table 4.17: Categories describing strengths of VisiStat

Category name	Explanation	Quotation	A	B
VisiStat does not provide sufficient knowledge of appropriate testing	Students cannot understand why VisiStat chooses a particular test for the situation and have difficulties to apply the tests themselves.	"But I had I found it a bit problem with the system. I couldn't understand I mean by using the system I couldn't understand which test should be used at which point of time so that was not that visible"	11	11
Help description is not detailed enough	On the one hand, the help description is not detailed enough to understand the concepts and should be accompanied by an example. On the other hand, it does not offer help descriptions for the graphs and functions of VisiStat.	"We once looked up one thing in help and it was described something like I do not understood it. I do not get this the concepts before a lectures and also not in last lecture here and eh I do not also understood this information it was only one sentence I do not can make further conclusions because I do not understood this basic words or."	10	9
Low interactivity	Users do not have the possibility to interactively change the input to be able to observe the corresponding changes. Furthermore, they want to be tested by the system and get feedback about their performance.	"Ehm in my opinion I think I would prefer a direct feedback because ehm (-) like now we had no feedback at all so we can validate what we did by seeing the lecture and understanding what we did but actually we had no feedback at all and if I for example have a question for understanding and I'm answering it maybe really in the software then the software will respond and say oh no you misunderstood something look again there"	10	7
Hypotheses cannot be confirmed in VisiStat	Hypotheses about statistical concepts, which have been developed when using VisiStat, cannot be confirmed with the help of VisiStat, because it does not provide an explanation.	"So you saw that there are different test and you kind of might form some assumptions why this test is used and the other one but you weren't sure about this"" "yes"	8	6
Students do not develop knowledge about over-testing	Students cannot develop any knowledge of over-testing when using VisiStat.	"No I think I didn't know it [over-testing] from the lecture and I didn't really get the the feeling of over-testing from the program"	5	9

Users are not encouraged to think about solutions	The automatic calculation of results in VisiStat does not encourage users to think about solutions themselves but only serves to answer the given questions.	"The problem when we solved it before was as you said there were no questions we had to VisiStat (-). We had not the intent to learn something but to solve this task so we just (-) trial and error a little bit until we get the right solution and the right results and we had no intention to learn what is meant there because there were several tools like this help (-) which would have explained us several things ehm but we just didn't use them because we wanted to conclude the task and not to learn something about statistics so having it in after the lecture (-)"	10	3
Help description is too difficult	The help text uses difficult language which cannot be understood by users and should make use of an example, which explains the concept.	"Another thing could be that when you hover over around the things that I said maybe if ehm the definitions were more in (plain) terms maybe it would help understand better because if it's too much in technical terms the definitions were too difficult to remember what it was"	5	8
Statistical concepts cannot be applied	After using VisiStat, the statistical concepts cannot be applied by users themselves.	"From just I ehm reading it [reporting text in VisiStat] a few times I could not use this/ like I could not produce this text on my own."	8	4
Previous knowledge is required to use VisiStat	Basic understanding of statistical concepts is necessary to use VisiStat so that it cannot be explored by novices.	"The system is not for a novice so for example I don't know anything about statistics I won't be able to to use it at all"	3	9
Users cannot try different statistical methods and error	As VisiStat automatically applies statistics on data, the user does not have the possibility to try a statistical concept him-/herself (like the choice of a statistical test) and recognize if this choice was correct and why.	"Ehm I think the automation in this case ehm is a little bit ehm constraint these because ehm then I can/ if I have a data set and it will already find ehm the statistical methods by its own (-) I can just not ehm like try and fail. I could not say like I want to run this statistical test on it and then it would tell me no you can't because of and then I would remember ok I can't do this	8	3

		because of that (-) but instead I just say here's my data here's what I want to test and then it just says here is your ehm (--) your method and you don't know why but it's correct (-) so ehm (-)"		
Reporting text does not enable to understand or apply reporting standards	Reading the reporting text in VisiStat does not enable users to fully understand how a report should be written and especially not, how to create a reporting text on their own.	"I've never seen anything like this is the standard template and I don't know which of these information (-) is basically important for the standard where what I could vary, if I can vary anything, and how this relates to each test was important for each of the test."	7	2
Help function is time consuming	The use of the help function is time consuming as it has to be clicked in the upper right corner, then hover over elements and eventually, deactivate the help again.	"You need to click on help and hover over that which would be lot time consuming if I was starting something or learning something. If I wanna learn something it would be in a small span of time because you had a limited amount of time for attention right"	3	6
VisiStat does not cover all statistical concepts	Not all statistical concepts can be explored and found by students in VisiStat, like the meaning of confidence intervals or Cohen's d.	"This this ehm I'm not sure if it was in the system but this graph showing the ehm confidence intervals what it means if we have a zero ehm if we have a 95% confidence interval and how this this ehm diagram was changing when he (-) ran through the statistical tests."	7	1
VisiStat does not support to gain satisfactory statistical knowledge	Users are not satisfied with the amount of statistical knowledge they could gain in the system and criticize the structure for learning.	"So when after using VisiStat ehm I did not really perform well on the ehm test as I did after the lecture. So I think using the tool before seeing the lecture is eh the wrong way around. That was my impression. Because on the tool okay you could see everything but you don't really/there's not/ you don't have the structure you just have like a I can see this and this and this and I click here and funny things happen then. At the moment it/when you use it it makes sense so you think yeah like of course it's like this and this but if	7	1

		you have to like (-) regain the knowledge (-) then (-) it's gone because it's basically all still in the tool. And it did not really transfer to that."		
Students do not want to learn with VisiStat but only solve tasks	Students can only learn with VisiStat when they are motivated to learn with it. Without further motivation, they do not intend to learn with it but only try to solve the tasks.	"The problem when we solved it before was as you said there were no questions we had to VisiStat (-). We had not the intent to learn something but to solve this task so we just (-) trial and error a little bit until we get the right solution and the right results and we had no intention to learn what is meant there because there were several tools like this help (-) which would have explained us several things ehm but we just didn't use them because we wanted to conclude the task and not to learn something about statistics so having it in after the lecture (-)"	7	1
Assumptions are not understood in VisiStat	Although users report that the assumptions can be easily remembered from VisiStat, VisiStat does not provide to understand the meaning of these assumptions.	"I liked ehhm=that one could see the assumptions for the statistical test ehm what is shown here with data normally distributed (-) ehm its kind of a check list (-) but (--) is it/ it doesn't help me (--) because I have not (-) ehm the knowledge (--) so if one should/ would only show me that I didn't know how=which test to chose"	6	3
Over-testing is observed in VisiStat but not understood	When observing over-testing in VisiStat, the user cannot make sense of its warning and does not entirely understand it.	"I was not sure that we can over-test/ that the over-testing depends on the test I think or I thought that it would be possible to test with every test so I was a little bit surprised that there was this sentence that we do not use a pairwise t-test because we do not want to risk over-testing."	5	0

Learning aim is not clear	The learning aim of exploring VisiStat is not clear for users so that they do not know what they want to learn from the system.	"I think my point is that I didn't know what do I wanna learn from this unless I did the test. And the reason I learned some stuff from this because I did the before test because I kinda knew what should I be learning"	4	0
Reporting text is difficult	The reporting text is difficult to read and the values which have to be reported are not marked as such.	"So you're saying that you liked that the description text [reporting text] you liked that? Because I considered that it was like like floating text and with numbers in it and it was hard to read for me […] yeah I think that text is unreadable"	3	0
No exploration but only telling	As users cannot decide how VisiStat calculates the results, it only tells solutions and does not support own exploration.	"Yeah I mean they touched a few different things I think so it complemented itself but in the same way just like giving me information and remember it so it was not that interactive"	2	0
Help description is not visual enough	The help text is too text-based and therefore, difficult to understand. The help function should integrate visualizations instead.	"So ehm there would be an easier way to navigate through different contents, probably say like an explorer or and those explorers could visualize for instance you have pair tests or something like that and can visualize what pair test is or can explain it graphically or animate it so you would rather understand it better when you see it visually and then proceed the text rather than just going through the text and then seeing ok this does this but does not give you a complete theoretical foundation"	0	2

Table 4.18: Categories describing weaknesses of VisiStat

Category name	Explanation	Quotation	A	B
Lecture provides knowledge to use appropriate tests	The lecture explains students how to use appropriate tests and their individual strengths.	"I think I know this [knowledge about appropriate testing] from the lecture"	12	14
Lecture provides deeper understanding	The lecture provides deeper knowledge of theoretical statistical concepts and helps to understand complex topics.	"I think that's the main part in the lecture there you get deeper knowledge"	11	12
Lecture explains assumptions	The lecture explains assumptions for parametric statistical tests and shows how to check them.	"From the lecture I'd say that the system we did after it and I already knew it from the lecture"	9	6
Lecture provides basic understanding	The lecture provides an overview of statistical concepts and raises students' understanding.	"So for me it helped me ehm (---) like getting this birds eye view on all the statistic methods and when to apply what because ehm there usually lots of different statistic ehm (--) yeah eh procedures techniques that you could use but it's not as obvious to know when to use what"	8	6
Content is explained well	Content is explained clearly and well in the lecture so that it is easy to understand.	"The main advantage of the lecture is that you have the conversation discussion with [the lecturer] and that he can better explain this stuff"	7	7
Lecture prepares for using VisiStat	Knowledge and basic understanding gained in the lecture prepares for using VisiStat.	"Yeah all this you have already an idea what is ANOVA, t-test because ehm (-) ok ANOVA I know from another term because I had this statistic lecture in German but t-test I do never know before and then you know oh ok it's something like this something you have already an idea because otherwise I think you do not can get information out of the the system because you miss too much information and I think it was very important to have this basic understanding."	0	14

Use of visualization demo	Use of demo of basic statistical concepts visualizes the lecturer's explanations.	"The demo I'm not sure what program was that but it basically had to complement whatever he was trying to tell us and sure it was well time what are the changes taken (--) that was the good part of it I guess"	7	4
Lecture explains concepts observed in VisiStat	Students observed terms or behavior in VisiStat, they did not understand. The lecture then explains and answers these open questions, confirming or rejecting their hypotheses.	"We didn't know at least I didn't know more stuff the (--) tests and what=what they actually mean so we when it just selects (base) done how the/ of course after the lecture we found out a lot of stuff like why normal distribution it needs to be tested in front particular way"	8	2
Flexibly addresses students' needs	The lecture flexibly addresses students' needs because students can ask as long as they understand and the lecturer explains until they understand.	"If you don't understand something then you can ask again and if you (-) then already didn't understand something then you ask again and here you have just this text and if the text doesn't help you then yeah, you need the internet, but yeah if you think it's more comfortable to discuss with people then reading text yeaaah the main advantage of the lecture is that you have the conversation discussion with [the lecturer]. and that he can better explain this stuff"	6	4
Lecture is presented well	The lecture and its contents are presented well by the instructor, especially the structure, the visualization of the decision tree and the combination of visual and auditory explanations.	"Yes because there was a running commentary when [the lecturer] was explaining certain/ he ran a simulation of means and variances and he was explaining it to us so it was more of a (-) visual and auditory combination that helped us understand in a better way and it gave us a different perspective as to ok this is what means."	5	3
Lecture provides knowledge how to report results	Lecture provides knowledge how to report results and which values have to be reported.	"Do you know standards of reporting your results?" "from the lecture"	4	4

Good first part of lecture	The beginning of the lecture and the introduction of basic statistical concepts is good.	"And ehm for the beginning the lecture was very good"	3	4
Lecture offers explanation for statistics papers	Statistics papers can be understood after the lecture as the terms are explained.	"I finally figured out what a lot of those terms meant […] and also when we go through the statistics papers"	3	2
Lecture explains the risk of over-testing	Students gain knowledge of the risk of over-testing from the lecture.	"Ehm where do you know this [over-testing] from, from the lecture or from the system" "from the lecture"	3	2
Clear consecutive structure	The lecture follows a clear and consecutive structure as basic concepts are introduced first and based on this knowledge, the complex parts are presented so that they can be followed well.	"Think the main advantage of the lecture is its (dadules?) so in the lecture you begin from very low level and then you build on it but eh in the system you just/ your thrown at the last level so yeah the lecture is (-) is has an advantage in this area"	3	1
Interesting content	The content in the lecture is interesting for students.	"I think the stuff that [the lecturer] presented was quite interesting"	0	4
Lecture is appreciated	The lecture is appreciated and evaluated as good.	"I I have to say I liked the lecture"	2	1
Lecture explains why values have to be reported	Lecture explains what the values which have to reported mean and why they have to be reported.	"Both I mean from the VisiStat I knew that they were in the report because it was there […] but here for example the CI I didn't knew it was confidence interval or what so I knew that those things must be there but for the understanding why there are important and what they really mean the lecture was essential"	2	0

Table 4.19: Categories describing strengths of the lecture

Category name	Explanation	Quotation	A	B
One Lecture is not enough for content	The lecture dealt with very much content which cannot be taught in only one lecture but should be distributed over at least two lectures.	"And I would suggest that the first lecture could just introduce concepts and the different methods and tests that are around both. and the second lecture could deepen your knowledge as okay you do this particular thing as so and so and because so and so."	13	17
No interactive practice part	Lecture is not interactive as it does not offer a practice part, which would have helped to deepen the knowledge.	"I think it would be nice if you had to do ehm (--) ehm some stuff ehm by yourself because [the lecturer] was showing a lot of things and it=it was a lot of stuff but eh difficult to remember (---) and since (-) ehm we had no time practice anything of this there was just eh (-) yeah a lot of stuff and one after the other and (-) yeah very difficult to remember everything (---) ehm in the end and because we hadn't done anything by this ehm ourselves yeah it's (---)"	9	10
Not spending enough time on tests and their characteristics	The second part of the lecture, which deals with the decision tree of statistical tests, was presented too fast. Instead, it should be spent more time on the choice of test and the tests' characteristics.	"That is once the when you use which test it was a graphics at the end of the lecture but only mentioned in one minute fast talk and it was over and it was too fast to see the whole tree with all the information and that was and the questions were a lot focused on this tree and the when you use which test and it wasn't uploaded and you cannot remember in one minute the whole tree [...] and to explain it why it is and eh yeah little bit more structured how the tree is"	8	7
Students do not develop knowledge about over-testing	Students do not develop any knowledge about over-testing in the lecture.	"No I think I didn't know it [over-testing] from the lecture and I didn't really get the the feeling of over-testing from the program"	5	9

