FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Uncertainty Visualization

Two User Studies

PRAKTIKUM

im Rahmen des Studiums

Visual Computing

eingereicht von

Andreas Roschal, Fabian Schwarzinger
Matrikelnummer 1225600, 1225307

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Mag. DI Dr. Theresia Gschwandtner

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Uncertainty Visualization

Two User Studies

PRACTICAL

in

Visual Computing

by

Andreas Roschal, Fabian Schwarzinger
Registration Number 1225600, 1225307

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Mag. DI Dr. Theresia Gschwandtner

Abstract

TODO Abstract

Contents

CHAPTER 1

Introduction

Data sets containing information retrieved from the real world usually also contain some amount of uncertainty. This uncertainty is often inherent to the data, for instance because certain measurements can never be exact or because some kind of aggregation is already done when acquiring the data. This is also true for temporal data. Sometimes the exact time of an event is not known (e.g., 'time of the big bang'), is given in an inexact way (e.g., 'since a few hours') or is an imprecise prediction of the future (e.g., 'it will take one or two days'). To incorporate these uncertainties into visual representations and make them visible to the user, several approaches have been proposed [9, 4, 12, 3, 7].

To find out more about the strengths and weaknesses of these techniques and to find out which technique fits certain tasks best, several studies have been conducted. In 2009 Sanyal et al. [15] asked 27 participants to solve four tasks with the help of four commonly used uncertainty visualizations. In 2012 Corell and Gleicher [5] compared four visual encodings of statistical uncertainty in a user study. In 2012 MacEachren et al. [11] conducted two studies, which targeted the intuitiveness of visual encodings and their performance in map reading tasks respectively. In 2015 Gschwandtner et al. [6] compared six visual encodings in a comprehensive user study.

To build upon the results of those studies, we are conducting two additional user studies. The first one (referred to as *Drawing Study*) aims to find out more about the intuitiveness of visual encodings. By asking people to draw visualizations for given tasks, we find out how people think about given problems and what kind of representations they think are most appropriate in those situations. The second study (referred to as *Evaluation Study*) is very similar in its design to the one by Gschwandtner et al. [6]. The difference is, that we do not only compare different visual encodings of uncertainty, but also include a representation in the comparison that completely omits the uncertainty of the underlying data. Through this approach we find out in which situations the visualization of uncertainty adds helpful information and in which situations it is only a counterproductive distraction.

In this report we thoroughly describe the design, execution and results of our user studies. Furthermore, we present some of the most relevant related work that has been done, which can be found in Chapter 2. In Chapter 3 we explain the design of our studies. This chapter is split into three main parts - the first is about the *Evaluation Study*, the second part regards the *Drawing Study* and in the third part the chosen participants and the evaluation approach is addressed. In the following Chapter 4 the results are presented, which are discussed in detail in Chapter 5. We conclude this report with a summary of our approach and its most important findings in Chapter 6.

CHAPTER 2

Related Work

Since we are designing and conducting two user studies, other similar studies are of great interest to us. Through those existing works, we can learn more about the state of the art of study design and evaluation.

Obviously, the user study that guides our work the most is the one by Gschwandtner et al. [6], as our aim is to build up on this work. This study compares six different techniques for the visualization of uncertainty in the temporal domain. To determine which technique works best for certain tasks, five different types of tasks were designed. The first type is about finding out how users interpret the different visualization techniques. In the second type of tasks the users are asked to read the boundaries of uncertainty intervals from the visualization. The third type is about determining the extent of an uncertainty interval. In the fourth type of tasks, the users have to gauge certain probabilities using the visualization and the last type of tasks asks the users for their opinion about the visualization.

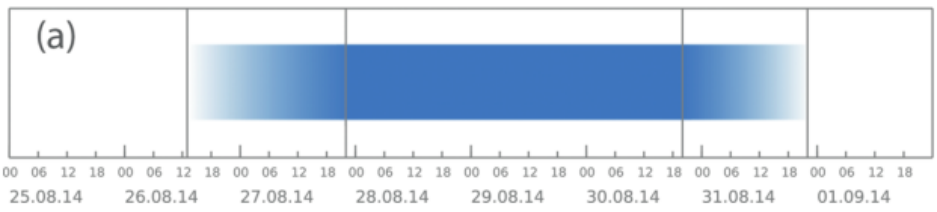


Figure 2.1: *A gradient plot shows the certain parts of an interval as a solid color, while the uncertain parts are represented by a color gradient. [6]*

The actual study was conducted with 73 participants, which all were bachelor students in computer science. The students were recruited from a course in information design and visualization, which implies a certain knowledge about this topic. To automatically track relevant data, such as completion time and accuracy, during the study sessions, the EvalBench software library

[2] was utilized. This library was designed especially for the evaluation of visualization. To analyze the results Gschwandtner et al. ran an analysis of variance(ANOVA) for each task and subtask and backed up their results with a non-parametric Kruskal-Wallis test. Their analysis showed, that the technique *ambiguation*, which can be seen in Figure 2.2, works best for tasks in which the user has to judge the exact duration and bounds of an uncertainty interval. If the user has to determine certain probabilities within the uncertainty interval, *gradient plots* (see Figure 2.1) work best.

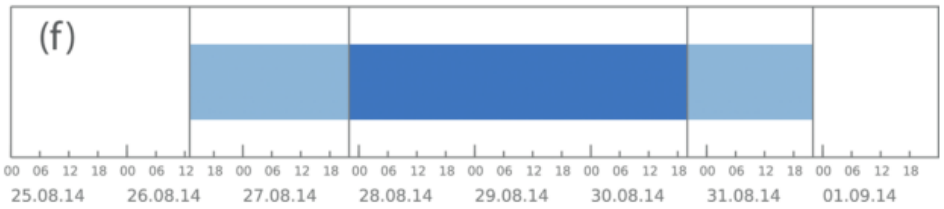


Figure 2.2: *This technique is called ambiguation and shows the uncertain parts of a time interval in a lighter color than the certain part, which is represented by a solid color. [6]*

In our *drawing study* our main focus lies on gauging how people think about certain situations and what kind of visualizations they associate with them. The goal is to find out what is intuitive for most people. These insights are valuable for the design of new visualizations, especially those aimed at non-expert users. MacEachren et al. [11] also tried to find out more about intuitive design of visualizations through user studies. Similar to us they conducted two separate user studies, which also have similar goals to ours. Their first study compares many different sets of symbols for the visualization of uncertainty, to find out which are most intuitive to people. Every set consists of three symbols, which encode high, medium and low uncertainty of 3 different kinds (accuracy, precision and trustworthiness), for 3 data domains (spatial, temporal and attribute). Some example sets can be seen in Figure 2.3. In total 102 symbol sets were rated by 31 undergraduate students for their intuitiveness on a scale from 1 to 7. After this first series of tests, the most unintuitive symbol sets were filtered out, which left 76 sets over. Those were again rated by 72 participants with a background in GIScience.

Figure 2.3: *Every column shows a set of three icons which represent high, medium or low uncertainty respectively. [11]*

After this first study about intuitiveness, the 20 highest rated symbols for every combination of uncertainty type and data domain were compared in a second experiment. The goal of this subsequent study is to compare the selected visualizations' performance, so the combined results

of both studies yield the best visualizations for a given task, which is intuitive and efficient at the same time. To compare the chosen symbols, two quadratic matrices with 9 symbols each were visualized side by side. The participants were asked to answer the question which of the two matrices featured a lower overall certainty, based on the presented symbols.

Walny et al. [16] conducted a study with the goal of providing deeper insights into the way people think about and use visualizations to communicate their ideas. This study is relevant to our work, because it features a similar approach as our *drawing study*. A total number of 69 researchers were observed using whiteboards during brainstorming, thinking, communication and similar actions. Whiteboards were chosen as a visualization medium, because they support a variety of thinking tasks, like personal and collaborative cognition, group meetings and planning. The results of the study feature interesting insights, such as different uses of emphasis techniques and the usage of ellipses as a focus and context technique. Our *drawing study* aims to provide similar insights through a similar approach, by also observing users in their creation of visualizations and reviewing those drawings.

In another study of greater exploratory nature, Walny et al. [17] asked 22 participants (mostly computer science students) to sketch visualizations of a given dataset. The data was provided in a table format and was about appropriateness ratings of certain behavior in given situations. The student's task was to create visualizations to find interesting patterns in the data and articulate them in a post-sketching questionnaire. The results were analyzed through multiple coding passes, which showed that, even though 9 out of 22 participants claimed to have no experience in visualization, most of the sketched representations could be classified as known types. As already stated, the study is of exploratory nature and therefore it does not answer many questions, but rather raises interesting questions and gives direction to future research.

