An Empirical investigation on the Visualization of Temporal Uncertainties in Software Engineering Project Planning

S. Biffl, B. Thurnher, G. Goluch, D. Winkler, W. Aigner, S. Miksch
Vienna University of Technology, Institute of Software Technology and Interactive Systems, Favoritenstr. 9-11/188; A-1040 Vienna, Austria
{biffl, thurnher, goluch, winkler, aigner, miksch}@ifs.tuwien.ac.at

Abstract

The success of software projects depends on the ability of a human planner to understand the relationships of tasks and their temporal uncertainty and hence the visualization thereof. In this paper we report on an empirical study that compares the performance of two techniques to visualize task relationships and temporal uncertainties: traditional "best-practice" PERT charts and recently introduced PlanningLines. Main results of the study are: (a) while PERT charts are well suited for reading single attributes, PlanningLines better support users in judging temporal task uncertainty; (b) both experiment rounds shows consistent results regarding the strengths and limitations of the techniques. Overall, these results suggest that a combination of PERT charts and PlanningLines has the potential to significantly improve the planning support of project managers and software engineers.

Keywords: empirical comparison of technique performance, visualization of temporal uncertainty, PlanningLines, PERT.

1 Introduction

Software project management aims at organizing a set of project tasks to meet goals on functionality, budget, and schedule. An inherent difficulty is to understand the impact of task dependencies [1,11,15] (e.g., some tasks can start only after others have completed) and temporal uncertainties, i.e., the range of (a) possible task start and end points in time as well as (b) possible task durations.

Project managers need to understand these temporal uncertainties to spot areas of risk to be able to plan appropriate counter measures in their project plan. A project manager can determine for single tasks the expected range of task durations and also describe how tasks depend on each other.

However, most planning methods used in practice, such as PERT or Gantt charts, are not well suited to express temporal uncertainty for intuitive planning (compare Section 2).

The PlanningLines visualization [1,2] is an approach to combine the advantages of PERT and Gantt charts to show all aspects of temporal task uncertainty in a project context. The PlanningLines technique was originally designed for medical treatment planning and has recently been adapted to project planning purposes: PlanningLines shows the possible distributions of start points, end points, and durations for each task in a project plan with several kinds of related bars. This allows the task planner to intuitively see which tasks have sufficient flexibility in the current overall plan and which tasks may have too little flexibility. Based on this observation the planner can assess the overall risk of the plan and, if necessary, consider focused changes to the plan, with immediate feedback of the impact of these changes to task flexibility.

With the introduction of any new technique an important issue is to measure the performance of typical users (apart from the inventors). In this paper we describe an empirical study that compares the performance of PlanningLines and PERT charts regarding the time needed to conduct a standard set of planning steps and the number of mistakes made. We choose PERT charts since they allow explicitly expressing temporal uncertainties. The focus of the experiment tasks was on cognitive abilities rather than project planning. The experiment was conducted as part of an academic workshop that teaches usability of software user interfaces with a series of interactive examples and empirical studies.

The remainder of the paper is structured as follows. Section 2 summarizes related work on techniques to visualize temporal task uncertainty. Section 3 describes research questions and Section 4 the experiment plan. Section 5 explains the data analysis and

0-7803-9508-5/05/$20.00 ©2005 IEEE

selected results. Section 6 concludes with suggestions for further work.

2 Representing Temporal Uncertainty

A recent empirical study about current project management practices [13] showed two major problems: among the top entries on the list of critical factors were "realistic schedules" and "difficulty to model the real world". Planners in application areas such as medical treatment planning and project management have to deal with inexact knowledge about future activities which translate often into temporal uncertainties.

A core difficulty of planning is to draw up a network of interrelated project tasks, to map the available data to the individual tasks, and to quickly understand, which tasks are truly critical and need focused attention.

2.1 Planning Methods: PERT and Gantt

There are many methods and tool-supported techniques to help planners to visualize their task networks; most widely used are techniques such as PERT (Program Evaluation and Review Technique) and Gantt charts [10].