Structure could be improved	The lecture's structure could be improved regarding the time for different topics and an easier beginning.	"I would say showing this decision tree before not after before everything just to know how mentally to think over the whole thing"	8	4
Knowledge about over-testing cannot be entirely understood in lecture	Students remember that the lecture dealt with over-testing but cannot entirely understand the meaning of this risk.	"In the lecture it was the example of the t-test were applying multiple times the t-test and raised the false positive probability but for me it's not really clear how applying a test several times as it does not change the eh (-) data set as when I remember this right effects my results. so this is what's not (---)"	6	5
Lecture does not provide sufficient knowledge about appropriate testing	The knowledge of how to choose a specific test cannot be fully developed by students in the lecture.	"I'd know it [appropriate test choice] was in the slides but I couldn't really say much about it"	6	5
Difficult to follow lecture	Due to many presented concepts, it is difficult to follow the lecture the whole time. Therefore, important parts should be announced by the lecturer so that it can be focused on these parts.	"Just split it into two parts because it was (-) really hard to follow all the time (--)"	6	5
Meaning of assumptions cannot be understood	The existence of assumptions for statistical tests can be remembered from the lecture but not the exact meaning of these assumptions and how they are connected to a specific test.	"I could remember that there were always those two conditions" "but you don't really remember how they are named or what they mean" "yeah how there were connected to the specific test"	4	7
Lecture does not provide sufficient statistical knowledge	The lecture's content is not detailed or explained enough so that students do not benefit from it sufficiently and are not satisfied with the gained statistical knowledge.	"There was also one question about Cohen's d in the test if ehhh they probably did/ what is the effect size if small or big. This was explained in the lecture ehm but (-) there was for small and big effect were both types with too much users or too less (-) ehm and this was in my opinion totally not explained in the lecture. So I could have answered ok small effect but it would have been pure guess if too much or too low number of participants. so there was several points	6	3

		were lecture was not enough for the post-hoc test.”		
Going through lecture is necessary for understanding	For really understanding the lecture’s content, follow up reading of slides or rewatching of the lecture video is necessary.	“Also it's a bit= I mean because you didn=i didn't go through the slides again it's hard to relate some of the stuff back to (-) you're not really sure I mean I have to read up to understand exactly where everything is”	3	6
No time to go through first part again for understanding of second part	After the first part, a break would be needed to go through this first apart again and deepen the understanding and then learn the second part.	“I think that [splitting the lecture into two] would be a great idea cause now we have some time to revue the first part so the basics would be very (?) when I I listen to the second part it would be easier”	3	5
Presentation of content is too difficult	The content in the lecture is not presented in an easy way and too difficult to understand, especially more concrete and easier examples should be used.	“What I would really like to have is more concrete examples (-) ehm in the lecture because for my learning it also helps if I can (--) like relate like construct a situation that is related to the question and then I can then if someone throws me a new situation I can think ok this is somewhat like the situation I already had and in this case I used that method and then I can make conclusions and relations between this and this would help me more than just to basically hear like (-) in this abstract situation you use this and yeah.”	4	1
Concept of non-parametrical tests cannot be understood	Students do not become aware of the characteristics of non-parametric tests and are not able to attribute specific tests to this category after the lecture.	“Ehm there [concepts not seen in lecture] were these two tests ehm (-) I already forgot the names." "Mann-Whithey-U-test and Kruskal-Wallis?" "yes exactly I don't know if they were mentioned in the lecture but to me they were completely new.”	2	3
Lecture does not explain reporting detailed enough	Although reporting results can be remembered from the lecture, they are not explained detailed enough as students do not understand how this	“I've never seen anything like this is the standard template and I don't know which of these information (-) is basically important for the standard where what I could vary, if I	2	2

	text varies for different tests.	can vary anything, and how this relates to each test was important for each of the tests.”		
Lecture does not prepare sufficiently for using VisiStat	One lecture does not prepare sufficiently for using VisiStat, but more training is necessary.	“I think I have to be trained before using the system so even after after attending the lecture once I didn't get all the (-) things.”	0	4
Time of lecture in semester is not ideal	The statistics lecture is not held at an ideal time. Students either claim it should have taken place earlier to be able to understand the papers or later after the mid-term as this distracted students.	“Ehm (--) I think we were a bit distracted with the mid-term so we didn't really took all that much from the lecture. If it were after the mid-term probably we would have paid more attention”	0	4
Not enough previous knowledge to follow lecture	The lecture does not begin on students’ level of knowledge. Students claim not to have the necessary previous knowledge to follow the lecture.	“The lecture, I found it to have ehm well if we're doing to really understand a lot of statistics I found it to start even from a little but not explaining well it does not take us our level”	1	2
Lecture does not enable to apply concepts	Lecture provides understanding but it is still difficult to actually apply these learnt concepts.	“For me it was also an overview and how one can use all those things and=ehm (-) interpret it but not (-) when I have to use something and: I don't think that I have learned so much ehm how I can use something or when I have to use something”	2	0
The consequences of violated assumptions are not discussed	Although the lecture deals with assumptions, the consequences of a violated assumption and how to deal with it are not discussed in class.	“Yeah the parts which said that you know the test conducted compensates for this compensates for the normal distribution so that I think that wasn't discussed in class what would happen if an assumption is violated”	0	2
Language problems	Due to different mother tongues and accents, students have difficulties to follow lecture.	“I can't I can't understand all (-) (?) that [the lecturer] eh (-) said maybe because my eh (---) my (?) […] my (--) eh: my listening skills is=are limited […] in English”	1	1

Table 4.20: Categories describing weaknesses of the lecture

Category name	Explanation	Quotation	A	B
Statistical knowledge definitely improved	After attending lecture and exploring VisiStat, students estimate that their statistical knowledge definitely improved.	"Definitely because when I eh first filled out the eh pre test I thought yeah ehm I have this knowledge from my bachelor thesis so it's okay but ehm I filled out the test and thought oh don't know anything really (-) and now there were some questions that I didn't know the answer to but ehm lots of questions I think I could answer so yeah I can feel the progress that I get to know much more knowledge about statistics."	10	12
Learning improved but room for improvement	The statistical knowledge improved but there is still room for improvements and students are not sure about every concept.	"It's eh only to get a main idea about I do not can answer complete questions it all some things I'm never sure I don't know perhaps 3 questions I'm sure the rest I think ok perhaps it's 80% like this and then I don't click anymore yeah I don't know but I think this was I was not most questions I was never so sure that I think it's 100% like this or this answer. To get a feeling of it but it was too short eh to get really into this stuff because it was very lot of new stuff and ehm yeah"	11	7
Treatments provide an overview and familiarity	The two treatments provide an overview of important statistical concepts and increase the familiarity of the terms so that students are able to read statistical papers but the applying on their own is still difficult.	"Yes and maybe if I do not know how many Cohen's d or whatever test there are but I think if there are some tests I do not know I would have difficulties to read the paper but if they used an ANOVA test I would have a clue why they did it and what the results tell me"	9	4
Further learning is necessary	To really understand the statistical concepts, further learning is necessary.	"But it's [knowledge from treatments] (-) not enough for the final exam I think that I have to learn it on my own"	6	7
Sequence Lecture → VisiStat is preferred	Attending the lecture first to learn the theory about statistical concepts and then explore VisiStat for applying this knowledge is the better sequence.	"So when after using VisiStat ehm I did not really perform well on the ehm test as I did after the lecture. So I think using the tool before seeing the lecture is eh the wrong way around. That was my	6	6

		impression […] But if you had the lecture before (-) then ehm I think the VisiStat tool would be a great help to like (-) deepen the knowledge and to just keep in mind by using it instead of just hearing.”		
Sequence VisiStat → Lecture is preferred	Exploring VisiStat first to get an overview of important statistical concepts and then attending the lecture to get detailed explanation is the better sequence.	“But if you are doing it [the system] in conjunction with the lecture (-) the/ this should be introduced first rather you get an overview of what was going on and then the lecture would give you the detailed knowledge as to what these introduced thins are.”	6	4
VisiStat should be used in class	VisiStat should be used for application during the lecture so that the instructor can give feedback.	“For example like the in class activities that we do in the middle of the lecture so there is a question and then we can explore it with the system and do it with the system and then we know the right answer or how should we tackle that so that's the instruction.”	3	6
Learning experience is insufficient	Students claim that the learning experience is insufficient and they still lack knowledge of fundamental statistical concepts.	“Yeah I think the whole experience was not as good as I expected it to be finally like I didn't get now I expected that I would learn way more but now I didn't learn as much as I expected”	3	7
VisiStat should be used before and after lecture	VisiStat should be used for learning statistics before as well as after the lecture.	“For me maybe one hour before the lecture and one after the lecture would help”	4	1

Table 4.21: Categories describing overall learning experience of VisiStat and lecture

Category name	Explanation	Quotation	A	B
Pressure is needed to be encouraged to learn statistics	Students are not motivated to learn statistics unless they are encouraged by external pressure, like the passing of an exam or an exercise.	"Or one thing I would say it would help to apply a little bit more pressure on the students and at same time release some pressure from other sides because over the last three weeks my focus was really not on learning statistics and doing the studies. this was just something that was on the side and also I could do it half hearted as it did not influence my grade and there were several things in the week that affected my grade more like the exercise and the mid term and that somehow like did not encourage me to eh focus on statistics. I think if I really tried to focus on it because ehm I needed to then my learning experience would have been better. but ehm with the exercise and the midterm exam there was just (--) too much things that were distracting."	3	6
The topic of statistics is difficult	Statistics is perceived to be a difficult topic.	"I think it's quite hard to like to explain statistics in only one lecture and it's not an easy topic, you have to remember the abbreviations and so"	1	3
Over-testing contradicts intuition	The concept of over-testing contradicts the students' intuition of how they understand statistics.	"And I'm also still a little bit confused about the over-testing. It seems counterintuitive like eh what I basically say is the less I test the more confident I can be. And that just sounds weird. (_) because eh I just make one test I don't know if it's good and then I just go away and say yaaay. And the more I test the less confidence I have"	1	0
Learning statistics is not liked	Students do not like to learn statistics.	"I mean it's statistics, no one really likes to learn statistics, and if someone does, they are weird people"	1	0

Table 4.22: Categories describing attitude towards learning statistics

4.3 Discussion

The user study described in Section 4.1 was conducted to evaluate the impact of VisiStat complementing a lecture on learning statistics. After the results were presented in the previous section, the following section deals with the interpretation of these results and how they can answer the research questions, which were introduced in Chapter 1. First of all, we examine if the drawn up hypotheses can be supported by the results (Subsection 4.3.1). Therefore, the overall test results as well as students' achievements in the different learning tasks are investigated. Furthermore, it is investigated how far VisiStat and the lecture can help to prevent students from making the mistakes Cairns [2007] declared as most popular in HCI research (Subsection 4.3.2. In Chapter 3, lecture and interactive statistical analysis system were analyzed regarding Garfield and Ben-Zvi's [2007] principles for learning statistics. In Paragraph 4.3.3, this analysis is reconsidered and revised by taking students' evaluation of their learning experience with both treatments into consideration. Based on this analysis and students' qualitative feedback concerning strengths and weaknesses of both learning treatments, we describe VisiStat's role in a statistics learning experience. This part is closed with an overall evaluation of VisiStat and lecture. Eventually, students' use of the help functionality in VisiStat as well as their corresponding feedback is discussed.

4.3.1 Effect of VisiSat and PFL

The study aimed to investigate 1) if students benefit more from the interactive statistical analysis system VisiStat in order to learn statistics than from a traditional statistics lecture, and 2) whether students treated with the Preparation for Future Learning approach outperform students learning with a traditional tell-and-practice sequence. This subsection evaluates the results against the background of students' test results, compared with their qualitative feedback. At first, the overall results are examined and in a sec-

ond step, we discuss the results for the different learning tasks.

Overall results

Hypotheses can be supported

The results support our two hypotheses as students using VisiStat scored higher than those attending the lecture. If the sequence of treatments was not important for the learning success, the two treatments would simply be additive and all students would achieve the same results in the post-test. However, after the second treatment, better results can be detected in favor of group A, who explored VisiStat before and then learned in the lecture. Students in group B could not catch up after exploring VisiStat, suggesting that the sequence has an impact on students' achievements. These findings signify that the PFL approach is successful and more effective than traditional education for learning statistics. Students are prepared for the lecture due to the use of the system before and can take advantage of this previous knowledge. Regarding VisiStat, these results provide preliminary evidence for its suitability for learning statistics.

Possible external factors

However, it has to be stressed that the effects between the two groups were only small to medium so that we recommended to repeat the study with more students to validate the results. Additionally, it is possible that these differences are attributed to external factors. For instance, although VisiStat and lecture covered the same statistical topics, students might have spent a different amount of time on different problems. Whereas the lecture spent most of the first half on explaining statistical basics like confidence intervals, the tasks in VisiStat started directly with different statistical tests. Apart from the content, the lecture lasted one and a half hour in contrast to the exploration of VisiStat, which could be explored for at most 50 minutes. This inequality is the result of the different characteristics of an interactive system and a lecture but might have caused differences in the results. Nonetheless, as students using VisiStat outperformed the learners in the lecture, it can be assumed that students benefit from using an interactive system.

Mixed qualitative feedback regarding sequence of treatments

In conclusion, the test results suggest that students benefit most from the Preparation for Future Learning approach. Students' qualitative feedback concerning the sequence of treatments, however, was mixed. On the one hand, six students in both groups prefer to learn the theory first and then apply the knowledge with VisiStat. On the other hand, six additional participants of group A regard the PFL approach as more suitable, compared with four group B members. Five students recommend to use VisiStat twice, before and after the lecture. As students were not asked directly about their opinion regarding the sequence unless they brought up the topic, only a part of students discussed the sequence. These different opinions indicate that students have difficulties to decide which learning approach is most effective for them. As the test results show higher achievements for the PFL group, we recommend to make use of this approach in future statistics courses. However, as VisiStat is appreciated for its opportunity to practice statistical concepts, which is discussed in detail in the course of this section, further repetitions of VisiStat use after the lecture should be taken into consideration.

Moreover, it is noteworthy that students' pre-test scores are quite low although all of them had gained statistical knowledge in lectures or books. Zieffler et al. [2008] found out similar results in their studies. These findings stress once again the necessity of new methods in statistical education and illustrate the difficulties students have to learn statistics. In a second step, the results for the different learning tasks are analyzed regarding the PFL approach and students' feedback.

Results for learning tasks

PFL is supported by learning tasks

We have seen that the PFL approach is successful for the overall results. Can this effect also be observed concerning the different learning questions in the tests? In summary, in the three questions types which yielded to significant results in the post-test, higher scores can be detected in favor of group A, indicating once again the advantages of the PFL approach. The other questions do not allow unambiguous

conclusions, especially as four question types are not representative due to their lack of amount of questions. In the course of the following subsection, these results and their implications are presented.

Preparation in VisiStat supports remember factual knowledge

On the level of factual knowledge, a marginal higher score for the PFL group can be measured for *remembering* questions after the post-test. These results could suggest that participants tend to remember facts, for example the three assumptions for parametric significance tests, easier when they have observed them in VisiStat first and then learned them in the lecture again. About 90% of group A students reported that they were prepared by VisiStat for the lecture so that they were familiarized with terms and concepts and knew what is important to concentrate on. This preparation might have helped them to figure out on which concepts they should focus on in the lecture, remembering these terms. A possible explanation, why the results for this question are comparably low, could be attributed to the fact that students did not have the possibility to go through the slides again and were only introduced to the concepts twice. Regarding understanding, the results signify the crucial meaning of the lecture for *understanding* factual knowledge. However, as only question addresses this dimension, the differences should only be interpreted cautiously.