There are user studies in the domain of information visualization which have the goal of determining if the visualization of a certain kind of information or a certain way of visualizing it is feasible or not. Our *Evaluation Study* is one of those studies, since it aims to answer the question if it is advisable to visualize temporal uncertainties or if this information is not used in decision making and should therefore be omitted. Another similar study by Xu et al. [18] is concerned with the feasibility of curved lines in graph layouts. In the study, graphs were visualized with either straight edges or three different kinds of curved edges and users were asked to perform certain tasks with a given graph. The completion time of those tasks and whether the final user decision was correct or incorrect, served as objective measures to rate the different graph layouts. Furthermore, the users were asked to give their personal opinion on which graphs they prefer and find visually more pleasing. An example graph in the three different layouts can be seen in Figure 2.4.

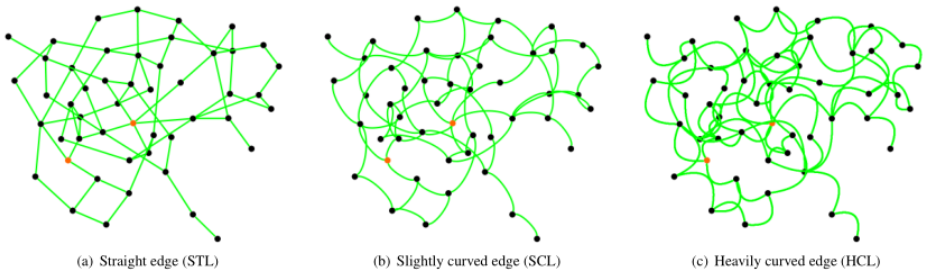


Figure 2.4: *The same graph is drawn in three different curve layouts. (a) shows the graph drawn with straight edges, while (b) and (c) use slightly and heavily curved edges respectively. [18]*

Robertson et al. [14] conducted a study to get insights about the feasibility of animations in trend analysis visualizations. To answer the question if animation helps in the comprehension of visualized trends and in the completion of corresponding tasks, users were provided with one dynamic and two static data representations and asked to perform tasks with the presented data. One of the static representations showed the change within the data in traces of the changing data points, as can be seen in Figure 2.5. The tasks were either questions regarding the presented data or some kind of analysis task. During every study session, the completion time and the accuracy of the given answers were automatically recorded. To evaluate the results of the study, four hypothesis were formulated and tested for support within the resulting data.

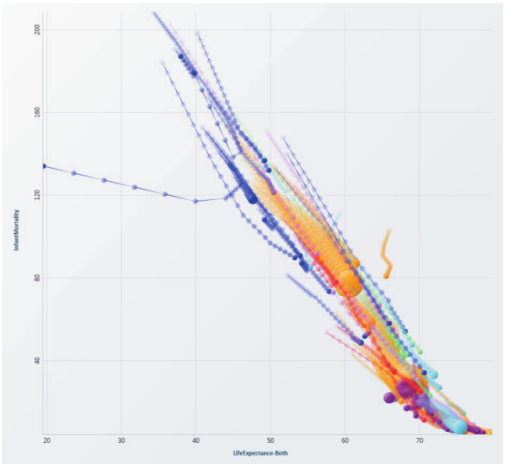


Figure 2.5: *The visualized data points are changing over time, which is statically represented by their traces. Traces are generated by drawing each point in its different stages over time and connecting these stages with lines. [14]*

CHAPTER 3

Method

In this chapter the study designs of both user studies are presented. Both studies have the aim of unraveling insights about the visualization of temporal uncertainty, but their approaches are wildly different. This is due to the varying research questions they try to answer. For this reason the two designs are presented in the following two separate sections.

3.1 Design of the *Evaluation Study*

Our *Evaluation Study* is building on the work of Gschwandtner et al. [6]. Gschwandtner et al. [6] compare different visual encodings for temporal uncertainties in order to find out which representations work best for which kinds of tasks. Our goal, on the other hand, is to evaluate for which kinds of tasks it actually makes sense to additionally visualize the information about the temporal uncertainty at all, i.e. in which cases does the user benefit from it and in which cases is the visualized uncertainty more of a distraction and actually not supporting the user in fulfilling his or her tasks. In order to examine those proposed deliberations, we designed our *Evaluation Study* and implemented it using the EvalBench framework [2]. Figure 3.1 shows an example of the user interface of our study implementation inside EvalBench.

Figure 3.1: *A screenshot of our study implementation using EvalBench. On the left side the task supporting visualization is depicted. On the right side the task description and inputs for answers are displayed.*

The study was conducted on three different user groups. We applied a within-subject-design on the tasks and a between-subject-design on the visualization type. Hence, all participants had to solve the same tasks, but depending on the assigned user group, the participants got different kinds of visualizations, supporting them in fulfilling their tasks. The first user group got Gradient Plots, the second user group got Ambiguation Plots, and the third user group only got visualized mean values, so no visual information about the temporal uncertainty was given at all. However, textual information of the uncertainty intervals, was always provided in the task description for all user groups.

We defined four different task types representing typical questions which might be asked when it comes to temporal uncertainties. Therefore, our study consists of four sequential sessions, covering a wide range of possible tasks. For the first task type, the uncertainty interval of the start- or finish-time of some uncertain time event is given. Furthermore there is a specified point in time, usually lying somewhere inside the uncertainty interval. The participants now have to estimate the probability of the time event having already started or ended at this specified point in time. Figure 3.2 shows how tasks of the first session look like for all user groups and how the answer is selected by the participant. Additionally to the task related question, we are also asking for the participant's confidence in his or her given answer.

For the second task type, there are always two parallel uncertainty intervals showing the uncertain finish-times of two possible time events. Again, there is a specified point in time. In these tasks, the participants have to compare the probabilities of both time events at the specified point in time and decide for which uncertainty interval the probability is higher. Figure 3.3 shows how tasks of the second session look like and how the answer is selected.

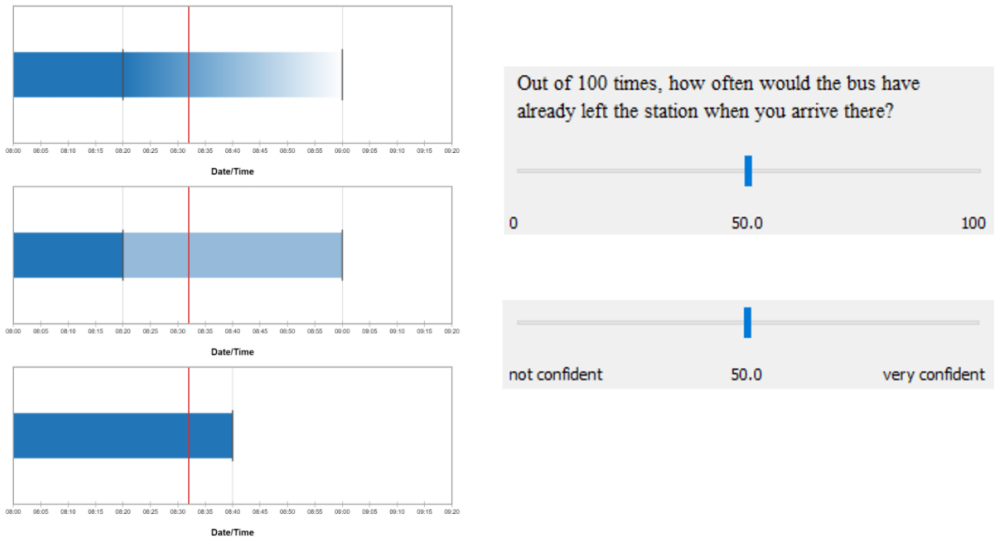


Figure 3.2: *On the left side the different task supporting visualizations for the first session are shown - from top to bottom: Gradient Plots, Ambiguation Plots, mean values. The specified point in time is marked by a red line in all plots. On the right side, there is an extract of the user interface showing the input fields for giving answers. We are not asking for a percentage value, but for some natural count instead, like suggested by Hullman [8].*

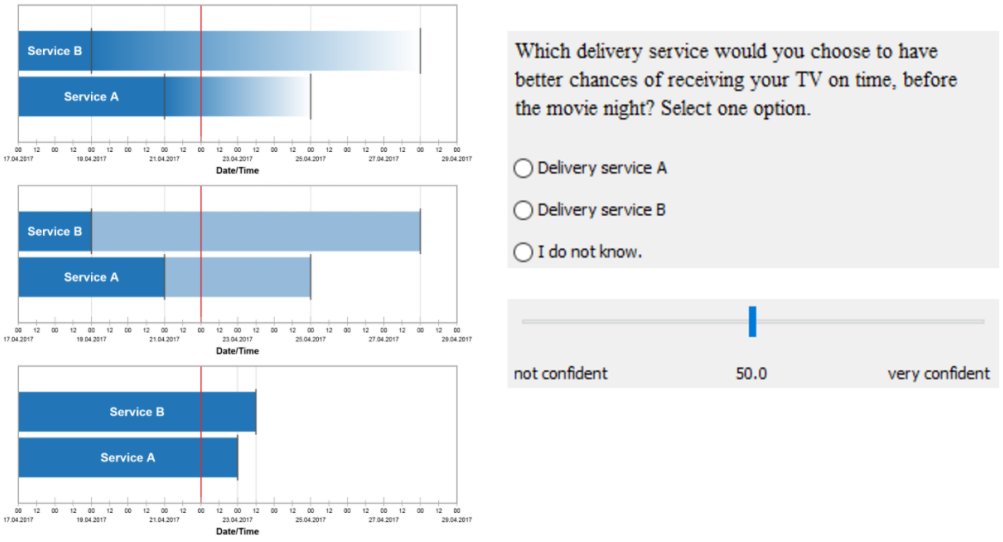


Figure 3.3: *On the left side the different task supporting visualizations for the second session are shown - from top to bottom: Gradient Plots, Ambiguation Plots, mean values. The specified point in time is marked by a red line in all plots. On the right side, there is an extract of the user interface showing the input fields for giving answers. The answer, i.e. the time event with the estimated higher probability, is selected from a set of radio buttons.*

For the third session, tasks are similar to the second session. Again, there are two parallel uncertainty intervals showing uncertain finish-times. However, this time there is no specified point in time, but instead the participants are asked to estimate which event will finish sooner on average. Figure 3.4 shows how tasks of the third session look like and how the answer is selected.