Higher achievements for PFL group regarding understanding conceptual knowledge

The post-test revealed significant differences for *understanding* conceptual knowledge in favor of the PFL group. These findings suggest that the participants could benefit more from the lecture after using VisiStat in contrast to learners in group B whose scores did not improve after the lecture. Students' qualitative feedback showed that they attribute basic knowledge and a first overview to VisiStat, whereas they regarded the lecture decisive for their deeper understanding of statistical concepts. These impressions cannot be regained by the test results as no differences between VisiStat and lecture could be identified in the mid-test. However, their feedback might suggest that it easier to get a first overview in VisiStat for understanding the concepts in the lecture. Exploring VisiStats prepares for the lecture by conveying basic knowledge, allowing the lecture to go into depth. *Analyzing* statistical concepts results in preliminary evidence for the usefulness of VisiStat as stu-

dents in each group reach higher achievements after using VisiStat. Once again, the evidence of these assumptions is not sufficient, because this type only consisted of one question.

PFL learners are better in understanding and evaluating factual knowledge

Analyzing the knowledge of statistical procedures, students in group A significantly outperformed the traditional learners concerning *understanding* and *evaluating*. When asked about the testing of assumptions, students appreciated the visibility of these assumptions and the procedure in VisiStat. The visual checking of assumptions could be connected to the test results which appeared after the assumptions were calculated so that students could recognize a pattern of this procedure. Furthermore, they evaluated the close combination of graph and results as strength of VisiStat, enabling them to form hypotheses about the connection between situation, graph, and result. These opportunities might be the cause for participants' better skills to *evaluate* procedures when using VisiStat. Students could then benefit from the lecture, to which they attributed to provide a structured overview of statistical tests. Furthermore, group A learners mentioned that the lecture could explain open questions arisen in VisiStat. As a result, the findings for the overall test results as well as the different learning questions indicate the usefulness of the PFL approach for learning statistics in a limited exposure of lecture and VisiStat. In a next step, the PFL approach is investigated regarding Cairns' four problems of statistical analysis in HCI research.

4.3.2 How VisiStat addresses Cairns' Four Problems

Cairns [2007] revealed four main problems, HCI researchers struggle with when using statistics. To prevent students from committing these mistakes, the user study focused on students' improvements after each treatment. Summing up the results, students' improved among all dimensions but further learning or practicing is necessary to help them achieve sufficient knowledge. Apart from reporting, group A students outperformed the traditional

learners in the other three problems, providing a preliminary evidence for the effectiveness of the PFL approach. In the following, the results from the tests for each of the four problems are analyzed and merged with students' qualitative feedback. In the end of this subsection, the results for the general questions part in the tests are presented as well, even though not named as one of the main problems. In order to figure out the different strengths and weaknesses of both VisiStat and lecture, the results from these students utterances in the interviews are compared to the results for Cairns' four problems.

Reporting

VisiStat is crucial for learning how to report results

Cairns' [2007] found out that insufficient reporting is the most frequent problem in HCI research. Even after learning with VisiStat and lecture, students do not even achieve a score of 20% of the reporting questions, suggesting to focus more on this topic in future teaching. A reason for this lack of knowledge could be that students have difficulties to remember all standards as they did not go through the slides again. The test findings provide preliminary evidence that the reporting function in VisiStat is crucial for gaining knowledge about reporting as the students who used VisiStat significantly outperformed the participants who attended the lecture. Although they could improve their knowledge in the lecture, the group A students were surpassed by group B students after they explored VisiStat. This assumption is supported by about 70% of each group, who state that VisiStat allows to build knowledge about reporting. In contrast to this, only about 25% of each group attribute knowledge about reporting to the lecture, criticizing that the lecture did not explain the reporting in enough detail.

Strengths of VisiStat and lecture regarding reporting

Furthermore, participants claim that VisiStat helped to create a reporting text for writing a result section. This can also be observed in the test results because students are best in creating a report directly after using VisiStat. On the other hand, especially students from group A contradict this statement, not feeling enabled to fully understand

and apply the standards for reporting after only reading the reporting text in VisiStat. In contrast to this, two participants in group A attribute explanations why values have to be reported to the lecture. These results suggest that for reporting the results it is helpful to attend the lecture first and get an overview of why reporting is important and then explore VisiStat to gain knowledge of how to write reports. A further indication for this hypothesis is that more students from group B described the reporting text as useful. A reason for these differences could be that the lecture showed only one slide of what is important about reporting results, whereas students created an automatic report several times in VisiStat.

Possible improvements

To improve students' reporting skills, we recommend to give them the possibility to practice as discussed in the previous subsection. The low results could be improved by asking students to create a reporting texts on their own and then compare it with the result in VisiStat to enable them to construct their own knowledge. To address VisiStat's inadequate ability to convey the meaning of the standards of reporting, a simple description of the characteristics of reporting results at the top of the reporting view could be inserted to explain to particistudents the aim of the reporting text. Moreover, it is interesting that more than 40% in group A and about 60% in group B used the reporting for understanding the results although this was not the intention of the report function. This behavior is discussed in detail in the following subsection.

Assumptions

Significant better results for PFL group

The result for checking assumptions is the most definite as students in group A score double of group B's results. Participants in group B can only improve slightly after exploring VisiStat, not achieving the score group A learners reached after using VisiStat. In contrast to this, the members of the PFL group accomplish nearly 50%, which is an improvement of more than 15%. These results provide a strong evidence that the sequence of treatments provides an advantage in favor of the PFL approach, implicating

that on the one hand, students gain more knowledge about assumptions from VisiStat than from the lecture. On the other hand, students benefit more from the lecture and its explanations of assumptions after they have explored the assumptions in VisiStat first.

Qualitative feedback supports test results

These conclusions are supported by students' qualitative feedback. More than 80% from group A state that they could remember two assumptions easily due to their visual representation in VisiStat. Before the test was chosen, the assumptions were visually checked by VisiStat so that students probably detected a connection between fulfilled assumptions and the chosen test. In the traditional learning group, 50% agree with this statement, indicating that participants without previous treatment might have focussed more on this visualization. Another reason for this difference could be that participants already knew about the assumptions from the lecture and therefore, did not pay attention to it. Students ascribe different strengths and weaknesses to VisisStat and the lecture, illustrating how VisiStat and lecture can complement each other. Whereas more than 55% of group A and 33.33% of group B participants state that the lecture explained the assumptions, students, especially in group A, could not develop full insight into the kind of connection between test and assumptions, not being able to understand the meaning of the assumptions. These statements suggest that learners in group A could develop hypotheses of the meaning of assumptions and but were not able to confirm these hypotheses. Nonetheless, this first examination of the topic, prepared them for the lecture, which then explained the meaning. However, students in group B could benefit less from the lecture as 40% claim having difficulties to understand the meaning of assumptions in the lecture, which was only reported by one quarter of group A students. Furthermore, two group B learners complain not having discussed the consequences of violated assumptions detailed enough in the lecture.

Integrating the third assumption in VisiStat

In conclusion, not the visibility of the assumptions in VisiStat alone is responsible for the gain of knowledge but the correct sequence of treatments is crucial for the successful learning process. Eventually, only two members of group B recognized the third assumption (interval data) in

VisiStat. The scale of data can only be selected at the first screen. Due to students' lack of knowledge regarding the interval data, it might be useful to show this assumption explicitly with the other assumptions. Although the results for the assumptions part in the test are comparably good, there is still room for improvement. Once again, students' knowledge could be improved by letting them practice the assumptions on their own.

Over-testing

Low knowledge about over-testing

The test results dealing with over-testing contradict students' qualitative feedback about their knowledge estimation. Whereas the test results show comparably good scores for group A and average scores for group B, more than 40% claim to have no knowledge about over-testing at all, leaving over-testing to be the mistake understood least. However, the differences in the test results can be explained in students' answers as 50% of group B learners claim to have no knowledge whereas only about 40% in group A address this problem. Positive feedback about knowledge concerning over-testing are reported by three group A and two group B participants, who attribute this knowledge to the lecture. Furthermore, two students in each group became aware of over-testing in VisiStat and explained that the system recommended to use post-hoc tests to avoid over-testing. On the other hand, about 30% of PFL group members observed over-testing in VisiStat but could not entirely understand it. The lack of students in group B describing this problem suggests that students in group B were aware to use ANOVAs instead of *t*-tests and therefore, did not encounter the over-testing warning. Furthermore, it is interesting to notice that most students in both group did not observe over-testing in VisiStat which might be an indication that VisiStat encourages them to use tests appropriately and prevents them from committing over-testing, which was communicated by some students as well.

Over-testing was barely encountered by students in VisiStat

Although this is a positive effect, almost thirty group B and forty group A claim to have difficulties to entirely anticipate the risk of over-testing in the lecture. These findings

might provide preliminary evidence that not exploring a problem in VisiStat makes it difficult to understand it in the lecture, supporting the PFL approach. Additionally, it can be observed in the test results that PFL learners, from which more students encountered over-testing in VisiStat, outperform the traditional learning students, stressing once again the advantage of the PFL approach. Large variances among students in both groups point out the differences between the knowledge levels and could be related to the exploring of over-testing in VisiStat. However, as only two students in group B report to have observed over-testing in VisiStat, it is surprising that their knowledge could improve nonetheless. A reason for these contradictions between students' feedback and test results could be that the over-testing section in the tests consisted only of two questions, of which one dealt with the use of ANOVA versus pairwise *t*-tests. Consequently, it might have to be differed between knowledge of over-testing in general and the understanding that ANOVA is preferred over pairwise *t*-tests. As found out by Garfield and Ben-Zvi [2007], it is important, not to overestimate the understanding as students might perform not bad in a test but have not understood the underlying principle. Furthermore, one students describes that over-testing contradicts his or her intuition which is well-known problem when learning statistics [Konold, 1995], indicating that over-testing is difficult to learn. To overcome these obstacles, it might be a possibility to push students with the tasks to explore the risk of over-testing in VisiStat and spent more on underlying problem regarding over-testing in the lecture.

Appropriate Testing

PFL students score higher reading appropriate testing

Regarding the choice of the appropriate test, the results provide preliminary evidence that students benefit from the PFL approach. Even after group B students completed both treatments, they cannot catch up with group A's result after only using VisiStat. In contrast to this, group A participants' knowledge increased about almost ten percent after they attended the lecture. More than three quarters of learners in both group attribute their knowledge about ap-

propriate testing to the lecture. 50% of group A students claim to have learned some basic characteristics of appropriate testing by using VisiStat. In group B, two thirds name having developed knowledge in VisiStat. These results might suggest that participants in group B are more aware of the strengths of VisiStat for their learning although students in group A benefited more from it. Furthermore, the different scores might be an indication that students can take more advantage of the lecture after developing their own hypotheses in VisiStat.

Dissatisfaction of current statistical knowledge

However, students criticize learning about appropriate testing in lecture and VisiStat as well, revealing their dissatisfaction with their current knowledge. Nearly 70% of group A members claim that VisiStat did not provide them sufficient knowledge about appropriate testing, in group B, at the beginning of 60% agree with this disapproval, making this the most frequently stated weakness of VisiStat. Concerning the lecture, about 40% of group A and about 30% of group B students comment on its inadequate contribution to their knowledge of the appropriate choice of tests. These results reflect that group A members seem to be more dissatisfied with their knowledge than group B participants, contradicting to the test results. On the one hand, this contradiction indicates the difficulties students have to estimate their knowledge. On the other hand, it reveals students, especially in group A, perceive choosing an appropriate test as difficult as they claim the knowledge gain in both treatment as insufficient. Furthermore, it is possible that the PFL approach prepares students with more questions about appropriate testing, which were not answered in the lecture. This perception can also be recognized when half of group A and 40% of group B students ask for more time for the different tests and their characteristics in the lecture.

Suggestions for improvement

As more than 80% group A and even almost 95% group B participants would have preferred to have more time in the lecture for the content, it could be helpful to split in the lecture into two parts and focus more on appropriate testing. Furthermore, as discussed in the previous subsection, students could be asked to fill out an assignment sheet after the first lecture and then get feedback about their

achievements, enabling them to estimate their actual statistical knowledge. Additionally, it might be a possibility to confront students with the solution VisiStat provides. In this case, students become aware of their current misconceptions and can focus more on their problems in the second lecture, as some students also reported successfully.

General Questions

Lecture is crucial for gaining general basic knowledge

Apart from Cairns' four problems, a fifth dimension was collected in the statistical knowledge tests, which dealt with general questions about basic statistical concepts. The findings suggest that in this case the lecture is crucial for gaining knowledge about basic statistical concepts. These results can be supported by students' qualitative feedback as students in acknowledge the lecture's clear and consecutive structure, which dealt with basic concepts first and then moved on to more complex topics. Furthermore, about 20% of participants in both groups emphasized that the first, introductory part of the lecture was good. In contrast to this, students started directly with performing statistical analysis when using VisiStat without preparation. However, after group A students attended the lecture, they could marginally surpass group B students in the post test, indicating that students could benefit even more from the lecture after they were prepared for it by exploring VisiStat first. These findings suggest that VisiStat offers the opportunity to explore basic statistical concepts without specific mentioning of these.

4.3.3 VisiStat's Role

In Chapter 3, Garfield and Ben-Zvi's learning principles [2007] were investigated regarding how they are fulfilled by a book, lecture, and an interactive analysis system, such as VisiStat. The fulfillment of these principles for a lecture and VisiStat are reexamined in this subsection, taking students' feedback into consideration. In the following sections, students' qualitative feedback towards each principle is presented.

Constructing knowledge

The ability to construct knowledge was only attributed to an interactive analysis system and not to a lecture in Chapter 3 as a teacher can only tell students about statistical concepts and students have to construct the knowledge themselves. Students' qualitative feedback supports this speculation. However, there some interesting findings which are discussed below.

VisiStat encourages to develop hypotheses about statistical concepts

Nearly three quarters of group B and more than 60% of group A participants stated that it is one of VisiStat's fundamental strengths to encourage students to develop hypotheses about statistical concepts, for example assuming the responsible characteristics of the data for a specific test choice. It is surprising that more members of group B name this advantage of VisiStat as one could have expected that students who explore the system without previous knowledge treatment tend to form more assumptions. Similar results are indicated by the video observations but further analysis is necessary to be able to interpret these results. The differences is emphasized by over 60% of participants in group A who claim that they were not encouraged to think about solutions in VisiStat themselves as VisiStat calculates the results automatically. In contrast, only 17% of members in group B agree with this statement. These results contradict Schneider et al.'s utterance analysis [2013]. A possible explanation could be the lack of freedom to try different tests on each variable combination. Moreover, participants' reflections in the interviews are limited and might differ from the actual utterance analysis of students' interactions with VisiStat.

Two students in group A state that VisiStat does not provide to explore statistics but also follows a telling approach as it calculates all results automatically. These findings suggest that group A students perceive it as more difficult to develop hypotheses. A possible explanation could be that the participants did not know what they should focus on in VisiStat, which was reported by four students in group A, claiming not to know the learning aim. Group B students on the contrary knew the main statistical concepts

they should explore in VisiStat, being able to use the gained knowledge to develop assumptions in VisiStat. However, this different assessments might be the result of divergent perceptions as the test results provide a preliminary evidence that students in group A are able to form assumptions when exploring VisiStat and benefit from this first learning in the lecture. In this case, group A participants' dissatisfaction should be addressed by giving students the possibility to interact more with the system and thereby, develop more assumptions.

Interactivity should be extended for VisiStat

At the current state of VisiStat, about 60% in both groups praise the interactivity VisiStat allows, being able to observe the results for different inputs. On the other hand, 60% of group A and about 40% of group B participants think that the system still lacks the possibility to interactively change the input, for example the variables or selected tests, so that the corresponding result can be connected to the input. Instead of the automatic checking of assumptions and choice of appropriate test, students want to be able to choose the test themselves and get feedback by the system about their performance. These statements indicate that VisiStat is generally able to encourage students to construct knowledge but can be improved to support students to tackle with the statistical concepts in more depth. One possibility to improve their learning is to implement a functionality in VisiStat so that students are forced to submit an estimation and then get feedback by the system if they performed correctly. This could solve the problem that hypotheses cannot be confirmed with VisiStat as well. A second and easier opportunity is represented by giving students questions to solve when exploring VisiStat. In contrast to the tasks in this user study, students should be asked to answer questions, for example about the appropriate test, *before* they see the answer in VisiStat and only check their answer against VisiStat's solution. However, students cannot receive an explanation why their solution is not correct in this variant. Nonetheless, it might be promising to try this easier possibility first, testing its success, and have the lecturer address arising problems.

Role of lecture and VisiStat

How VisiStat and lecture complement each other in order to construct knowledge can be observed in the following

statements. Whereas VisiStat enables participants to yield hypotheses, half of group A and one third of group B students criticize that these developed assumptions cannot be confirmed with the help of VisiStat. A possible reason for this could be the insufficient help description, which is discussed in detail in Subsection 4.3.5. The difference between the two groups concerning this statement could be explained by the fact that group B students had already developed knowledge in the lecture and therefore, are able to confirm their assumption easier. To confirm or reject the hypotheses elaborated in VisiStat, the lecture can be used, explaining VisiStat's behavior to half of group A students. Additionally, two group B participants regard the lecture useful for this approval. More than two thirds in both group A and B estimate to have gained deeper statistical knowledge in the lecture. This result contradicts the test results which attribute more understanding developed through VisiStat and the possibility to construct knowledge only in VisiStat. These findings could indicate that students are uncertain about their achievements after using VisiStat and are more familiar with the traditional learning situation. To overcome this obstacle, students could be given feedback, which is also discussed in depth later this section, about their learning gain. Moreover, it can be confirmed that VisiStat enables students to construct knowledge but could be developed further to actually encourage students to form assumptions, which is considered in the following subsection. Regarding the lecture, students do not claim to be able to construct knowledge in the lecture but it can be used to confirm or reject the hypotheses developed in VisiStat, which could be supported more so that all open questions can be answered. A possibility to improve the learning experience, could be that students interactively ask the instructor about their open questions or students are asked to fill out an assignment sheet after using VisiStat, giving the instructor the possibility to address wrong misconceptions in the lecture.

Active involvement

To improve learning statistics, the possibilities to generally actively involve students are given in VisiStat as well as in the lecture. As analyzed in the previous section, students can interact with VisiStat to observe statistical calculations and form hypothesis. Evaluating the strength of the lecture, almost 40% of group A and more than 20% of group B participants describe that lecture flexibly addresses students' need by giving them the possibility to ask until they understand a concept. However, Garfield and Ben-Zvi [2007] report only literature dealing with group work to actively involve students. Assessing the use of group work to learn statistics, the lecture as well as VisiStat do not fulfill this principle. VisiStat does not offer students a group mode, enabling them to discover VisiStat together. Group work with VisiStat is only possible when several students use one version of VisiStat together. Despite the constructive interaction used in our study, students did not make any statements whether they think they could benefit from exploring VisiStat together, which might be interesting to assess in a following study. However, about 55% in each group criticize the lack of an interactive practice part in the lecture, indicating that students miss the active involvement part. Consequently, future lectures could consider active involvement in form of small group works for the lecture. Furthermore, about one quarter of all students proposes to use VisiStat in class. It might be possible to have students develop answers for specific tasks together in small group with each group using VisiStat to control their solutions.

Consider active involvement in small groups for lecture and VisiStat

Encourage Practice

Practicing is main strength of VisiStat

The third principle requests teachers to encourage students to practice the learned statistical concepts in varying ways. More than three quarters of students ascribe this ability to VisiStat so that, in students' opinion, the possibility to apply and practice learned statistical concepts can be considered as VisiStat's main role. It is noteworthy that all group B participants appreciate this strength, whereas in group A this skill is only mentioned by half of the mem-

bers. This could be put down to the cause that students in the traditional group used VisiStat after the theoretical input to practice these concepts whereas participants in the PFL group did not have the necessary knowledge to actually practice. Consequently, these results provide a strong evidence that VisiStat is suitable for practicing. Additionally, Garfield and Ben-Zvi [2007] stressed the necessity of a hands-on experiment to encourage practice. Participants attribute VisiStat this advantage as well, acknowledging that VisiStat allows to do something on their own, like experiment with the data, and going into depth. Once again, mostly group B students recognized this opportunity. The test results indicate that students benefit most from the learning experience when they explore VisiStat before the lecture. However, as Garfield and Ben-Zvi [2007] emphasize practicing as an important part of the learning experience, students' qualitative feedback gives reasons to include VisiStat a second time in the learning process as a tool for practicing and applying concepts. This could be helpful because half of group A participants complain about their lack of knowledge to apply statistical concepts after using VisiStat, whereas only one fifth of group B students report this weakness. By giving students the opportunity to benefit from VisiStat as a tool for applying statistical concepts, this practicing could enable them to gain further knowledge.

Lecture should include practice part

Regarding the lecture, it is frequently criticized by students of both groups that it did not contain a practical part, preventing them from deepening the knowledge. Moreover, two students claimed that the lecture did not enable them to apply concepts, which could be overcome by more practice as well. On the other hand, students lamented that one lecture is not enough time for the amount of content. Thus, a time-consuming practice part during the lecture does not seem to be a promising solution. Instead, students could be asked to practice at home by using VisiStat. As several participants in both groups described to need more pressure, like an exercise sheet, to be encouraged to learn statistics, the practice should be accompanied with an exercise sheet. To liven up the lecture and make it more interactive, students' suggestion to use VisiStat during the lecture after a theory part could be taken into consideration. This recom-

mendation is supported by students' quantitative feedback, which rates VisiStat as fun and enjoyable. In the interview, about 45% of group B and more than 30% of group A members appreciate VisiStat as a satisfying experience. These results and the finding in the questionnaire, in which group B students evaluated better in general, suggest the preliminary conclusion that students using VisiStat for practicing are more satisfied with their learning experience. However, as only small differences can be observed at the moment, this should be investigated in more detail.

Be aware and confront with errors

VisiStat does not provide enough knowledge to confront with errors

Literature shows that students suffer from various misconceptions about statistical concepts (cf. Chapter 2.2. To overcome these misconceptions, Garfield and Ben-Zvi [2007] recommend to let students form assumptions about the meaning of a concept first and then contrast their results with the actual meaning. More than 60% of participants in both groups explained that VisiStat supported them to develop hypotheses about statistical concepts. But can VisiStat help to confront students with their misinterpretations? In the video observations, we found that students are confronted by the automatically calculated results, the help text, the visualizations, or the reporting text after they formed predictions of VisiStat's statistical behavior. However, whether these confrontations lead to correct conclusions is yet to be investigated. This can be done in an in-depth analysis of session recordings as already proposed for three videos in 4.2.2. We recommended to do this in-depth examination in future analysis of the data.

Nevertheless, students' qualitative feedback indicates that they used the reporting text to understand the results and could draw explanations from VisiStat's help function. However, about 45% of group A and 30% of group B students stated that they were not able to confirm these hypotheses in VisiStat. Furthermore, both groups report that the help description is not detailed enough, which is discussed in more detail in Subsection 4.3.5. The difference between the two groups suggests that it easier for group B

students to confirm or reject their hypotheses as they have already gained previous knowledge from the lecture.

Students want to try and error

Moreover, students, especially half of group A, criticize that they are not able to try and error when using VisiStat as VisiStat automatically applies the correct statistical concepts. Instead of using VisiStat to validate the assumptions, the lecture could explain the underlying concepts and thus, confront students with errors. Half of group A students report that the lecture revealed reasons for VisiStat's behavior. Additionally, the test results suggest that the lecture is able to confirm or reject group A students' assumptions. However, students' qualitative feedback indicates that there is still room for improvement and not all open questions can be answered in the lecture. The results of the quantitative feedback questionnaire imply a similar impression as well, revealing that PFL students are more dissatisfied with the lecture although nearly all students in group A report that they could follow the lecture easier as they have been prepared for it by using VisiStat.

Suggestions for improvements

These results as well as students' direct feedback indicate that they would prefer to be able to interactively try and error in VisiStat, for example by letting them choose the test and then get detailed feedback from the system why this test choice was correct or not. However, this would require to change VisiStat, making it an intelligent learning question-answer tool. Before this approach is put into practice, students could be asked to form predictions and write them down before VisiStat automatically calculates the assumptions and the appropriate test. Furthermore, the help description should be extended and offer the possibility to look up all concepts so that students, who suspected to use a specific test, but realized that VisiStat uses another one, are able to contrast the descriptions of both tests. A further advantage of this approach is that students are *forced* to think about the solution as currently nearly half of group A members stated that they did not try to learn with VisiStat but only wanted to solve the tasks. To support the lecturer to in uncovering students' misconceptions, students could submit their sheet with assumptions from the first exploration of VisiStat so that the instructor detects common mistakes and can deal with these problems in depth in the lec-

ture.

Do not underestimate the difficulty

Mixed opinions regarding the level of difficulty in VisiStat

Several researchers found out that students have severe difficulties with learning statistical concepts (cf. Chapter 2.2). Regarding VisiStat, students have different opinions if VisiStat underestimates the difficulty. On the one hand, they regard it as positive that VisiStat can be explored at individual speed in contrast to the lecture, where members of both groups complained that it was difficult to follow the lecture. Furthermore, almost half of group A students appreciated the low necessary background knowledge required for using VisiStat. Due to the automatic application of correct statistics, some students state that VisiStat compensates for their lack of statistical knowledge. As a consequence, even students without considerable previous knowledge can learn with VisiStat so that it can be assumed that VisiStat does not underestimate students' difficulties. On the other hand, half of group B and more than 60% of group A participants criticize the help description as not detailed enough. These findings indicate that students, especially when not attending the lecture first, need more help description to understand the results. Moreover, it is reported that the help description is too difficult as it does not use easy language and does not provide an example. Consequently, it is recommended to elaborate and improve VisiStat's help description, making it easier and more detailed to address students' difficulties.

Most students found lecture well explained, some had difficulties

About 40% of students in both groups appreciated that the content in the lecture was explained well and easily. The clear consecutive structure, which introduced basic concepts first and then moved on to more complex topics, helped students to follow the presented concepts. But there is still room for improvement and some students disagree with these strengths. One quarter of group A participants as well as one group B student claim that the presentation of content was too difficult, contradicting the other students who praised the presentation. It is interesting that this is mentioned mostly by group A students who

have been prepared for the lecture. Furthermore, three students affirmed that they did not have the necessary previous knowledge to follow the lecture. Thus, most students did not feel overtaxed by the lecture but there are students who had difficulties. Splitting the lecture into two parts and allowing weaker students to repeat the first part of the lecture before attending the second part could help to overcome the current problems.

Do not overestimate the understanding

The analysis of students' qualitative feedback does not allow to draw conclusion whether VisiStat or the lecture overestimate students' understanding. Regarding the overall learning experience, the findings suggest that this limited exposure of VisiStat and lecture is not sufficient to provide them an adequate statistical knowledge for research. This assumptions is supported by the test results, which show that there is still room for improvement. Students' feedback indicate similar assumptions as students are not satisfied with the knowledge they gained after using VisiStat, attending the lecture, and both treatments. Furthermore, they emphasize that further learning is necessary to use statistical concepts in own work or pass the exam. These concerns could be addressed by picking up participants' suggestion to split the lecture into two parts and include more practice.

Give consistent and helpful feedback

Students want to get feedback in VisiStat

The importance of getting helpful feedback can be observed in students' qualitative feedback as especially group A participants complained not to have gained sufficient knowledge in VisiStat and the lecture. These statements contradict the test results, which suggest that the PFL students benefitted more from the learning experience. These results might unveil group A students' dissatisfaction with their statistical skills. Claiming that they still lack fundamental knowledge, almost 40% of group B members show frustration as well. A possible reason is that students were still un-

certain about many questions when filling out the post-test and did not get feedback about their achievements. To enable students to estimate their knowledge and show them their improvements, detailed feedback should be given so that students get to know their scores. Detailed feedback could also help to prevent students form developing a negative attitude towards learning statistics and regarding it as too difficult, which is currently expressed by five students.

To include feedback in the learning experience, self-assessment tests after using VisiStat can be used, which is recommended by Ardito et al. [2006]. However, the use of assignment sheets after the learning experience to practice applying statistical concepts is another possibility. After evaluating students' skills in these test, the lecturer can address the occurred problems. Students in both groups named the advantage of the lecture to flexibly address students' needs as students can ask until they understand. These statements suggest that it is possible to receive helpful feedback in the lecture.

Technology to visualize and explore data

VisiStat is suitable for learning statistics

Garfield and Ben-Zvi [2007] found out that technology can be useful for learning statistics, but stress that teachers have to take advantage of the system's strengths, like visualizations and the possibility to let students explore the data. The presented results indicate that VisiStat is suitable for learning statistics. These findings are supported by participants' statements. More than 60% from group B and almost 70% from group A students describe VisiStat's advantage of visualizing statistical concepts. Additionally, about 60% of both group members acknowledge VisiStat's interactivity to observe the results, especially different graphs, after changing the input. Apart from the visualizations, students describe the exploration of VisiStat as a hands-on experiment. However, especially group A students ask for more independent interactivity to be able to change the input data on their own and observe the corresponding changes. Due to this lack of interactivity, two group A students criticize that VisiStat does not allow exploration but tells re-

sults, comparable to a lecturer. In this case, the positive feedback outweighs the negative concerns. Nonetheless, as argued before, it might be taken into consideration to enhance the interactivity of VisiStat, offering students the opportunity to explore VisiStat in more depth.

It is evident that a traditional lecture does not fulfill this principle as students cannot explore data on their own, while the instructor introduces new concepts. However, the use of a visualization demo in the lecture was widely appreciated by students. Furthermore, participants characterize the lecture as well presented and acknowledge the use of a visualization of the test decision tree as well as the combination of visual and auditory explanations.

Summing up, the findings indicate that the assumptions about fulfilled learning principles in lecture and VisiStat (cf. table 3.1) can be confirmed. On the one hand, it has been seen that students criticize parts of lecture and VisiStat, in which the learning principles are not applied. On the other hand, these results suggest that VisiStat and lecture can complement each other because they compensates for each other's weaknesses. However, there is still room for improvement and Garfield and Ben-Zvi's learning principles [2007] can help to translate these suggestions into practice, improving students' learning experience.