In the fourth session, tasks revolve around overlapping uncertainty intervals of two successive event. So there is some time event with an uncertain finish-time and some time event with an uncertain start-time and those time events are overlapping to some extent. Here the participants have to estimate the probability of the overlap. Figure 3.5 shows how tasks of the fourth session look like and how the answer is selected.

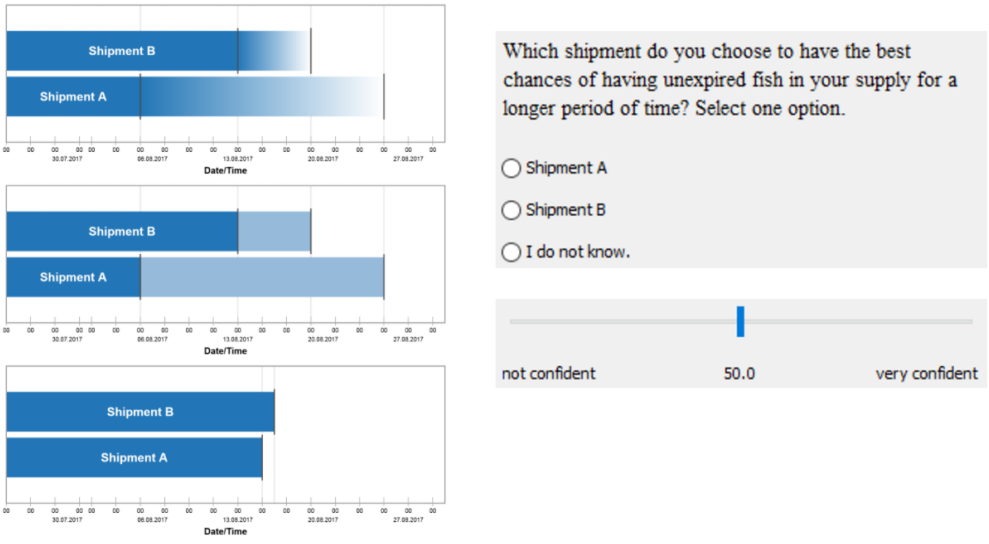


Figure 3.4: *On the left side the different task supporting visualizations for the third session are shown - from top to bottom: Gradient Plots, Ambiguation Plots, mean values. On the right side, there is an extract of the user interface showing the input fields for giving answers. The answer, i.e. the time event which is estimated to end sooner, is selected from a set of radio buttons.*

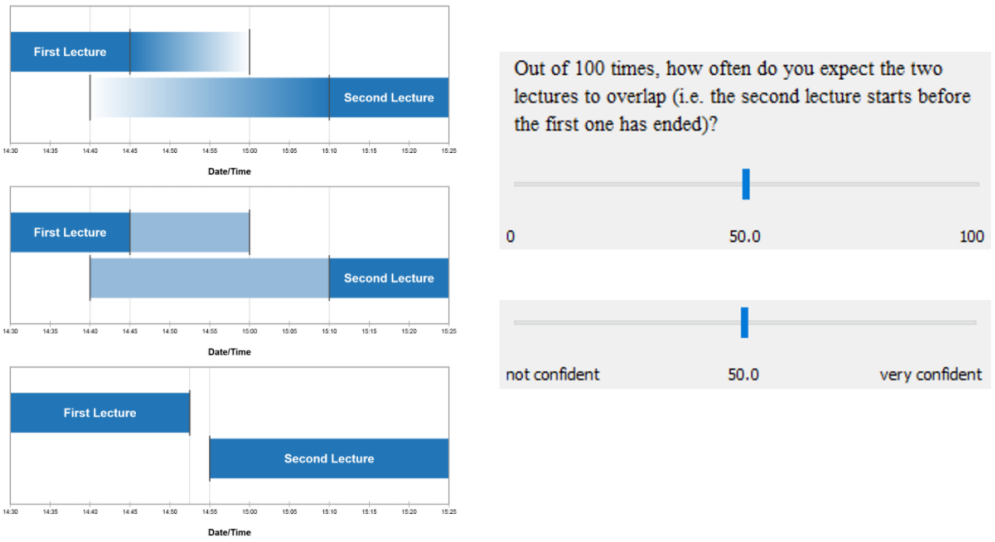


Figure 3.5: *On the left side the different task supporting visualizations for the fourth session are shown - from top to bottom: Gradient Plots, Ambiguation Plots, mean values. On the right side, there is an extract of the user interface showing the input fields for giving answers. Like in the first session, we are again not asking for a percentage value, but for some natural count instead, like suggested by Hullman [8].*

Evaluation

The evaluation of this study is based on three variables, we gathered during the study for each participant:

- error rate
- task completion time
- confidence in given answer

Since each participant performed multiple instances of a task type per session, we calculated the average values of the measured variables per session, resulting into a total of 12 variables for each participant (three average values for each of the four sessions). On those variables we then performed Kruskal-Wallis tests to check for their statistical relevance between the individual user groups.

While inspecting the participants' answers for the individual sessions, we noticed that there have been some obvious misunderstandings, regarding the asked probabilities. Some participants occasionally answered with the complementary probability. In order to still use those test instances, we decided to re-invert the given answers, assuming that the participants knew what they were doing. The criteria for identifying those answers which should be inverted were specified like this: If the correct probability and the participant's answer are at least 20% apart, we check if the complementary probability of the given answer is at most 5% off from the correct

probability. If so, we invert the given answer. If the correct probability and the participant's answer are at least 70% apart, we use a more generous tolerance interval of 10% instead of 5%. After undergoing these manual corrections on misunderstood tasks, the average values of testing variables have been calculated and Kruskal-Wallis tests haven been performed as mentioned before.

3.2 Design of the *Drawing Study*

The *Drawing Study* is designed to be of exploratory nature. This means that the research question it aims to answer is not as concrete as for instance the one of our *Evaluation Study*. The goal is to gain insights into the intuitiveness of visual encodings and to find out how people would visualize temporal uncertainty by themselves. This information could consequently be used in the design of novel visualizations, aimed at expert and especially non-expert users.

To gain these insights, we describe predefined scenarios, which encompass some kind of temporal uncertainty, to our study participants and ask them to draw a visualization sketch that intuitively represents this given scenario. Furthermore, the participants are always provided with a certain task a hypothetical user should be able to efficiently solve given an implementation of the sketched design.

To elicit the desired sketches of temporal visualizations from our study participants, we have to ask the right questions and also have to pay close attention to ask them in the right way. This means that it is imperative to not suggest any possible answers or solution approaches while communicating the task that should be solved, because this would greatly affect their given answers [8]. For this reason we try our best to make the given scenario and the task as clear as possible to our participant without suggesting anything that would help in the solution of the task and would steer them to a specific answer.

The scenarios and task are chosen to be as representative as possible to many typical tasks that can be solved through the visualization of uncertainties. This specification matches the one we already had for our *Evaluation Study*. Hence, we use the same 4 main types of tasks:

1. The first task is to create a visualization that makes it possible to gauge the probability of something for a given point in time in the uncertainty interval. The concrete scenario to be visualized is as follows: "*A bus should arrive at 12:00, but may be running late for up to 10 minutes. How would you visualize this scenario, so that you can estimate the probability of still catching the bus if you arrive at the bus station at a given point in time?*". Figure 3.6 shows an example sketch for this scenario.

Figure 3.6: *The image shows a sketch drawn to represent scenario one of the Drawing Study. It is a conventional line graph for the decreasing probability of catching the bus over time.*

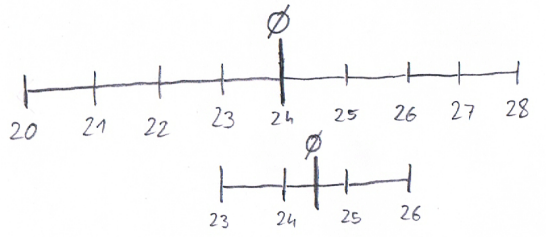


Figure 3.7: *The sketch in the image represents scenario two of the Drawing Study. The extent of the uncertain end of the two given project approaches is represented by two juxtaposed time lines. Their respective average end times are explicitly marked, since these values are important for the user to solve the given task.*

2. The second task is about the comparison of two given uncertain end times of intervals. The task is to judge which of the two intervals will end earlier on average. The concrete scenario is as follows: "*There are two possible approaches to a given project. The first approach will take 20 to 28 days, while the second one will take 23 to 26 days. How would you visualize the scenario, so you can effectively judge which of the two approaches will on average lead to an earlier completion of the project?*". Figure 3.7 shows an example sketch for this scenario.