Strengths and weaknesses of VisiStat

VisiStat is easy to use

In the previous sections, VisiStat's role regarding the sequence of VisiStat and lecture was analyzed, which suggested to use VisiStat before the lecture. In the following, VisiStat's strengths and weaknesses are summed up, examined and determined, evaluating how these already complement a traditional lecture and where there is still room for improvement. The results of the feedback questionnaire suggest that VisiStat is easy to use, supported by students' qualitative feedback. The findings show a better evaluation on the part of group B in qualitative as well as quantitative feedback dealing with the intuitiveness of VisiStat, implying that it is easier for users with previous knowledge to

navigate in VisiStat's user interface. Students suggested several improvements, which are not presented here but considered for further development on VisiStat.

VisiStat is useful

In general, the results of the feedback questionnaire suggest that VisiStat is regarded as useful by students for learning statistics. Four students would like to learn with VisiStat again to prepare for the exam. Furthermore, VisiStat is attributed to be useful for conducting research, when writing a thesis for example. These findings support the previous findings that VisiStat is suitable for learning statistics. Regarding the lecture, participants replenish that the lecture helped them to finally understand the result section of research papers.

Main advantage: practicing

The most frequently stated advantage of VisiStat is practicing and applying statistical concepts by experimenting with the data on their own and at individual speed. Especially group A students state that VisiStat provides information to learn by exploring. As the quantitative results from the feedback questionnaire imply that students enjoyed the use of VisiStat as it made fun and the times flew during the exploration. On the opposite, students were not completely absorbed during the learning experience. A possible explanation is that participants did not perceive the exploration as challenging or that the work with a partner and the resulting distraction prevented them from completely absorbing.

Strength: visualizations

VisiStat's continuous visualizations of statistical concepts are regarded as another crucial strength. By comparing different situation and graphs, it could be observed that students interact with the visualizations and try to predict VisiStat's behavior and form hypotheses about statistical concepts. However, as the first analysis of three videos indicated, students were not always able to contrast their prediction with the actual solution and draw conclusion or generate rules. These observations are also supported by students qualitative feedback, which reveals the central critic of low interactivity. Students complain about not being able to try and error or the lack of possibility to interactively change data and observe the corresponding outcome. As a consequence, almost half of group A students

state that VisiStat did not convey all statistical concepts so that they did not gain satisfactory statistical knowledge. These findings contradict the test results, as observed several times in this chapter, which could be an indication that students are dissatisfied that they cannot verify their assumptions only by using VisiStat. Another reasons could be the lack of feedback so that students were not aware of their learning gain. However, these results underline once again the necessity of a lecture to complement VisiStat.

Lecture can compensate for VisiStat's weaknesses

The lecture can be positioned at these weaknesses of VisiStat, being attributed to provide deeper knowledge. Moreover, students claimed that the lecture can explain VisiStat's behavior and flexibly addresses their open questions. The most often stated disadvantage of the lecture is perceived to be the amount of content for one lecture so that most students ask to split the lecture into two parts. To improve the lecture's structure further, the participants recommend to have more time for complex topics.

Interaction between VisiStat and lecture

It was shown that VisiStat and lecture complement each other concerning their strengths and weaknesses. It was mentioned that the lecture can answer open questions, occurred during the exploration of VisiStat, implying that VisiStat and the lecture are not two individual learning treatments but form a mutual learning experience. Not only the lecture can answer questions arisen when using the system, but VisiStat can also be used to check if a concept was understood correctly. Further interactions include that one treatment prepares for the other one. Almost 80% of group B students claim that the knowledge they gained in the lecture prepares them for using VisiStat. It is interesting to notice that half of group B students mention the necessity to have this previous knowledge to be able to use VisiStat. About 20% of group B students even think that they would have needed more training to be able to use VisiStat sufficiently. In contrast to this, less than 20% of group A students agree with this statement, whereas more than 40% of them thinks that it is an advantage of VisiStat that only low background knowledge is required to explore it. These remarks as well as the test results in favor of group A indicate that it is possible to use VisiStat without previous knowledge and that students can even benefit from

this, proving group B participants wrong.

Nearly 90% of group A participants reported that the use of VisiStat prepares them for the lecture, as they were familiar with concepts and could follow the lecture better. Furthermore, they stated that they focused on the solution of problems they encountered in VisiStat, which might be an explanation for the better test results of the PFL learners. On the other hand, it is surprising that they evaluated the lecture worse than group B participants in the feedback questionnaire and on several occasions in the qualitative feedback. These results could be an implication that the lecture did not answer all their questions. To overcome this problem, the lecture could be split into two parts, as asked for by several students, to have more time to deal with students' questions. Based on the different strengths and weaknesses of VisiStat as well as lecture, an overall evaluation of the two learning treatments is given in the next subsection.

4.3.4 Overall evaluation of VisiStat and lecture

The results of the quantitative feedback questionnaire show that students' overall evaluation of their learning experience is positive. More than 60% of participants in each groups stressed in the interviews that their statistical knowledge definitely improved. The overall evaluation of VisiStat is satisfying, whereas the lecture is evaluated slightly worse, especially by group A students, yet still satisfactory. These results are surprising, as students in group A are more prepared for the lecture and could develop previous knowledge in VisiStat. A reason for this evaluation of the lecture could be students' complaints that there was too much content for one lecture and they would have preferred to split the lecture into parts. Furthermore, the findings suggest that questions arouse during using VisiStat which could not be answered completely in the lecture. One could also speculate that group A learners were more knowledgable and knew which parts were missing in the lecture whereas participants in group B discovered further parts in VisiStat. Although participants in group A benefited more from the learning experience, more group B

students described VisiStat as satisfying experience in the interviews.

On the contrary, the results provide insight that the limited exposure of one lecture and one exploration of VisiStat is not sufficient for learning statistics. Nearly 70% of group A participants emphasized that there is still room for improvement and about 40% of both group members explain the necessity for further learning to be able to apply statistical concepts. These outcomes provide preliminary evidence that students in the PFL group are more dissatisfied than students in group B, which can be observed in the results of the feedback questionnaire as well as students in group A evaluated slightly worse than group A participants. Whereas more students in the PFL treatment claimed to be dissatisfied with the individual learning experience in lecture and VisiStat, almost 40% of the traditional learners reported that the overall learning experience was insufficient. How this dissatisfaction can be addressed and overcome is discussed in Subsection 4.3.3, using Garfield and Ben-Zvi's learning principles [2007].

At the moment, 40% of students assess the results of the learning experience to have provided an overview and raised the familiarity with statistical concepts, whereas the own application is still difficult. Regarding the amount of gained knowledge, more than two thirds of all participants regards VisiStat's role to give a first overview of statistical concepts, which can serve as starting point. However, some students do not agree with this limitation of VisiStat's abilities but one quarter attributes to have learnt new concepts to VisiStat. Half of the PFL and one third of tell-and-practice participants ascribe basic knowledge to the lecture. However, almost 70% of group A and two thirds of group B learners think that it is one of the lecture's main strengths to provide deeper understanding in contrast to VisiStat.

4.3.5 In search of help

A frequently discussed problem in VisiStat is the lack of sufficient help, which students report. About one third of

students in both groups claimed that VisiStat's help function offered them explanation, but more than 60% of group A and half of group B students thought that the help description is not detailed enough to understand the concepts. Being named second most of VisiStat's weaknesses suggests that this limitation is crucial for the learning experience. In Section 4.2.2, it was shown that students, especially in group A, considerable time with the help function in VisiStat. Furthermore, we have seen that students were not able to confirm or reject all their predictions. That students in group A tend to use the help function more is not surprising as they have no previous knowledge and try to get this knowledge by reading the help text. However, students' qualitative feedback indicated that the help function is not sufficient. Some participants suggested to address this drawback by including examples in the help function. Another possibility is to use easier language as students claimed that the help description was too difficult. Participants also proposed to use shortcuts to activate the help function as it was time consuming to turn it on and off again.

As the help function does not provide enough explanation, more than 40% of group A and 60% of group B participants stated that the reporting text helped them to understand the results. However, as discussed in section 3.3, the reporting text uses a low coherence structure. Therefore, it is not surprising that three students in group A name the reporting text as difficult to read. Due to their lack of previous knowledge, students in group A do not benefit from the low coherence but have difficulties to understand it. Consequently, we recommend to extend the help description to address students' need for explanation and to adapt the description to students' level of knowledge.

Summing up, it was shown that VisiStat, complementing a traditional lecture, is suitable for learning statistics. Furthermore, students benefit from the PFL approach, which suggests to explore VisiStat before the lecture to get a first overview of statistical concepts, which prepares for the lecture. These findings were supported by the test results for each of Cairns' [2007] problems. These results provide preliminary evidence that VisiStat can help to prevent

students from conducting one of Cairns' reported problems. Whereas VisiStat can encourage to construct knowledge, develop assumptions, and practice statistical concepts, the lecture addresses students' needs and explains in more depth until students understand. A combination of VisiStat and lecture is able to fulfill all of Garfield and Ben-Zvi [2007] learning principles. However, we recommend to adapt the current learning experience to students' feedback as there is still for room for improvement and the participants seem to be dissatisfied with their amount of knowledge. Eventually, we noticed that students try to contrast their own predictions with the actual answers but do not have sufficient possibilities to do so as the help description is not adequate. These findings have to be interpreted against the background of the limitations of our study, which outlined in the following chapter. Afterwards, the results are summed up again in the last chapter.

4.4 Limitations

In Chapter 4, the different methods and their appropriateness for the user study were presented. However, these methods have limitations, which have to be considered to guarantee adequate interpretations. Furthermore, minor mistakes occurred during the course of the user study, which have to be reflected in the following.

Motivation differences

At first, students' attitude towards learning statistics and the user study can influence their performance. Zieffler et al. [2008] reported several studies which showed that students' behavior and test scores depended on their motivation and attitude towards learning statistics. Students reported that an exam can increase this motivation, but as they were not graded for their achievements in the user study, they were not encouraged to invest effort in learning statistics. Furthermore, four students criticized the time of the lecture in the semester because the statistics lecture was the last lecture before the mid term exam. Although no user studies took place in the last days before the exam, group B students, who participated the days shortly after the test seemed quite unmotivated. If this motivation directly influ-

enced their exploration has not been examined yet, but can be analyzed with the video recordings. Another indication of students' different motivations is the duration of their exploration of VisiStat, which lasted from 25 to 53 minutes (after 45 minutes they were asked to come to an end). The lack of motivation might have affects on the tests, for example, many students did not try to write a reporting text on their own in the last question.

Limitations of tests

Regarding the tests, it has to be stressed that it has not been tested if the questions are appropriate and representative for Cairns' four problems. The evaluation suggested that the over-testing part did not cover the range of the underlying problem. It is recommended to revise the test and investigate its suitability for following studies. As students did not score more than 30% on average and seemed to be dissatisfied with their current statistical knowledge, it might be considered that the knowledge test was too difficult and could be adapted correspondingly. Moreover, two minor mistakes occurred regarding the knowledge tests. At first, the last question in the mid-test made use of different examples in the text and corresponding table, which was changed after one students noticed the mistake. As a consequence, answers with both examples were accepted in the evaluation. Secondly, the post test included two questions in the appropriate testing part, whose correct answers were two-way ANOVA, whereas the other two tests consisted of two one-way ANOVA questions and only asked about two-way ANOVA once. This difference might have affected the test results as students used more one-way ANOVAs in VisiStat.

Effects due to patner

Another possible influence is given by the exploration and interview together with a partner. It is possible that students did not want to admit difficulties when learning cooperatively and are blocked in their progress by the partner. On the other hand, a motivated partner could also have an encouraging affect, which would not have occurred without the collaboration. Moreover, students tended to form a mutual opinion so that teaming up with another partner could have leaded to divergent feedback.

Interviewer effect

Similar problems can have been evoked by an interviewer

effect so that students felt observed during the learning experience. Although students were promised that the results were treated completely anonymously, they have had fear to be graded and judged by the instructor. Furthermore, it is possible that students did not want to criticize their instructor or wanted to help the investigators and therefore tried to assume what the correct answers [Lazar et al., 2010].

External factors

Time constraints due to organizational reasons prevented to have the same period of time between exploration of VisiStat and lecture for all students. A pair of students exploring VisiStat the day before or after the lecture might have remembered more knowledge from the first treatment than students who had a break of a week between the two learning experiences. In addition, students in group B conducted the interview directly after using VisiStat whereas students in group A had to be interviewed on a second date after they attended the lecture. This different point of time could have influenced students' opinions and memory. However, as students in group A provided a lot of feedback, this criticism can be disregarded. After exploring VisiStat, students directly filled out the test. To make the results comparable, students were asked to answer the test after the lecture within 24 hours after the lecture. One student forgot to do the test and did it later. Furthermore, it could not be checked if students answered the questions without help.

Small sample size

Due a sample size of 34 regarded participants, only first assumptions about VisiStat's role in a learning experience can be developed, which have to be reexamined in following studies. Furthermore, these studies could detect possible third influences which were responsible for the different test results of the two groups. The influences presented in this chapter have to be taken into consideration for the evaluation of the results. Against the backdrop of these limitations, the next chapter sums up the results of the user study and proposes future works.

Chapter 5

Summary and future work

This thesis tried to determine the role of the interactive analysis system VisiStat on learning statistics. Therefore, it was investigated how VisiStat can complement a traditional lecture. Furthermore, VisiStat's ability to prevent students from conducting common mistakes in HCI statistical analysis was examined. In the course of this user study, 36 HCI students were asked to evaluate their learning experience with a limited exposure of VisiStat and a lecture. In this chapter, the findings are summed up and against the backdrop of the research questions evaluated. In the second part, potential future works are discussed, which can follow this user study.

5.1 Summary

Addressing problems of statistical analysis in HCI research

Cairns [2007] found out that HCI researchers struggle with statistical analysis in their works. Insufficient statistical education was identified as the underlying problem. Chapter 2.2 showed that several researchers of various disciplines investigated the problems and misconceptions students on college level have regarding statistics. Garfield and Ben-Zvi [2007] elaborated eight learning principles based on lit-

erature to improve learning statistics. We claimed that a combination of lecture and VisiStat fulfills all learning principles and can help to improve statistical education among HCI researchers (Chapter 3.2. Chapter 2.3 revealed that the use of technology to enhance learning experiences has already been used successfully in different fields. In statistical education, e-learning tools proved to be promising, as well. However, these studies did not address advanced statistical analysis, which tries to overcome Cairns' four problems. Our user study tries to overcome this gap and provides an in-depth analysis of students feedback regarding their learning experience to determine the reasons for possible strengths and weaknesses of lecture and VisiStat.

Experimental Design

To evaluate the role of VisiStat to complement a traditional lecture, a similar approach to Schneider et al. [2013] was chosen (Chapter 4.1.1. Students were divided into two groups, receiving two different treatments. The first group (A) followed the preparation for future learning approach, which was presented in Chapter 2.4, exploring VisiStat without previous knowledge attending the lecture afterwards. Contrary to this, the group B attended the lecture first to learn the theory and then practiced with VisiStat (tell-and-practice). To evaluate students' knowledge gain, participants were asked to fill out tests after each treatment. Eventually, a feedback questionnaire and qualitative interview aimed to assess students' feedback and evaluation.

Automatically generated report in VisiStat

The tested version of VisiStat addresses all four of Cairns' problems in order to prevent users from conducting inappropriate statistical analysis. The automatically generated report texts was developed as part of this thesis (Chapter 3.3. Based on APA's standard guidelines [2010], a sufficient set of necessary values was determined. With the help of Sandig's pattern for text types [1997] prototypical characteristics of reporting texts were elaborated. Jakobs' principles of communicative usability were applied to ensure the text's comprehensibility. These requirements resulted in an automatically generated and sufficient reporting text.

VisiStat is suitable for learning statistics

The evaluation of participants' test results, their quantitative as well as qualitative feedback, and a first insight into the observations of students' explorations indicates

PFL approach is promising

VisiStat's suitability for learning statistics. The test results suggest that students benefit most from VisiStat if they explore it first without previous knowledge and then attend a lecture, supporting the Preparation for Future Learning approach.

VisiStat helps to prevent Cairns' four problems

Furthermore, VisiStat can help to prevent students from committing Cairns' four major mistakes in HCI research. In case of three of the four problems, students in the PFL group outperformed students in the traditional learning group. Although students achieved higher results for appropriate testing after using VisiStat, they attributed this knowledge mainly to the lecture in the qualitative feedback. Furthermore, we found out that students would like to interactively choose the appropriate the appropriate test themselves to be able to try and error and spend more time on this subject in the lecture. Regarding checking assumptions, VisiStat's visibility of the assumptions supported to easily remember them and form first hypotheses about the connection between fulfilled assumptions and test choice. A deeper explanation of this connection was finally given by the lecture. The interviews revealed that the comprehension of over-testing is comparably low. A reason for this low understanding could be that especially the students in the second group did not encounter the over-testing warning in VisiStat and therefore could not develop deeper knowledge in VisiStat. The findings indicated that VisiStat is crucial for the sufficient reporting of results. In this case, students in the traditional learning group scored higher. As VisiStat was attributed to develop practical knowledge to create reporting texts, the lecture provided information why sufficient reporting is important, why could be an explanation for the better results of the traditional tell-and-practice group.

VisiStat and lecture complement each other to fulfill learning principles

Moreover, we discussed that a combination of VisiStat and lecture fulfills Garfield and Ben-Zvi's learning principles [2007]. Our findings suggest that VisiStat and lecture complement each other and can address all learning principles. However, the lecture in this user study did not actively involve students but this could be generally possible in lecture and was also demanded by students. Nonetheless, there is still room for improvement to support the learning principles. These suggestions for improvements mainly address higher interactivity when using VisiStat so that stu-

dents have more possibilities to develop hypotheses. Regarding the lecture, the main criticism is the amount of content for the time so that we propose to split the lecture into two parts with more practice. In conclusion, the results indicate that the role of VisiStat is mainly to practice statistics in an hands-on experiment and to construct own knowledge by developing assumptions although students are not always aware of this strengths. The lecture flexibly addresses students' open questions and discovers misconceptions by providing deeper knowledge. As a consequence, we recommend to use VisiStat for exploration before attending a lecture for future statistics education, but also include participants' feedback. The next chapter formulates suggestions for this inclusion.

5.2 Future work

Iterative improvement process

The qualitative interviews revealed students' request to be able to interactively change inputs in VisiStat as well as the possibility to try and error. For example, students want to choose the statistical test by themselves and then get detailed feedback why their choice was correct or not. To address these suggestions, we advise to follow an iterative improvement process, which evaluates how much change is necessary to improve VisiStat. In a first step, students could explore VisiStat for a while and then are asked to make predictions about for example the appropriate test for a situation. After they have written down the answer, they can contrast their own solution with the one in VisiStat. Therefore, it is important that the help description is extended and easier language is used. In a second step, the lecturer can evaluate students' responses in the task sheet and address possible misconceptions in class. In this case, students results should be checked with a post test again. If this approach is not sufficient, further adjustments could be considered, which include direct changes in VisiStat. By evaluating this process of improvements, a balance between VisiStat's former aim to support researchers to conduct appropriate statistics and VisiStat as a learning tool can be determined.

Furthermore, this thesis only provided a first overview of results and future works can address individual parts in more detail. For example, an in-depth analysis of each student, his or her tests results, and corresponding feedback can be conducted to conclude specific problems. Furthermore, students' utterances during the observation of VisiStat should be examined to find out whether differences between the two treatment groups exist and if the test results depend on the amount and quality of utterances. This analysis could also yield students' misconceptions when dealing with statistical concepts (in VisiStat), which could help to improve VisiStat and provide a basis how to difficult topics can be addressed in statistical education in HCI research.

In addition, future works evaluating VisiStat could assess the learners' attitude towards statistics, its impact on students' achievements and if it changes after using VisiStat. It could also be interesting to investigate student groups of other fields, which need comparable statistical analysis (e.g. psychologist or sociologists), and compare their results and feedback with the current findings, as statistical education is not only a problem of HCI researchers (cf. Chapter 2.2.

The previous part focused on incentives for future evaluations. This chapter now closes with suggestions for the reporting function in VisiStat, which could be extended. Apart from the use of a short explanation of the meaning of the reporting function (cf. Chapter 4.3.2, more text patterns could be provided, which are used alternately, so that all reports in the history can be copied in the result section without necessary changes. Otherwise, a longer result section only consists of the same text, which is not boring to read and unprofessional. Furthermore, to enhance the comprehensibility, tables instead of the second descriptive result sentence could be used.

Appendix A

User study

This appendix includes

- the statistical knowledge test (in the pre-test form, the other two are homogeneous to this one)
- the tasks users dealt with when exploring VisiStat
- the feedback questionnaire
- the interview questions

All other data can be found on the DVD.

Pre-Test

Dear student,

Thank you for participating in our user study. We appreciate your time and hope that you will benefit from our study as well. If you do have any questions, do not hesitate to write a mail to: sarah.voelkel@rwth-aachen.de

This study is anonymous. Complete participation in this user study results in 3% of the overall course score. As an alternative to the user study, you opt-in for an optional statistics assignment to receive the same score as well. Your performance in the study will not influence your final grade.

In this first step we would like you to fill out the following demographic questionnaire and pre-test so that we know your level of previous statistical knowledge. If you do not know the solution to a question, this is absolutely no problem: Just skip the question or write/mark "I don't know".

Please do not look up any of the questions on the web or in a book, etc., because we want to know the knowledge you have until now and are not testing you. Of course, if you do not understand an English word, you can look this up ;-)

Filling out the questionnaire takes approx. 15-30 minutes.

Thank you!
Sarah & Chat

* Required

1. **Your ID** *
   In order to be able to match your answers correctly while guaranteeing anonymity, we would ask you to think of a personal code or ID like your mother's name and your house number. For example: If your mother's name is Tracy and your house number is 42, your personal ID would be Tracy42.

Demographic Questionnaire

First of all, we want to know a little bit about you, like your demographic background and statistics experience.

2. **How old are you?** *

3. **Please indicate your gender.** *
   Mark only one oval.
   - ◯ female
   - ◯ male
   - ◯ other

4. **What are you studying?** *
Check all that apply.

- [ ] B. Sc. Computer Science
- [ ] M. Sc. Computer Science
- [ ] M. Sc. Media Informatics
- [ ] M. Sc. Software System Engineering
- [ ] M. Sc. Technical Communication
- [ ] Other: .....

5. **What is your current semester of studying?**
Please note that counting is resetted in a new course of study, so if you're in your second Master's semester, please write 2 (instead of e.g. 8).

.....

6. **How would you estimate your prior statistical knowledge?**
Mark only one oval.

	1	2	3	4	5	
very low	◯	◯	◯	◯	◯	very high

7. **How have you already developed statistical knowledge?**
Multiple answers are possible.
Check all that apply.

- [ ] school
- [ ] one university lecture
- [ ] more than one university lecture
- [ ] read books about statistics
- [ ] used statistics in a seminar work, thesis, paper, etc.
- [ ] interactive statistics learning systems
- [ ] Other: .....

8. **Your preference of learning**
Think back to your last learning situation, e.g. your last exam. How did you prepare for it? Did you do the exercises first and then tried to understand the theory or did you work through the theory first and then practiced your knowledge by doing exercises?
Mark only one oval.

- ◯ first practical application, then learning theory
- ◯ first understanding theory, then practicing
- ◯ Other: .....

Pre-Test

Now we're starting with the actual pre-test. If you do not know an answer, just skip the question or write/mark "I don't know". Please do not guess.

9. **What's the X in t(X)?**
   You conducted a t-Test. Your result is t(22) = -1.68. Name or shortly describe what 22 refer to?

   .....

   .....

   .....

   .....

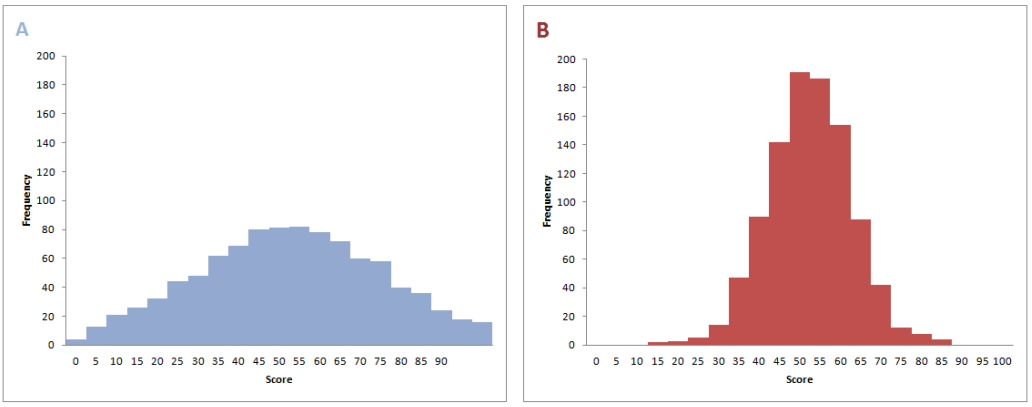
   .....

10. **What's X and Y in F(X,Y)?**
    In a statistical report, you found "F(2,12) = ...". Which of the following phrases are conclusions you can draw from this result?
    Check all that apply.

    - [ ] The number of users participated
    - [ ] The number of measured data points
    - [ ] The number of independent variables
    - [ ] The number of conditions of the independent variable
    - [ ] The name of statistical test
    - [ ] The cut-off level to determine whether the statistics is significant or not
    - [ ] I don't know

11. **Standard Deviation**
    Assign the correct standard deviation distribution to the pictures below assuming that both have the same mean
    Mark only one oval.

    - ◯ Standard Deviation of A > B
    - ◯ Standard Deviation of A < B
    - ◯ Standard Deviation of A = B
    - ◯ I don't know



12. **p-Value**
You want to test the hypothesis that students drink a different amount of beer than working people. You conduct a significance test comparing the amount of beer that each group drinks, which results in a p-value = 0.02. What can you conclude from this result?
Mark only one oval.

- ◯ There is a 2% chance that students and workers drink the same amount.
- ◯ There is a 2% chance that students and workers drink a different amount.
- ◯ If students and workers drink the same amount, there is a 2% chance that this result occurs.
- ◯ If students and workers drink the different amount, there is a 2% chance that this result occurs.
- ◯ I don't know

13. **What can you conclude from this result?**
You asked 10,000 people about their opinion towards the convenience of touchscreen mobile phones. You found out that there are statistically significant differences between men and women. The effect size is Cohen's d = 0.05. What can you conclude from this result?
Mark only one oval.

- ◯ small effect size: the survey needs more participants
- ◯ small effect size: the survey used too many participants
- ◯ large effect size: the survey needs more participants
- ◯ large effect size: the survey used too many participants
- ◯ I don't know

14. **Assumptions for parametric significance tests**
Please name the three assumptions your data has to fulfill in order to conduct parametric significance tests (e.g. t-test or ANOVA).

.....

.....

.....

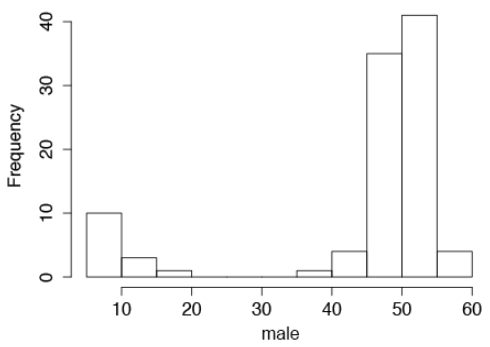
.....

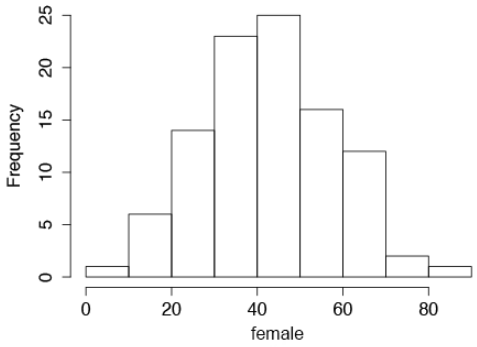
.....

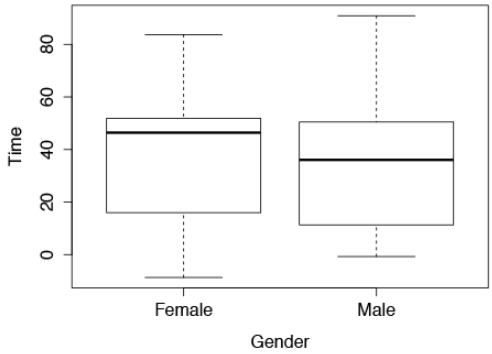
15. **Assumptions for unpaired parametric tests**
When you are carrying out an examination with an unpaired parametric test, is it important that your data is normally distributed...
Mark only one oval.

- ◯ overall
- ◯ within each group
- ◯ overall and within each group
- ◯ overall or within each group
- ◯ I don't know

Use the following graphs for the next question.







16. **Are the assumptions for unpaired t-test fulfilled?**
You are asked to evaluate an augmented reality app which lets the user of an online shopping portal try glasses on their own face. You want to conduct a t-test in order to find out whether there are differences between male and female and their task completion time with the app. The graphs above show the distribution of your data and the variances. Are the assumptions for an unpaired t-test fulfilled? If not, name the assumption which is violated.

.....