3. The third task works with the same scenario as the second one and only adapts the user task that the visualization should support. Instead of judging the overall average completion time, the user should be able to make a decision which approach is better to finish the project until a given date. In other words, the user has to compare the completion probability of both approaches in a given point in time. This scenario usually did not lead to helpful answers or additional sketches, but more on that in the Results Chapter.

4. The fourth and last task of the study is about judging the probability of an overlap of two uncertain events. One of the given events has an uncertain end time, while the other event start within an uncertain time frame. The concrete scenario is as follows: "*Two lectures are taking place after each other. the first lecture will end between 11:50 and 12:05, while the second lecture will start between 12:00 and 12:15. How would you visualize this scenario to be able to judge the probability of an overlap of the two lectures? Furthermore, it should be possible to accurately judge the interval in which an overlap can possible take place from your visualization.*". Figure 3.8 shows an example sketch for this scenario.

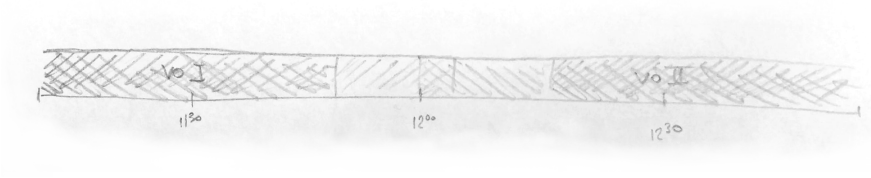


Figure 3.8: *This sketch visualizes the fourth scenario of the Drawing Study. Both lectures are superimposed on the same time line and are marked through their own colors (hinted at through hachures). The overlapping part is colored by mixing the two colors of the lectures.*

The scenarios are not only given through the mentioned texts, but are also more thoroughly explained to every participant, to make sure that everything is understood correctly and the participant knows what the given task is about. These individual explanations make it hard to not say anything that affects the participants way of thinking about the given problem and therefore have an impact on the study results. But none the less we believe that it is important to individually explain the scenarios to every participant to make sure that the tasks at hand are completely clear.

Before the participants are given any task, they are all provided with the same introductory information. This is supposed to set the preconditions of every study session as equally as possible. During the introduction the following three main points are always made clear:

1. The goal is to think of (interactive) computer visualizations. This means that colors, animations and any interaction techniques can be freely incorporated into the sketch. Since the visualizations are only represented as sketches, all this has to be hinted at as far as possible and explained to the study supervisor.

2. The second point tackles a similar problem, since it emphasizes the fact that no participant should omit anything just because it is hard to sketch. The goal of the sketch is to communicate the idea of a visualization design. If any parts of this design are hard to sketch onto paper, it should be done as far as possible and verbally explained.

3. The third point is about the usage of language within the drawing. Since almost every study participant speaks German natively, we did not want to make the explanations of their design and thoughts unnecessarily complicated, by forcing them to speak English. Since we want to use their sketches in our reports and papers though, we asked them to keep everything they write down in their sketches in English.

Evaluation

Our evaluation approach for this study is very similar to the open coding approach of Walny et al. [17]. In that study two researchers separately categorized the collected sketches and afterwards discussed their individual categorization until a consensus was reached. Their final result was a continuum from numeric to abstract representations, in which every sketched was ranked. Since our goal is different, we adapted their approach to fit our need.

Demographic	**Sex**		**Age**		**Computer Usage**	**InfoVis Knowledge**
Male	20	*<30*	24	*Low*	5	16
Female	12	*>30*	8	*Average*	10	11
				High	17	5

Table 3.1: *This table summarizes the demographic of our study participants. A more specific description of it can be found in the previous section.*

Our goal is to find similarities and distinctive features, to find visualization approaches that seem to be intuitive to people. Since we do not know the features by which we are categorizing the sketches, we do not believe a separate evaluation by us would yield good results. Instead we evaluate the collected sketches together and discuss them to find fitting categories. After the categories are determined, every sketch is evaluated. The next step is to count how many visualizations fall in a certain category and to find patterns and trends within these categorizations, in order to finally end up with hypotheses, which could reasonably explain them.

3.3 Participants

In sum we had 32 participants. Two of them only participated in the *Drawing Study*, but not in the *Evaluation Study*. Hence, we collected 30 and 32 datasets for the two studies respectively.

The participants were chosen people from our friends and family. To make the results as representative as possible, we tried to recruit a heterogeneous group. 20 of our participants are male, while 12 of them are female. Recruiting a heterogeneous group in terms of age was not as easy. We ended up with 24 people under the age of 30 and 8 older ones.

To further given an idea of the participant's demographic we estimated their daily computer usage and knowledge in the field of information visualization and ranked those two factors from in the categories low, average and high. Our participant group shows a rather high density of heavy computer users, with a sum of 17 of them. 10 of them are ranked as average users and only 5 of them use computers seldom.

Since some of the recruited participant are study colleagues of us and also study computer science, they have some knowledge about information visualization. In sum 5 of them completed a bachelor course about this topic and were therefore ranked 'high' by us. 11 participants were ranked average, because they either encounter many data visualizations in their daily work life or studied something technical that involves such visualizations, even though there is no specific information visualization course. The remaining 16 users have no education in this field and work with visualizations only seldom.

An overview of the demographic of the participants is presented in Table 3.1.

CHAPTER 4

Results

This chapter encompasses the results of both user studies. For the *Evaluation Study* these are the quantitative results from the statistical tests, that check the validity of our hypotheses. The results of the drawing study are the categorizations of the collected sketches.

4.1 Results of the *Evaluation Study*

For the analysis of results gathered in our *Evaluation Study*, we used R [13] which is a well-known language and environment for statistical computing. We conducted non-parametric Kruskal-Wallis tests [10] in order to verify the statistical relevance of our results. A Kruskal-Wallis test gives information about the likelihood that samples originate from the same distribution or not. In our case, we are evaluating this test on each individual testing variable (error rate, completion time, confidence) for all user groups (visualization types). The output of a Kruskal-Wallis test is a p-value which represents the probability that the labeled samples are from the same distribution. Hence, the smaller this p-value is, the higher is the probability that there are some differences between the individual groups. Our expectations for the results are formulated in the following hypotheses.

H1 Gradient Plots and Ambiguation Plots will perform better than just visualizing the mean value for the tasks of the first and second session.

H2 The visualization of means alone will result into a better and faster performance compared to Gradient Plots and Ambiguation Plots for the third task type.

H3 The fourth task type represents a mathematically more complex problem for which the quantitative result values can not be read directly from the visualizations we used in our study. Therefore we expect all three user groups to have problems with solving these tasks.

The following tables show the determined p-values of the Kruskal-Wallis tests for each testing variable and session. Table 4.1 contains the p-values for the whole result data, i.e. for all three user groups. Table 4.2, 4.3 and 4.4 show the p-values for a pairwise comparison between Gradient and Ambiguation Plots, Gradient Plots and mean values, and Ambiguation Plots and means values respectively. This pairwise comparison was determined by conducting post-hoc Nemenyi-tests.

Session / p-value	**Error-rate**	**Completion-time**	**Confidence**
Session 1	0.4425	0.5034	0.5973
Session 2	**0.1954**	0.5199	0.7977
Session 3	**0.2414**	**0.09689**	**0.1686**
Session 4	0.8466	0.823	0.6358

Table 4.1: *The determined p-values on all testing variables and sessions, between all user groups.*

Session / p-value	**Error-rate**	**Completion-time**	**Confidence**
Session 1	0.60	0.97	0.95
Session 2	0.34	0.81	0.80
Session 3	0.71	0.89	0.99
Session 4	0.85	0.99	0.77

Table 4.2: *The determined p-values on all testing variables and sessions, between Gradient Plots and Ambiguation Plots.*

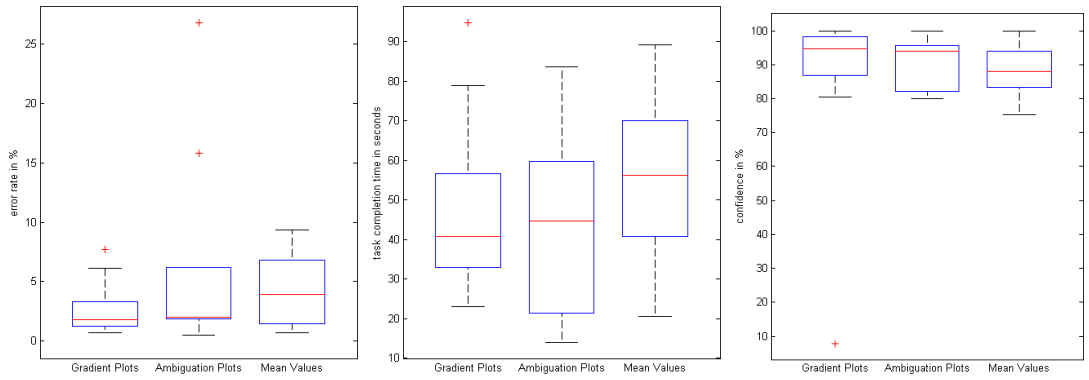
Session / p-value	**Error-rate**	**Completion-time**	**Confidence**
Session 1	0.45	0.66	0.58
Session 2	0.97	0.49	0.87
Session 3	**0.22**	**0.25**	**0.27**
Session 4	0.89	0.83	0.97

Table 4.3: *The determined p-values on all testing variables and sessions, between Gradient Plots and mean values.*

Session / p-value	**Error-rate**	**Completion-time**	**Confidence**
Session 1	0.97	0.50	0.77
Session 2	**0.22**	0.85	0.99
Session 3	0.66	**0.099**	**0.21**
Session 4	1.00	0.88	0.63

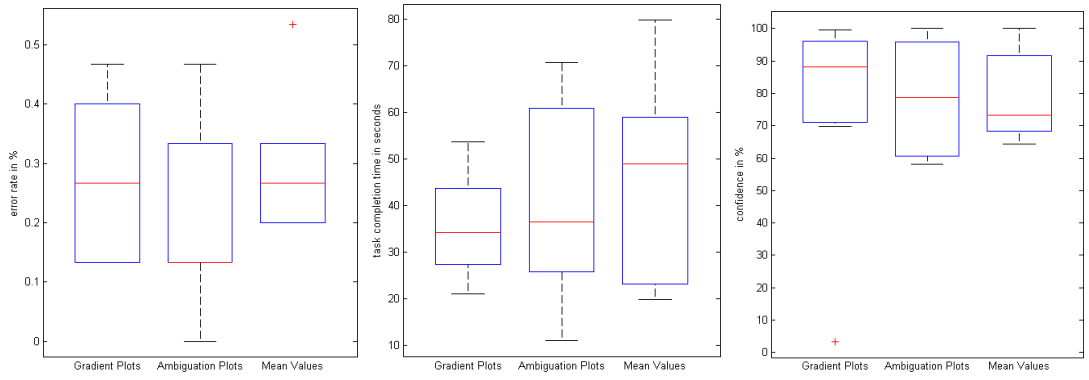
Table 4.4: *The determined p-values on all testing variables and sessions, between Ambiguation Plots and mean values.*

Additionally to this analytical evaluation with p-value tests, we also visually represent the gathered data with boxplots in order to give more information about the data's distribution properties. Figure 4.1, 4.2, 4.3 and 4.4 visualize the results for the first, second, third and fourth session respectively. For the creation of the boxplots, MATLAB [1] has been used.



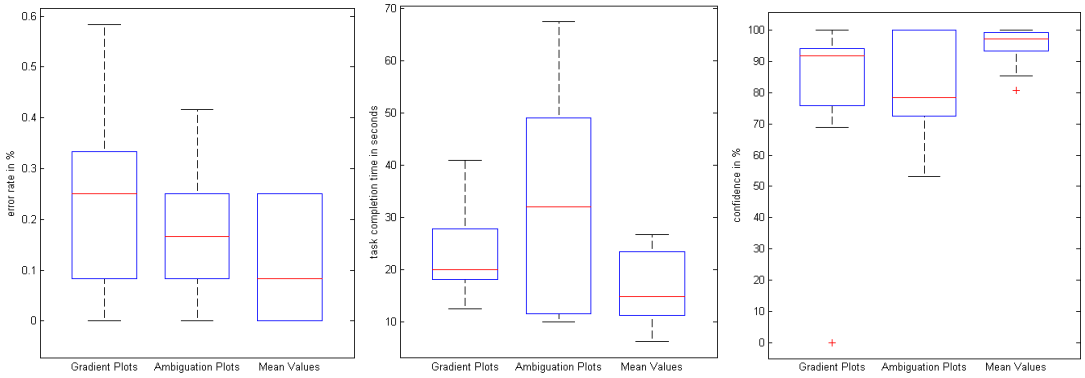
(a) Error rate (b) Completion time (c) Confidence

Figure 4.1: Boxplots visualizing the gathered results of the first session.



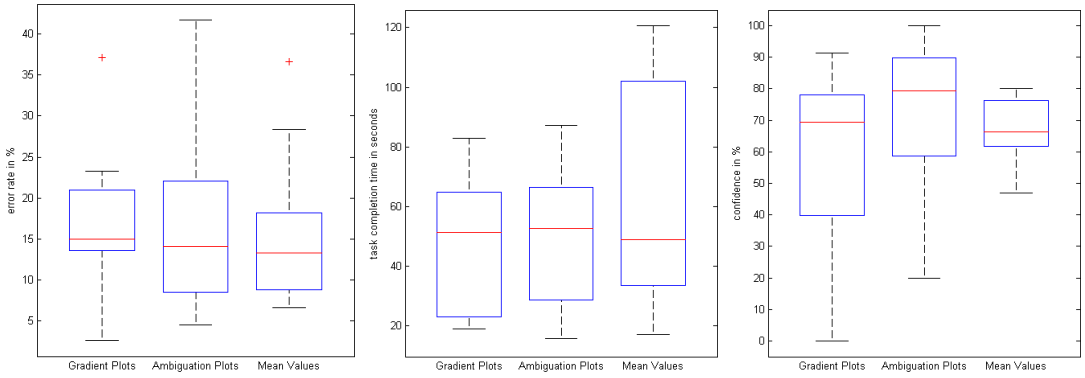
(a) Error rate (b) Completion time (c) Confidence

Figure 4.2: Boxplots visualizing the gathered results of the second session.



(a) Error rate (b) Completion time (c) Confidence

Figure 4.3: Boxplots visualizing the gathered results of the third session.



(a) Error rate (b) Completion time (c) Confidence

Figure 4.4: Boxplots visualizing the gathered results of the fourth session.

4.2 Results of the *Drawing Study*

The results of the *Drawing Study* are presented separately for Task 1 and 4, but Task 2 and 3 share a single table. Originally the study design was meant to lead to four sketches per participant, but since Task 3 built up on the previous one, it only led to verbal comments by our participants or just slight modifications to the existing sketches from Task 2. This means every study session led to three sketches instead of four.

Task 1

An overview of the results of Task 1 can be found in Table 4.5. The first row and column list categories, while the numbers in the table are the sums of sketches that fall into this category. If a sketch matches the **Explicit...** category, it also falls into at least one of the categories of the

first column, since these describe in which way the uncertainty was represented explicitly. The individual categories are described in the following list:

- **Graph** Sketches that count towards this category feature some kind of conventional line graph. Figure 4.5 shows an example sketch, that fits this category. It is also important to notice that every graph visualization also counts toward the **Explicit...** category and *...Length/Height* sub-category.

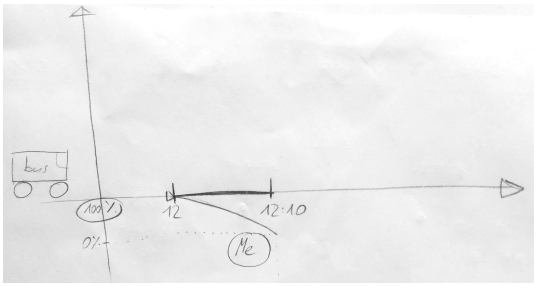


Figure 4.5: *This conventional graph visualization shows the probability of reaching the bus over a given time interval.*

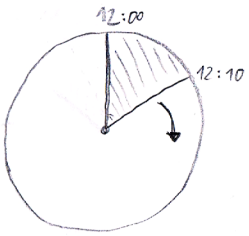
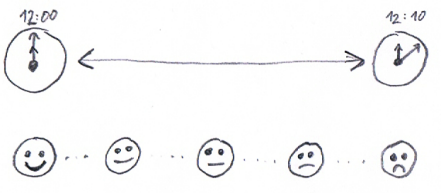
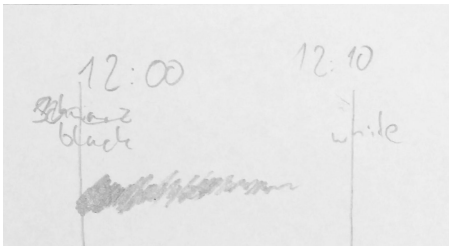


Figure 4.6: *The clock metaphor used in this visualization quickly makes clear, that temporal data is presented. In this case the uncertainty is only given through the bounds of the uncertainty interval.*

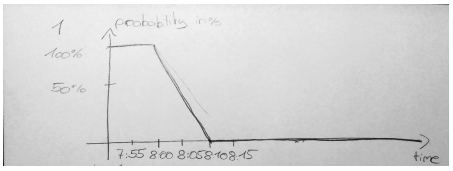
- **Clock** As shown in Figure 4.6, some participants utilized a clock metaphor to visualize time intervals.
- **Explicit...** A sketch counts towards this category, if the uncertainty for a given point in time is somehow represented explicitly, instead of just by the relative position to the bounds of the uncertain time frame. To further split up this category, the following four sub-categories describe the way the uncertainty is represented. Sometimes multiple representations were used in combination, which is why the sum of sub-category counts do not add up to the total number of explicitly represented uncertainties. Figure 4.7 shows an example for every sub-category.
  - *...Icons* This means that the uncertainty for a given point in time was represented by an icon. Often smileys were used for this type of encoding.
  - *...Color* This means that the uncertainty was represented by a color value. This could for instance be a color gradient from one color, representing uncertainty, to another color, representing certainty.
  - *...Length/Height* This means that the uncertainty was given through position of something (e.g. a line). A common example would be a conventional line graph that represents values through the height of a line at a given point.
  - *...Interaction* Sometimes the uncertainty of a given point in time was directly stated in a percentage value. To find out about the probabilities of different time points, user interaction was needed.