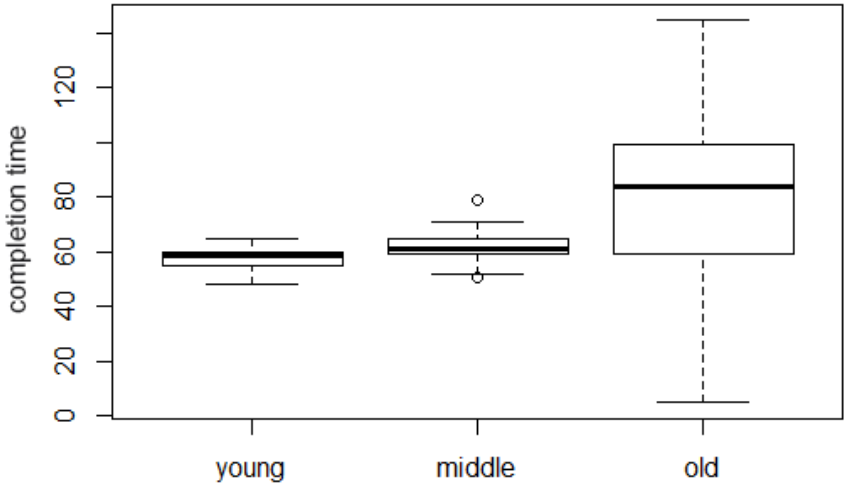
.....

.....

.....

.....

Use the following boxplot for the next two questions.



17. **Are the assumptions for one-way ANOVA fulfilled?**
Now you want to know whether there are differences of speed of navigating through menu between age groups (young: 0-25; middle: 26-49; old: 50-99). You have gathered data from 15 participants. The boxplot above shows the means and standard deviations. Are the assumptions for one-way ANOVA fulfilled? If not, name the assumption which is violated.

18. **Can you conduct a one-way ANOVA?**
Keep the scenario from the question before in mind. Assume that each of the three age groups consists of 50 participants. Can you conduct the one-way ANOVA? Give a short explanation.

19. **Characteristics of one-way ANOVA**
Please mark the correct characteristics of one-way ANOVA.
Check all that apply.

- [ ] within-groups design
- [ ] between-groups design
- [ ] one independent variable with two conditions
- [ ] one independent variable with more than two conditions
- [ ] two independent variables
- [ ] parametric test
- [ ] non-parametric test
- [ ] I don't know

20. **Characteristics of Paired t-Test**
Please mark the correct characteristics of paired t-test.
Check all that apply.

- [ ] within-groups design
- [ ] between-groups design
- [ ] one independent variable with two conditions
- [ ] one independent variable with more than two conditions
- [ ] two independent variables
- [ ] parametric test
- [ ] non-parametric test
- [ ] I don't know

21. **Unpaired vs. paired significance tests**
When do you use a paired instead of an unpaired significance test?

.....

.....

.....

.....

.....

22. **Parametric vs. non-parametric tests**
Describe one characteristic of the experiment that requires using a non-parametric instead of a parametric significance test.

.....

.....

.....

.....

.....

23. **One-Way vs. Two-Way ANOVA**
Under which conditions do you use a two-way ANOVA instead of a one-way ANOVA?
Mark only one oval.

- ◯ one variable with three conditions
- ◯ two variables each has one condition
- ◯ two variables each has two conditions
- ◯ I don't know

24. **Which statistical test do you use?**
You want to investigate three different objects (mouse, trackpad, joystick) as input devices. Therefore you ask every participant to use each object and note their task completion time with each. Your data is normally distributed and the variances are homogeneous. Which among the following tests is the most appropriate to compare the task completion time of the different objects?
Mark only one oval.

- ◯ Paired t-test
- ◯ Unpaired t-test
- ◯ One-way ANOVA
- ◯ One-way repeated-measured ANOVA
- ◯ Two-way ANOVA
- ◯ Non-parametric test
- ◯ I don't know

25. **Which statistical test do you use?**
You investigate interactive tabletops and have the following variables: Independent variable: gender (male/female); dependent variable: error rate (in percent). Your data is normally distributed and the variances are homogeneous. You want to find out whether there are significant differences between male and female error rate. Which is the most appropriate test?
Mark only one oval.

- Paired t-test
- Unpaired t-test
- One-way ANOVA
- One-way repeated-measured ANOVA
- Two-way ANOVA
- Non-parametric test
- I don't know

26. **Which statistical test do you use?**
You developed a mobile guide and want to find out how satisfied users are with it. Therefore, you split your participants into three groups (very experienced user, slightly experienced user, inexperienced user) and rate their satisfaction (from 0 = "not satisfied at all" to 5 = "very satisfied"). The variances are homogenous but your data is NOT normally distributed. Interested in whether there are differences between the three user groups, which test would you suggest as the most appropriate?
Mark only one oval.

- Paired t-test
- Unpaired t-test
- One-way ANOVA
- One-way repeated-measured ANOVA
- Two-way ANOVA
- Non-parametric test
- I don't know

27. **Which statistical test do you use?**
You developed an IDE. Now you want to know if your system is better than another already existing system. Therefore, you ask your participants to use both systems and measure their time to deal with a specific task. Your data is normally distributed and the variances are homogeneous. Which among the following tests is the most appropriate in order to find out whether there are significant differences between the two systems?
Mark only one oval.

- Paired t-test
- Unpaired t-test
- One-way ANOVA
- One-way repeated-measured ANOVA
- Two-way ANOVA
- Non-parametric test
- I don't know

28. **Which statistical test do you use?**
You developed a serious game for seniors which should help them to stay physically active. In order to investigate the effects of your game, you perform a longitudinal study. You measured your participants' accuracy of performance six months ago, three months ago and last week. Now you want to find out whether there is a trend of improvement. Your data is normally distributed and the variances are homogeneous. Which among the following tests is the most appropriate?
Mark only one oval.

- ◯ Paired t-test
- ◯ Unpaired t-test
- ◯ One-way ANOVA
- ◯ One-way repeated-measured ANOVA
- ◯ Two-way ANOVA
- ◯ Non-parametric test
- ◯ I don't know

29. **Which statistical test do you use?**
You want to investigate the time per day that people use social networks sites. You want to compare whether there are differences between men and women as well as between different age groups (3 groups: young, middle-aged, old). Your data is normally distributed and the variances are homogeneous. Which test would you suggest as the most appropriate?
Mark only one oval.

- ◯ Paired t-test
- ◯ Unpaired t-test
- ◯ One-way ANOVA
- ◯ One-way repeated-measured ANOVA
- ◯ Two-way ANOVA
- ◯ Non-parametric test
- ◯ I don't know

30. **Meaning of ANOVA results**
Assuming that you receive F and p value as a result from ANOVA, what of the following can you conclude?
Check all that apply.

- ☐ Identify whether or not there differences across groups
- ☐ Identify the group(s) that differ(s) from others
- ☐ Identify the directions of differences (more than, less than)
- ☐ Identify magnitude of the differences
- ☐ I don't know

31. **Meaning of post-hoc test**
Assuming that your ANOVA showed a significant result and now you perform pairwise post-hoc tets, which of the following conclusions can you ALWAYS make from pairwise post-hoc tests?
Check all that apply.

- [ ] Identify whether or not there are differences across groups
- [ ] Identify the group(s) that differ(s) from others
- [ ] Identify the directions of differences (more than, less than)
- [ ] Identify magnitude of the differences
- [ ] I don't know

32. **What is the risk of using a 10-way ANOVA?**

.....

.....

.....

.....

.....

33. **ANOVA and pairwise t-Tests**
You want to compare three different aging groups: young, middle-aged and old. Which test is the most appropriate?
Mark only one oval.

- ◯ Pairwise t-tests (young vs. middle-aged; middle-aged vs. older; young vs. older)
- ◯ Pairwise t-tests with Bonferroni correction
- ◯ 1-way ANOVA with all three conditions
- ◯ I don't know

34. **Reporting 2-way ANOVA**
You have to report a 2-way ANVOA result with the independent variables gender and technically experienced (3 groups: low, middle, high) and the dependent variable performance time. How many F-values do you have to report? Name the effect for each of the F-values.

.....

.....

.....

.....

.....

35. **Reporting Unpaired t-Test**
You conducted an unpaired t-test in order to find out whether there are significant differences between participants who used two different keyboard layout (k1 and k2) on their completion time in a typing test. As you are writing a paper for a conference, you do not have much space. Which of the following values would you include in order to satisfy the minimum requirements assuming that the result is significant?
Check all that apply.

- [ ] exact p-value
- [ ] $p < 0.05$
- [ ] t-value
- [ ] n (n = number of participants)
- [ ] n_k1 (n = number of participants)
- [ ] n_k2 (n = number of participants)
- [ ] degrees of freedom
- [ ] SE_k1
- [ ] SE_k2
- [ ] M
- [ ] M_k1
- [ ] M_k2
- [ ] independent variable
- [ ] dependent variable
- [ ] SD
- [ ] SD_k1
- [ ] SD_k2
- [ ] Var
- [ ] Var(k1)
- [ ] Var(k2)
- [ ] Median
- [ ] Median_k1
- [ ] Median_k2
- [ ] effect size (r^2 or Cohen's d)
- [ ] confidence interval
- [ ] I don't know

36. **Reporting Paired t-Test**

Assume now you would have asked the same participants to perform the typing test between two keyboard layouts. In this case, you would have to report a paired t-test in order to compare their completion times. Which of the following values would you NOT include unter these conditions in order to satisfy the minimum requirements assuming that the result is significant.

Check all that apply.

- [ ] exact p-value
- [ ] $p < 0.05$
- [ ] t-value
- [ ] n (n = number of participants)
- [ ] n_k1 (n = number of participants)
- [ ] n_k2 (n = number of participants)
- [ ] degrees of freedom
- [ ] SE_k1
- [ ] SE_k2
- [ ] M
- [ ] M_k1
- [ ] M_k2
- [ ] independent variable
- [ ] dependent variable
- [ ] SD
- [ ] SD_k1
- [ ] SD_k2
- [ ] Var
- [ ] Var(k1)
- [ ] Var(k2)
- [ ] Median
- [ ] Median_k1
- [ ] Median_k2
- [ ] effect size (r^2 or Cohen's d)
- [ ] confidence interval
- [ ] I don't know

37. **Reporting significance test**
You want to write a paper about the differences between experts, regular users and novice users and their correctness of performance with your system. You have determined the results below and checked that the data fulfills the assumptions of parametric significance tests. Write down how you would report these results and do not forget to mention the test you applied.

.....

.....

.....

.....

.....

One-Way ANOVA
$F(2,72) = 12.43$
$\eta^2 = 0.41$
$p = 0.0089$

	All	Expert	Middle	Novice
n	75	25	25	25
M	33.091	39.012	32.456	27.713
SD	5.062	3.125	6.812	5.249
Confidence intervals	[23.169, 43.013]	[32.887, 45.137]	[19.104, 45.808]	[17.425, 38.001]

Team:

Tasks

Dataset: Keyboard Layouts Comparison

1. Find out whether there is a significant influence of **gender** on **speed**.

2. Investigate whether there are differences between the three **keyboard layouts** QWERTY, Dvorak and Colemak and the participants' **speed**.

3. You want to write a thesis with your results from the two tasks above. Were the differences significant? Discuss the implications with your partner.

Task	**Significant**	
Task 1: gender => speed	**O** Yes	**O** No
Task 2: keyboard layout => speed	**O** Yes	**O** No

Dataset: Effect of Food on Test Scores

1. Does the **eaten food** influence the **verbal score**?

2. Now find out whether the **verbal score** depends on **gender**.

3. You want to publish both your results from the two tasks before in a paper. Which of the two tasks reported a significant result? How large were the effect sizes? Discuss the implications with your partner.

Task	**Significant**		**effect size**
Task 1: gender => speed	**O** Yes	**O** No	_____
Task 2: keyboard layout => speed	**O** Yes	**O** No	_____

4. Does **gender** also have a significant influence on the **math score**?

Dataset: Weight Lost

1. Do the three different **conditions** influence the amount of **weight loss**?

2. To be able to write a paper, report the results. Which effect size has been measured? Discuss the meaning of this result and each reported statistics.

   Effect size: _____

Dataset: Effect of OS on Stress

1. Are there differences between the **phone OSs** and their resulting **stress Score**?

2. If the difference in task 1 is significant, speculate which OS causes the significance and compare it with each of the other OSs.

3. You want to publish your results from the two tasks above in your paper. Discuss the statistical procedure.

Dataset: Weight Lost

1. Do the three different **conditions** influence the amount of **weight loss**? Notice the test that is used.

2. Have the different **conditions** an influence on **BMI**? Notice the test that is used.

3. Investigate the effect of **condition** on **user rating**. Notice the test that is used.

4. Furthermore, find out how **condition** and **exercise** together influence the **weight loss**.

5. To be able to write a paper, report the results from task 3. Discuss the results and focus on which effects are compared.

Feedback Questionnaire VisiStat

Dear participant,

Thank you for staying with us till this point of time :) This is finally the last written questionnaire we kindly ask you to fill out.

Therefore please answer the following questions providing feedback how you evaluate the learning experience with VisiStat. There are no right or wrong answer but we value your honest and personal opinion. Of course, your answers will be treated completely anonymously.

Filling out the questionnaire will take approx. 5 minutes.

Thank you!
Chat
& Sarah

* Required

1. **Your ID** *
   Please fill in your personal ID. Remember this was the ID you thought of at the first questionnaire. Please make sure this is the same ID you used in the other questionnaires so that we will be able to match them correctly. For example: Your mother's name and the house number you used to live in as a child.

   .....

Usefulness of VisiStat to learn statistics

You have used VisiStat before/after the statistics lecture. Do you think that VisiStat is useful in this situation?

2. **I believe that using VisiStat would improve my course performance.** *
   Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

3. **I find VisiStat useful in my lecture.** *
   Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

4. **Using VisiStat would enhance my effectiveness in learning statistics.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

5. **Using VisiStat would increase my productivity in dealing with statistics.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

6. **Using VisiStat would make it easier to understand concepts in statistics.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

7. **Using VisiStat would make it easier to use statistics.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

8. **I think that VisiStat should be part of each course concerning learning statistics in the university.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

Usability of VisiStat

The second part of this questionnaire deals with your perceived ease of use of VisiStat as a learning device for statistics.

9. **Learning to operate VisiStat was easy for me.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

10. **My interaction with VisiStat was clear and understandable.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

11. **It would be easy for me to become skillful at using VisiStat.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

12. **I found VisiStat easy to use.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

Enjoyment while using VisiStat

Apart from your learning results we would like to know whether you enjoyed using VisiStat.

13. **I had fun interacting with VisiStat.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

14. **Using VisiStat was pleasant.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

15. **I found using VisiStat enjoyable.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

16. **Using VisiStat bored me.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

Working with VisiStat

In the following part we would like you to tell us about how working with VisiStat felt for you.

17. **Sometimes I lost track of time when I was using VisiStat.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

18. **Time flew while I was using VisiStat.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

19. **When I was using VisiStat, I am able to block out most other distractions.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

20. **While using VisiStat, I was absorbed in what I am doing.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

Overall evaluation of Lecture

Now we would like to know how you evaluate the learning experience in the lecture.

21. **Attending the lecture revealed my misunderstandings in statistical concepts.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

22. **The lecture helped me understand concepts in statistics that I am familiar with.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

23. **The lecture introduced me to unfamiliar concepts in statistics.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

24. **How would you rate the lecture in order to learn more about statistics in HCI?** *
Mark only one oval.

	1	2	3	4	5	6	7	
very bad	◯	◯	◯	◯	◯	◯	◯	very good

Overall evaluation VisiStat

Furthermore, we would like to know how you evaluate the learning experience with VisiStat and whether you would intend to use it in the future.