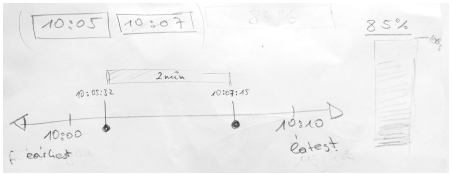
(a) *The clock represent the relevant interval, while the uncertainty over this interval is encoded in smiley faces of varying happiness.*



(b) *The two vertical lines mark the relevant time interval. The uncertainty over this interval is given through a color gradient between those lines.*



(c) *This sketch shows a conventional line graph, that encoded the uncertainty over time in the height of the graph.*



(d) *This sketch shows an interactive visualization. The user has to pick a time interval through sliders to see the probability of catching the bus in this interval.*

Figure 4.7: *The four Figures (a) through (d) show examples for the four sub-categories of explicit uncertainty representation. (a)=Icons, (b)=Color, (c)=Length/Height and (d)=interaction*

- **Bounded** If the uncertainty was not given explicitly, but only through the bounds of the uncertainty interval, the sketch counted towards this category.
- **Time left-right** Here we count how often there is some sort of time line from left to right, instead of any other direction.
- **Time top-bottom** This is the same as the last category, just from top to bottom. These two directions were chosen, because we expected them to be typical directions for time lines.

Task 1	**Graph**	**Clock**	**Explicit...**	**Bounded**	**Time left-right**	**Time top-bottom**
Σ	9	2	19	11	26	0
...Icons			4			
...Color			4			
...Length/Height			11			
...Interaction			3			

Table 4.5: *This table shows the results of the first task of the Drawing Study. The first row of numbers shows the respective counts of every category. The **Explicit...** category is further split up into its four sub-categories and their respective counts.*

Task 2 & 3

The results of Task 2 and 3 can be found in Table 4.6. It features the same categories as the table of Task 1 with two additional ones. The **Superimposed** category counts how often the two project approaches are encoded in the same space in the sketches. Figure 4.8 shows an example for a superimposed representation. If the two approaches are drawn next to each other in their own space, as can be seen in Figure 4.9, the sketch counts towards the **Juxtaposed** category.

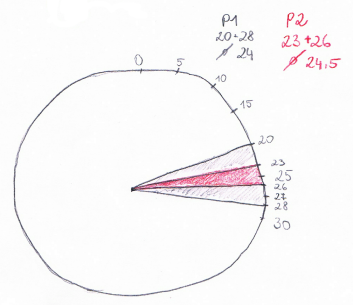


Figure 4.8: *This example shows the superposition of both intervals in the same clock metaphor.*

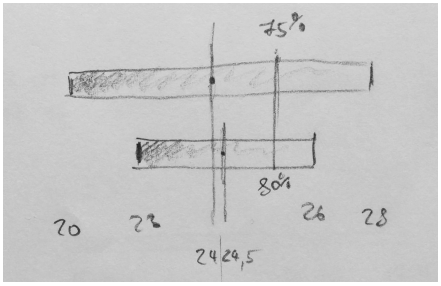


Figure 4.9: *As most sketches this one features a juxtaposition of the two approaches.*

Task 2 & 3	**Graph**	**Clock**	**Explicit...**	**Bounded**	**Time left-right**	**Time top-bottom**
$\sum$	5	1	17	13	25	1
...Icons			0			
...Color			5		**Superimposed**	**Juxtaposed**
...Length/Height			10		7	23
...Interaction			5			

Table 4.6: *This table is structured the same way as the table of Task 1 and features the results of Task 2 and 3. Additionally there are categories for superposition and juxtaposition with their respective counts.*

Task 4

The results of the fourth task are presented in the same way as the last tasks in the corresponding Table 4.7. Additionally, the **Superimposed** category features additional information about the use of color. Since many visualizations featured a superposition of intervals, we counted how often the two intervals were visually separated through color. This count is given in the red number next to the count of superpositions.

Task 4	**Graph**	**Clock**	**Explicit...**	**Bounded**	**Time left-right**	**Time top-bottom**
$\sum$	4	7	10	21	23	1
...Icons			0			
...Color			5		**Superimposed**	**Juxtaposed**
...Length/Height			7		22(11)	9
...Interaction			2			

Table 4.7: *This table encompasses the results of Task 4. It is structured the same way as the previous tables. The individual category descriptions can be found in the list for Task 1.*

CHAPTER 5

Discussion

In this chapter the results of our studies will be thoroughly discussed. The results of the *Evaluation Study* do not show a satisfying amount of statistical significance. The reasons for this and also why these results are still valuable is discussed in the following sub-chapter. In the discussion chapter about the *Drawing Study* the collected sketches will by analyzed, which leads to hypotheses why certain approaches were popular and which representations are intuitive to people.

5.1 Discussion of the *Evaluation Study*

Looking at the evaluation results of the p-value tests, we realize that we did not reach a sufficient significance level for most of our expectations. In the following subsection, we will discuss the results in relation to each of our hypotheses.

Our first hypothesis of the *Evaluation Study*, **H1 - Gradient Plots and Ambiguation Plots will perform better than just visualizing the mean value for the tasks of the first and second session.**, needs to be **rejected**. Our reasoning behind this hypothesis is based on the fact that for these task types we are asking for probabilities at given points in time or ask the user to compare two probabilities. Visualizing the uncertainty intervals might help reading those probability values and therefore make an estimation easier. The smallest p-value regarding this hypothesis represents the error-rate of session 2 with a value of 0.1954 in Table 4.1. However, looking at the tables for pairwise comparison, as well as the boxplots, it seems that this statistical difference in error-rate only holds between the user group with Ambiguation Plots and the user group with visualized mean values. Between the user group with Gradient Plots and the user group with visualized mean values, there is a less noticeable difference. Furthermore, p-values for the other testing variables of the first and second session are quite high (above 0.5) in general. Hence, our hypothesis **H1** cannot be confirmed.

Our second hypothesis, **H2 - The visualization of means alone will result into a better and faster performance compared to Gradient Plots and Ambiguation Plots for the third**

task type., seems to be **plausible** but cannot be confirmed with absolute confidence. For tasks of the third session, participants had to decide between two events which will finish sooner on average. So for these tasks, the mean value actually holds enough information for giving an answer. Showing all the uncertainty information might be unnecessary for the user and maybe just makes the task more complicated. Looking at the results, the error-rate as well as the needed task completion time for the user group having visualized mean values are clearly lower compared to the other user groups, and also the confidence in the given answers is noticeable higher. However, the evaluation by p-value tests does not provide a high enough significance level. While the p-value for task completion time is slightly below 0.1, the p-value for the error-rate is about 0.24. For absolute confidence, these values are not low enough.

Our third hypothesis, **H4 - For the fourth task type we expect all three user groups to have problems with solving these tasks.**, seems to be **plausible**, as p-values (around 0.8) suggest that all samples come from the same distribution, i.e. all user groups are having the same difficulties solving those tasks. Furthermore, the error-rates are about three times higher compared to the first session, were we also ask for quantitative values for estimated probabilities. The error-rates of the second and third session cannot be compared directly in this matters, since users are choosing between two options there instead of specifying a value.

When it comes to statistical significance, we realize that a total of 30 participants, 10 persons per user group each, is not enough for a quantitative user study. The variance in our gathered results tends to be very high, as many people might not answer that precisely. Furthermore, misunderstandings as mentioned before regarding complementary probabilities also affect the variance. However, the difference in performance between the individual visualization types might be small compared to the variance. Hence, we would need a lot of people in order to come to a statistical significant conclusion. For now, we performed the *Evaluation Study* under supervision, allowing the participants to ask us if there are any misunderstandings or ambiguities. In order to conduct our study with a much larger amount of people, we would need to make it more understandable and robust in order to lower the variance in the results and also allow for an unsupervised study.

Another aspect which probably has an influence on the results is the study length. We received feedback from some participants saying that the study, especially the first session, was too long. A longer study consisting of more tasks might be tedious for the user and leads to them losing their focus and motivation. Hence, results might be biased by the participant's mood or concentration. On the other hand, too short studies will produce less results per participants, increasing the variance and therefore decreasing the stability of the results. Also, if the learning phase of a task type takes too long, this affects a shorter study more than a longer study. If we extend our work in the future and make a follow-up study, we need to be careful when choosing the number of tasks.

5.2 Discussion of the *Drawing Study*

Since this study is of exploratory nature, there are no predefined hypotheses we are trying to proof. The goal is rather to interpret the collected drawings and generate hypotheses from this

analysis. These hypotheses will come up during the following discussion and will be highlighted. None of them were proven in any way in the context of this work, which makes them possible topics for future research.

The results of task 1 show that almost two thirds of all drawings feature an explicit representation of uncertainty of some kind. This makes sense, since the task description directly asked to support the user in determining the uncertainty at a given point in time. In this context, especially graph visualizations are common. This leads us to our first hypothesis: **H1 Graph visualizations are intuitive representations to support the user in judging a specific probability value of a given point in time.**

Another explicit uncertainty representation we encountered multiple times is the Gradient Plot, like the one shown in Figure 5.1. This is interesting, since Gschwandtner et al. [6] identified these plots to work very well for this kind of task. If the following hypothesis **H2** holds true, they indeed seem to be a very good choice for those tasks, also if the target user group encompasses non-experts. **H2 Gradient Plots are intuitive representations to support the user in judging a specific probability value of a given point in time.**

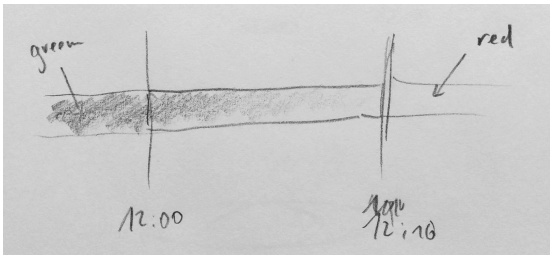


Figure 5.1: *This sketch utilizes a color gradient from green to red to represent the probability of a specific point in time.*

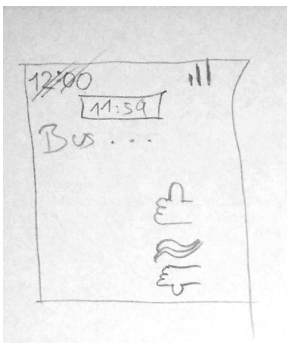


Figure 5.2: *In this design the participant worked with user interaction and icon representations. The user has to enter a point in time and receives an icon, representing the corresponding probability, as feedback.*

Other explicit representations of uncertainty utilize icons to convey the actual probability values to the user. An example for this can be seen in Figure 5.2. We believe this approach to be especially intuitive, but lacks the precision to represent exact values. Therefore we think: **H3 Icon representations, like smiley faces, are a good approach to represent probability values in a very intuitive way, as long as these values do not have to be judged very precisely.**

Even though most visualizations feature an explicit representation of uncertainty, 11 sketches, like the one in Figure 5.3, are of a bounded nature, which only shows the bounds of the uncertain interval, rather than explicit values. If this is the case because this is seemed to be the best way for these participants, or if they simply had no good idea to represent uncertainty explicitly, we do not know. Either way, these representations seem to be intuitive to most people, even though they are not well suited for the task at hand. Gschwandtner et al. [6] showed that this approach

is well suited to convey durations and temporal bounds to the user, which leads us to the following joint hypothesis: **H4 Bounded visualizations are intuitive and effective ways to convey durations and temporal bounds of events with uncertain start and end times to non-expert users.**

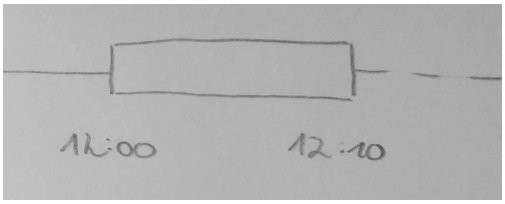


Figure 5.3: *The broader part from 12:00 to 12:10 marks the uncertain part of the event, while the continuous line on the left marks the certain part.*

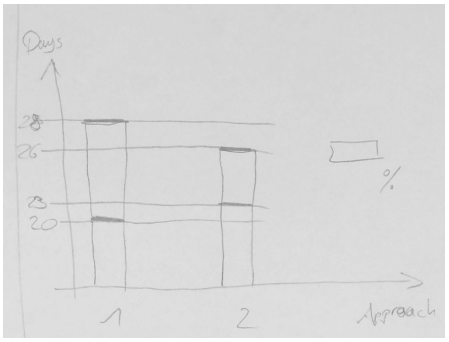


Figure 5.4: *In this vertical bar chart the time is mapped to the y-axis going from the bottom to the top. The two bars represent the two project approaches.*

The assumption that most people would represent time on the horizontal axis from left to right is confirmed by our results. From the total amount of 93 sketches we collected and analyzed, 80% represented time in this way. We also assumed that time would usually be represented from top to bottom, if the time line occupies the vertical axis. This assumption was falsified, since only two sketches featured time this way. The rest of the drawings showed time either in a clockwise manner (clock metaphors) or from bottom to top, like in Figure 5.4. These findings indicate that the question regarding the truth of the following hypothesis can not trivially be answered: **H5 It is more intuitive to a non-expert user group to vertically map time from the bottom to the top, than vice versa.**

In Table 4.6 the results of Task 2 and 3 can be seen. It shows that only five drawings feature graph representations in these tasks. We believe that this is due to the task we asked our participants to solve with their visualization. To judge the average time an event takes, people seem to prefer to see the two uncertain intervals next to each other, instead of superimposed graphs. **H6 If two or more events are compared to each other, it is more intuitive to show them in a juxtaposition, than superimposed in the same space.** This hypothesis is also supported by the numbers of superpositions (7) and juxtapositions (23) in the collected drawings.

Another thing that can be seen in the results table of Task 2 and 3, is that most sketches feature an explicit representation of uncertainty, but there does not seem to be one favorite way to do so. Within the collected drawings uncertainty is represented using graphs (see Figure 5.5), other representations that encoded it in length or height (see Figure 5.6), through color (see Figure 5.7) and through interaction (see Figure 5.8). All of those approaches came up five times within our experiment. Hence, the results do not show any indication of one of those techniques

being more intuitive than others in the context of a comparative visualization. What can be seen though, is that there was not a single icon representation used for these tasks. We believe that the reason for this is that icons do not lend themselves to comparisons, since they also do not represent single values accurately. **H7 Icon representations are not well suited for direct comparison.**

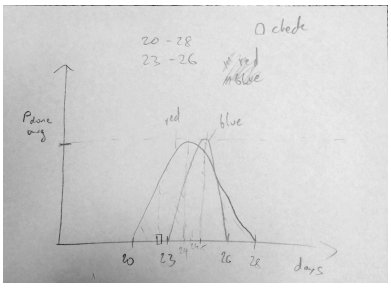


Figure 5.5: *This is a conventional graph visualization, that features two superimposed graphs. The graphs show the possibility of the event ending at the corresponding point in time.*

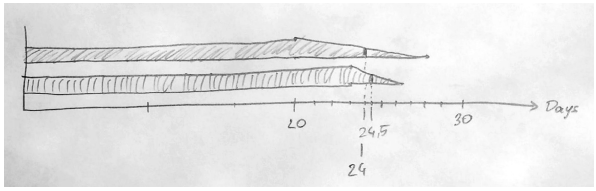


Figure 5.6: *The uncertain time intervals are bounded by the sloping part of the horizontal bars. Additionally the thickness of the bar at a given point in time represents the possibility that the event is still going on at that point in time.*

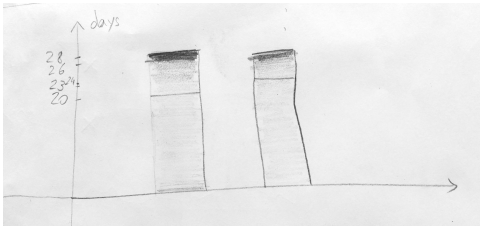


Figure 5.7: *In this vertical bar chart color is used to additionally show the uncertainty explicitly for every point in time.*

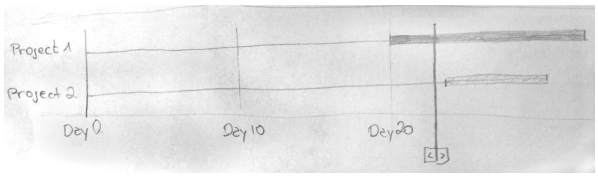


Figure 5.8: *This visualization only shows the bounds of the uncertain time frames, but also shows explicit values through user interaction. The user can slide the slider on the bottom to a specific time point to get information about the respective possibilities of both events.*

Even though most sketches feature uncertainty in an explicit way, there are still many bounded visualizations. We believe that this is due to our task description. To support the user in the goal of Task 2, only the average value of the two intervals are needed. Hence, there is no need to visualize uncertainty to a given point in time explicitly. The fact that uncertainty was drawn so many times, even though it is not relevant for the task at hand, might be an indication that: **H8 Most people prefer to have the underlying uncertainty of data presented to them, even if it is not directly relevant for the task at hand.** An example for the additional visualization of uncertainty can be seen in Figure 5.6.