25. **Interacting with VisiStat revealed my misunderstandings in statistical concepts.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

26. **VisiStat helped me understand concepts in statistics that I am familiar with.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

27. **VisiStat introduced me to unfamiliar concepts in statistics.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

28. **How would you rate the learning experience with VisiStat in order to learn more about statistics in HCI?** *
Mark only one oval.

	1	2	3	4	5	6	7	
very bad	◯	◯	◯	◯	◯	◯	◯	very good

29. **I intend to use VisiStat in order to prepare for future exercises, exams, thesis, etc.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

30. **I intend to recommend VisiStat to my friends when they have to use statistics.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

Overall evaluation VisiStat and Lecture

Finally, we would like to know how you evaluate the whole learning experience with VisiStat and the lecture.

31. **VisiStat and Lecture complement one another.** *
Mark only one oval.

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

32. **How would you rate the whole learning experience with VisiStat and lecture to learn more about statistics in HCI?** *
Mark only one oval.

	1	2	3	4	5	6	7	
very bad	◯	◯	◯	◯	◯	◯	◯	very good

Interview Questions

---------- part 1 ----------

1.) Could you tell me about your experience in the statistics lecture?

2.) Could you tell me about your experience in using the interactive system with a partner?

• In aspect of …, how do you compare the experience in the lecture and the experience with VisiStat?
• Could you describe a reason that … is better in {lecture, VisiStat}?
• What are advantages of VisiStat that the lecture did not offer?
• What are advantages of the lecture that VisiStat did not offer?

Follow–up for clarification in each condition:

• Can you give me an example of …? (e.g., concepts that they said they don't understand)
• When you said … did you mean …? (try to elicit precise responses)

3.) Do you think that the exploration of VisiStat last week influenced your learning in the lecture?

- Have there been parts of the lecture you could understand easier because you're already explored them in VisiStat?
- Have there been parts you wanted an explanation for while using VisiStat and got this explanation during the lecture?
- Have there been parts in the lecture which were completely new to you and you have not observed in VisiStat?

4.) Overall in the last 3 weeks that you participated in the user study, do you think that your statistical knowledge improved?

- Why/Why not?

5.) How would evaluate the overall learning experience of VisiStat and lecture?

---------- part 2 ----------

We're now going to talk about some specific parts we focus on in learning statistics.

1.) Please read the following research question. (Appropriate testing)

- Do you know how to choose statistical test?
- Where do you know it from?
- Follow--up: Which part of the lecture or VisiStat did you learn this from?

2.) Please read the following research question. (Assumptions)

- Do you know which assumptions have to be checked before conducting a test?
- Where do you know it from?
- Follow--up: Which part of the lecture or VisiStat did you learn this from?

3.) Do you think you know the risk of over-testing? Please read the following research question. (Over-testing)

- Do you know the risk of overtesting?
- Where do you know it from?
- Follow--up: Which part of the lecture or VisiStat did you learn this from?

4.) Please read the following research question. (Reporting)

- **Do you know the standard of reporting your results?**

- **Where do you know it from?**
- Follow--up: **Which part of the lecture or VisiStat did you learn this from?**

---------- part 3 ----------
Imagine now that you're a teacher for statistics.
1.) Would you like to make any improvements to VisiStat?
- would you add any more functionalities?
- would you remove something from VisiStat?
2.) Would like to make any improvements to the lecture?

---------- part 4 ----------
Is there anything else you would like to tell me?

Bibliography

Christopher L Aberson, Dale E Berger, Michael R Healy, Diana J Kyle, and Victoria L Romero. Evaluation of an interactive tutorial for teaching the central limit theorem. *Teaching of Psychology*, 27(4):289–291, 2000.

Ritu Agarwal and Elena Karahanna. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, pages 665–694, 2000.

Aly Amer. Reflections on Bloom's revised taxonomy. *Electronic Journal of Research in Educational Psychology*, 4(1): 213–230, 2006.

Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, New York, 2001.

APA. *Publication Manual of the American Psychological Association*. American Psychological Association, Washington DC, 6th edition, 2010.

Carmelo Ardito, Maria Francesca Costabile, Marilena De Marsico, Rosa Lanzilotti, Stefano Levialdi, Teresa Roselli, and Veronica Rossano. An approach to usability evaluation of e-learning applications. *Universal access in the information society*, 4(3):270–283, 2006.

Melanie Autin, Hope Marchionda, and Summer Bateiha. Attitude adjustment in introductory statistics. In *Proceedings of the 41th Annual Meeting of the Research Council on Mathematics Learning 2014*, page 80, 2014.

S-P Ballstaedt. Textoptimierung: Von der Stilfibel zum Textdesign. *Fachsprache*, 21(3-4):98–124, 1999.

Rachel M Best, Michael Rowe, Yasuhiro Ozuru, and Danielle S McNamara. Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders*, 25(1):65–83, 2005.

Benjamin Samuel Bloom and David R Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans, 1956.

John D Bransford and Daniel L Schwartz. Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, pages 61–100, 1999.

Franz Breuer. *Reflexive Grounded Theory: Eine Einführung für die Forschungspraxis*. Springer, 2009.

Luc Budé, Margaretha WJ Van De Wiel, Tjaart Imbos, MJJM Candel, Nick J Broers, and Martijn PF Berger. Students' achievements in a statistics course in relation to motivational aspects and study behaviour. *Statistics Education Research Journal*, 6(1):5–21, 2007.

Paul Cairns. Hci... not as it should be: inferential statistics in hci research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, pages 195–201. British Computer Society, 2007.

J Clark, Gertrud Karuat, David Mathews, and Joseph Wimbish. The fundamental theorem of statistics: Classifying student understanding of basic statistical concepts. *Unpublished paper*, 2003.

Ruth C Clark and Richard E Mayer. *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. John Wiley & Sons, 2011.

Jacob Cohen. The earth is round ($p < .05$). *American psychologist*, 49(12):997, 1994.

Deborah Cotton and Karen Gresty. Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology*, 37(1):45–54, 2006.

Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

Fred D Davis Jr. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD thesis, Massachusetts Institute of Technology, 1986.

Robert Delmas, Joan Garfield, Ann Ooms, and Beth Chance. Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2):28–58, 2007.

Pierre Dragicevic, Fanny Chevalier, and Stéphane Huot. Running an hci experiment in multiple parallel universes. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 607–618. ACM, 2014.

Mark D Dunlop and Mark Baillie. Paper rejected ($p > 0.05$): An introduction to the debate on appropriateness of null-hypothesis testing. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 1(3):86–93, 2009.

Taciana Pontual Falcão and Sara Price. What have you done! the role of 'interference' in tangible environments for supporting collaborative learning. In *Proceedings of the 9th international conference on Computer supported collaborative learning-Volume 1*, pages 325–334. International Society of the Learning Sciences, 2009.

Andy Field. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

Sara J Finney and Gregory Schraw. Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, 28(2):161–186, 2003.

Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

Norm Friesen. *Re-thinking e-learning research: Foundations, methods, and practices*, volume 333. Peter Lang, 2009.

Iddo Gal and Lynda Ginsburg. The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, 2(2):1–15, 1994.

Joan Garfield. How students learn statistics. *International Statistical Review/Revue Internationale de Statistique*, pages 25–34, 1995.

Joan Garfield. The statistical reasoning assessment: Development and validation of a research tool. In *In the Proceedings of the 5 th International Conference on Teaching Statistics*. Citeseer, 1998.

Joan Garfield and Andrew Ahlgren. Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for research in Mathematics Education*, pages 44–63, 1988.

Joan Garfield and Dani Ben-Zvi. How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75 (3):372–396, 2007.

Joan Garfield, M Aliaga, G Cobb, C Cuff, Rob Gould, Robin Lock, Tom Moore, Allan Rossman, Bob Stephenson, Jessica Utts, et al. Guidelines for assessment and instruction in statistics education (gaise): College report. *Alexandria, Virginia: The American Statistical Association*, 2005.

Joan Garfield, Robert delMas, and Beth Chance. Using students' informal notions of variability to develop an understanding of formal measures of variability. *Thinking with data*, pages 117–147, 2007.

Eleanor Jack Gibson. *Principles of perceptual learning and development*. Appleton-Century-Crofts, 1969.

Barney G Glaser and Anselm L Strauss. *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers, 2009.

Frederick Gravetter and Lori-Ann Forzano. *Research methods for the behavioral sciences*. Cengage Learning, 2011.

Wayne D Gray and Marilyn C Salzman. Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human–Computer Interaction*, 13 (3):203–261, 1998.

Norbert Groeben and Peter Vorderer. *Leserpsychologie: Textverständnis-Textverständlichkeit*. Aschendorff Münster, 1982.

Jens Grossklags and Nathan Good. Empirical studies on software notices to inform policy makers and usability designers. In *Financial Cryptography and Data Security*, pages 341–355. Springer, 2007.

Heiko Haller and Stefan Krauss. Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, 7(1):1–20, 2002.

Eva-Maria Jakobs. Kommunikative Usability. *Sprache und Kommunikation im technischen Zeitalter: Wieviel Internet (v) erträgt unsere Gesellschaft?*, 2:119, 2012.

Eva-Maria Jakobs and Katrin Lehnen. Hypertext - Klassifikation und Evaluation. In *Websprache. net: Sprache und Kommunikation im Internet*. Walter de Gruyter, Berlin, 2005.

Douglas H Johnson. The insignificance of statistical significance testing. *The journal of wildlife management*, pages 763–772, 1999.

Maurits Kaptein and Judy Robertson. Rethinking statistical analysis methods for chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1105–1114. ACM, 2012.

Carolyn M Keeler and R Kirk Steinhorst. Using small groups to promote active learning in the introductory statistics course: A report from the field. *Journal of Statistics Education*, 3(2):1–8, 1995.

Clifford Konold. Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education*, 3(1):1–9, 1995.

R. Koper. *Handbook of research on educational communications and technology*, chapter Open source and open standards, pages 355–365. Spector, J.M., Merrill, M.D., Jeroen van Merrienboer, J.V., Driscoll, M.P., Mahwah, NJ, 2007.

David R Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

David M Lane and Zhihua Tang. Effectiveness of simulation training on transfer of statistical concepts. *Journal of Educational Computing Research*, 22(4):383–396, 2000.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. John Wiley & Sons, 2010.

A Leavy, O Fitzmaurice, and A Hannigan. If you're doubting yourself then, what's the fun in that? an exploration of why prospective secondary mathematics teachers perceive statistics as difficult. *Journal of Statistics Education*, 21(3), 2013.

MC Lovett. A collaborative convergence on studying reasoning processes: A case study in statistics. *Cognition and instruction: Twenty-five years of progress*, pages 347–384, 2001.

Kori Lloyd Hugh Maxwell. *The Effects Of Using Visual Statistics Software On Undergraduate Students' Achievement In Statistics And The Role Of Cognitive And Non-Cognitive Factors In Their Achievement*. PhD thesis, Georgia State University, 2014.

Danielle S McNamara, Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch. Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1):1–43, 1996.

Thomas L Naps, Guido Rößling, Vicki Almstrum, Wanda Dann, Rudolf Fleischer, Chris Hundhausen, Ari Korhonen, Lauri Malmi, Myles McNally, Susan Rodger, et al. Exploring the role of visualization and engagement in computer science education. In *ACM SIGCSE Bulletin*, volume 35, pages 131–152. ACM, 2002.

Jakob Nielsen. *Usability engineering*. Elsevier, 1994.

Donald A Norman. *The design of everyday things*. Basic books, 2002.

OED online. e-learning, n., 2014. http://www.oed.com/view/Entry/261522?redirectedFrom=e-learning, accessed on June 25, 2014.

A Pang. The educational effectiveness of dynamic and interactive data visualisation and exploration in geographical education. *An academic exercise in partial fulfilment of*

the requirements for the degree of Master of Science in Geographical Information Science at Birkbeck College, University of London, 2001.

Adam Perer and Ben Shneiderman. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 265–274. ACM, 2008.

Raafat Saadé and Bouchaib Bahli. The impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: an extension of the technology acceptance model. *Information & Management*, 42(2):317–327, 2005.

Raafat Saadé, Fassil Nebebe, and Weiwei Tan. Viability of the "technology acceptance model" in multimedia learning environments: a comparative study. *Interdisciplinary Journal of E-Learning and Learning Objects*, 3(1):175–184, 2007.

Barbara Sandig. *Formulieren und Textmuster Am Beispiel von Wissenschaftstexten*. Schreiben in den Wissenschaften, 1997.

Bertrand Schneider, Jenelle Wallace, Paulo Blikstein, and Roy Pea. Preparing for future learning with a tangible user interface: The case of neuroscience. *Learning Technologies, IEEE Transactions on*, 6(2):117–129, 2013.

Daniel L Schwartz and Taylor Martin. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2):129–184, 2004.

Daniel L Schwartz, Catherine C Chase, Marily A Oppezzo, and Doris B Chin. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, 103(4):759, 2011.

Dino Schweitzer and Wayne Brown. Interactive visualization for the active learning classroom. *ACM SIGCSE Bulletin*, 39(1):208–212, 2007.

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg R Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, et al. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*, 2009.

Ben Shneiderman and Catherine Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Addison Wesley, Boston, USA, 5th edition edition, 2010.

Krishna Subramanian. Visistat: Visualization-driven, interactive statistical analysis. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 987–992, New York, NY, USA, 2014a. ACM.

Krishna Subramanian. Visistat: Visualization-driven, interactive statistical analysis. Master's thesis, RWTH Aachen University, 2014b.

Pei-Chen Sun, Ray J. Tsai, Glenn Finger, Yueh-Yang Chen, and Dowming Yeh. What drives a successful e-learning? an empirical investigation of the critical factors influencing learner satisfaction. *Computers and Education*, 50(4): 1183 – 1202, 2008.

Thierry Volery and Deborah Lord. Critical success factors in online education. *International Journal of Educational Management*, 14(5):216–223, 2000.

Ruud Wetzels, Dora Matzke, Michael D Lee, Jeffrey N Rouder, Geoffrey J Iverson, and Eric-Jan Wagenmakers. Statistical evidence in experimental psychology an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3):291–298, 2011.

Mun Y Yi and Yujong Hwang. Predicting the use of web-based information systems: self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model. *International journal of human-computer studies*, 59 (4):431–449, 2003.

Panagiotis Zaharias and Angeliki Poylymenakou. Developing a usability evaluation method for e-learning applications: Beyond functional usability. *Intl. Journal of Human–Computer Interaction*, 25(1):75–98, 2009.

Weimo Zhu. Sadly, the earth is still round ($p < 0.05$). *Journal of Sport and Health Science*, 1(1):9–11, 2012.

Andrew Zieffler, Joan Garfield, Shirley Alt, Danielle Dupuis, Kristine Holleque, Beng Chang, et al. What does research suggest about the teaching and learning of introductory statistics at the college level? a review of the literature. *Journal of Statistics Education*, 16(2):1–23, 2008.

Alain F Zuur, Elena N Ieno, and Chris S Elphick. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1):3–14, 2010.

Index

Typeset August 15, 2014