Table 4.7 shows the results of Task 4. The most apparent difference between these results and the previous ones is that over two thirds of sketches feature the two relevant time intervals in

a superimposed view, instead of a juxtaposition. We believe that this is due to the nature of the task. In contrast to the previous scenario, the two intervals are both happening after each other, with the possibility of overlap and there is not direct comparison of the two. **H9 To represent the amount of overlap between events, it is intuitive to superimpose them in the same view.**

It is also noteworthy that there are more clock metaphors, like the one in Figure 5.9, used in this task, compared to the previous ones. We believe that this is because: **H10 Clocks lend themselves to show two superimposed time intervals, as long as the overlapping area does not exceed a one hour time frame.**

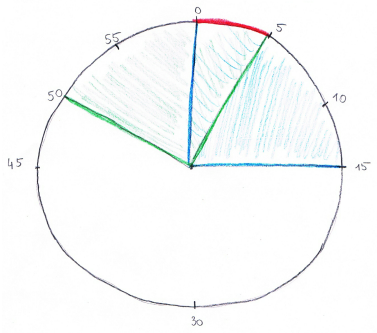


Figure 5.9: *This sketch shows a simple bounded clock visualization. Both uncertainty intervals are colored wedges on the clock. The two colors are mixed within the overlap of the two intervals.*

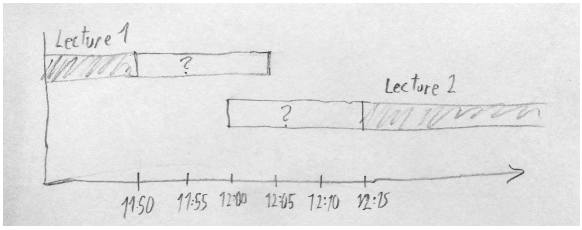


Figure 5.10: *In this sketch both uncertainty intervals are only given by their temporal bounds. The user is not directly supported in figuring out what the possibility of overlap of the two events is.*

Another significant difference to the last task is the lower amount of explicit uncertainty representations. About two thirds of all sketches of Task 4 are bounded ones, like the one in Figure 5.10. We believe that this is due to the complexity of the underlying goal. It is not trivial to visualize the joint probability of overlap of two intervals. Hence, most people do not have a good idea how to do it and simply visualize the bounds of the two intervals. **H11 An elaborate way of visualizing the joint probability of two uncertain events, to represent their overlap possibility is not very intuitive for most people.**

Taking a closer look at the sketches that utilize a superposition of the two intervals, shows that half of them use color to distinguish the intervals from each other. The other half does not need color to distinguish them, since they are characterized by their shape or position, like in Figure 5.11 and 5.12. **H12 Color is an intuitive way of separating two overlapping objects of the same shape.**

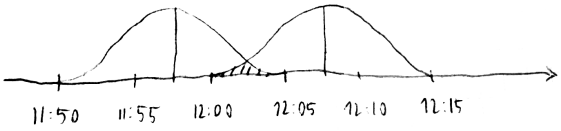


Figure 5.11: *The two graphs, which represent the respective uncertainty intervals of the two events, do not need to be distinguished by color, because they are separated by their position.*

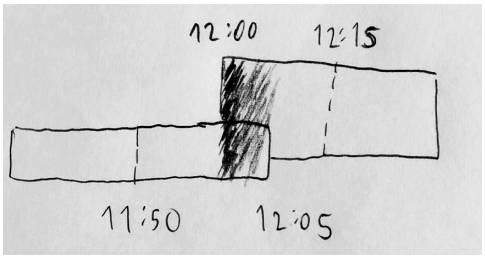


Figure 5.12: *In this case the participant used different shapes/sizes for the two uncertainty intervals, which left the color dimension to encode something else than the distinction between the two events.*

CHAPTER 6

Conclusion

We conducted our *Evaluation Study*, which is building on the work of Gschwandtner et al. [6]. Our goal was to evaluate for which kinds of tasks it makes sense to additionally visualize temporal uncertainty and for which tasks it does not bring any advantages. The study was conducted on three user groups, where each user group was supported by a different visualization type. The different visualization types were Gradient Plots, Ambiguation Plots and visualized mean values. Each group performed the same tasks, representing typical questions which might be asked when it comes to temporal uncertainties. Those tasks have been separated into four successive sessions. We defined three hypotheses which we wanted to verify by this user study. For evaluation of the gathered results, we performed statistical p-value tests, in particular Kruskal-Wallis tests, and also visually investigated the data by using boxplots.

Unfortunately, our study results did not confirm our hypotheses with absolute confidence. While, we had to reject our first hypothesis, our second and third hypotheses can only be considered plausible. The main reason for this is that the amount of participants we asked is too low for a quantitative user study. Still, the results show a visible trend, suggesting that the second and third hypotheses could hold. In order to verify our assumptions, future work might be needed where a more elaborate and more extensive study is conducted.

Additionally to our quantitative *Evaluation Study* we conducted a exploratory study called the *Drawing Study*. This study has the goal of gaining insights about the intuitiveness of visual encodings for temporal uncertainty. Since the study is of exploratory nature, we did not proof any hypotheses we posed beforehand, but rather generated possible hypotheses from the study results, which could be the focus of future research. During the study we asked the participants to think of possible visualizations for given scenarios and tasks, and sketch them. We collected these drawings and analyzed them with an open coding approach. This means that we looked for similarities and distinctive features and defined categories, in which we could classify the sketches. The respective count of every class is the basis for our analysis.

Through our analysis we generated 12 hypotheses, which can be found in Chapter 5. Most of them are only vaguely supported by our results so far. Because of this it is important to address

these in future work and test them through quantitative studies. Since most of the proposed hypotheses regard the intuitiveness of visual encodings in a certain context, we believe that even if they are proven to be true they do not hold much value on their own. The true value lies in the joint insights that can be generated from multiple hypotheses. An example for this is our hypothesis **H2 Gradient Plots are intuitive representations to support the user in judging a specific probability value of a given point in time.** The knowledge that this visualization technique is intuitive is not valuable on its own, but if we combine it with the study results of Gschwandtner et al. [6] that tell us that Gradient Plots are also well fit to support a certain task, we get valuable and deployable insights.

TODO sort out Bibliograpy (delete unused papers)

Bibliography

[1] (2013). *MATLAB version 8.1.0.604 (R2013a)*. The Mathworks, Inc., Natick, Massachusetts.

[2] Aigner, W., Hoffmann, S., and Rind, A. (2013). Evalbench: a software library for visualization evaluation. In *Computer Graphics Forum*, volume 32, pages 41–50. Wiley Online Library.

[3] Aigner, W., Miksch, S., Thurnher, B., and Biffl, S. (2005). Planninglines: novel glyphs for representing temporal uncertainties and their evaluation. In *Information Visualisation, 2005. Proceedings. Ninth International Conference on*, pages 457–463. IEEE.

[4] Chittaro, L. and Combi, C. (2001). Visual definition of temporal clinical abstractions: A user interface based on novel metaphors. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 227–230. Springer.

[5] Correll, M. and Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151.

[6] Gschwandtner, T., Bögl, M., Federico, P., and Miksch, S. (2016). Visual encodings of temporal uncertainty: A comparative user study. *IEEE transactions on visualization and computer graphics*, 22(1):539–548.

[7] Harris, R. L. (2000). *Information graphics: A comprehensive illustrated reference*. Oxford University Press.

[8] Hullman, J. (2016). Why evaluating uncertainty visualization is error prone. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 143–151. ACM.

[9] Kosara, R. and Miksch, S. (2001). Metaphors of movement: a visualization and user interface for time-oriented, skeletal plans. *Artificial intelligence in medicine*, 22(2):111–131.

[10] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

[11] MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., and Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505.

[12] Messner, P. (2000). *Time shapes: a visualization for temporal uncertainty in planning*. Citeseer.

[13] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[14] Robertson, G., Fernandez, R., Fisher, D., Lee, B., and Stasko, J. (2008). Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1325–1332.

[15] Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., and Moorhead, R. (2009). A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE transactions on visualization and computer graphics*, 15(6):1209–1218.

[16] Walny, J., Carpendale, S., Riche, N. H., Venolia, G., and Fawcett, P. (2011). Visual thinking in action: Visualizations as used on whiteboards. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2508–2517.

[17] Walny, J., Huron, S., and Carpendale, S. (2015). An exploratory study of data sketching for visual representation. In *Computer Graphics Forum*, volume 34, pages 231–240. Wiley Online Library.

[18] Xu, K., Rooney, C., Passmore, P., Ham, D.-H., and Nguyen, P. H. (2012). A user study on curved edges in graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2449–2456.