PERT charts are usually used for the Critical Path Method (CPM). They consist of boxes and arrows, where boxes represent tasks and arrows depict the temporal and logical relationships of tasks (e.g., predecessors and successors). Exact temporal information, like earliest start, latest start, earliest end, latest end, minimum duration and maximum duration are represented as text in the boxes, but not depicted graphically. This data allows for computing the flexibility of tasks in a network. However, the textual notation needs mental calculation and makes intuitive analysis of a task network often rather difficult and time consuming. Compared to Gantt charts, relationships and order of tasks are visualized explicitly and more clearly. Therefore, PERT charts are often used for determining critical paths of a project or depicting them visually. A flaw of PERT charts is that they do not provide a notion for displaying task hierarchies.

Most project management tools provide Gantt and PERT charts as visualization techniques. However, Gantt charts do not represent temporal indeterminacies: they operate on the idea of fixed task duration. Thus they give the impression of exact knowledge about begin, end, and duration of tasks. This can easily mislead the planner to abstract temporal uncertainty from her planning, although the degree of uncertainty can differ widely in a set of tasks in a typical project.

Consequently, planners in practice often neglect their knowledge on variations of tasks' starting, finishing times, and durations, which looses important input on modeling the real world and subsequently makes it harder to draw up realistic schedules.

2.2 The PlanningLines Concept

The goal of introducing a new technique for representing temporal uncertainty, so-called PlanningLines [2], was to provide visual representation of temporal uncertainty for a single activity, to support identification of (un)defined attributes, to help maintaining logical constraints, and to give a direct intuitive visual impression of the uncertainties of a single task in the context of a task network. A technique that supports the above properties well, can also greatly aid to support the identification of critical areas, facilitate the understanding of activity interrelationships and the comparison of activities in (parts of) a project plan.

In this section we want to provide enough information of the notation of PlanningLines to convey the potential of the technique as part of the empirical study. Planning Lines build up on a set of visual representation methods (LifeLines, Paint Strips, Temporal Objects, and Time Annotation Glyphs [11,15]) and combine major strengths while avoiding most limitations. For a more detailed description see [1] and [2].

Figure 1 shows the concept of PlanningLines: a task is modeled as a set of related bars along a calendar scale to provide temporal context. For a single task PlanningLines consist of two encapsulated bars, representing minimum and maximum duration, that are bounded by two caps that represent start and end intervals. The caps are colored in solid black to emphasize their fixed position. The minimum and maximum duration bars are drawn in lighter color to represent some flexibility.

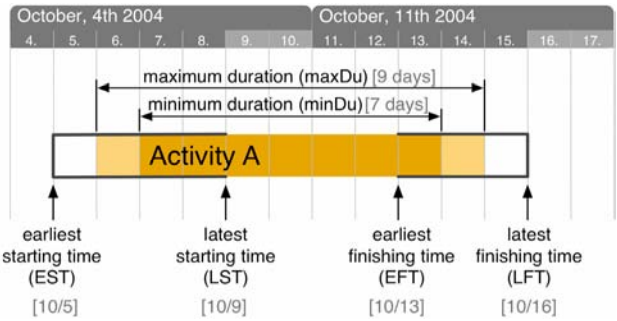

Figure 1. PlanningLines visualization concept [1].

Temporal uncertainties regarding starting, finishing time, and duration of tasks are modeled, similar to

PERT, as intervals including a set of six attributes: earliest/latest starting time, earliest/latest finishing time, and minimum/maximum duration. This implies that the actual start of a task may be any instant within the start interval and a task's end any instant within the end interval while the duration of the task can be any span between minimum and maximum duration.

The visual representation can be remembered easily with simple mental model: The two black caps representing begin and end interval are solidly mounted at the time scale. These caps must hold the minimum and maximum duration bars, which can be shifted within the constraints of the two mounted caps. This mental model corresponds to maintaining a valid attribute set, a number of logical constraints regarding the allowed range of calendar dates and durations (compare [1]).

The design rationale of PlanningLines shows some clear advantages compared to traditional approaches. However, as many potential users are much more familiar with traditional techniques, there is a need to empirically investigate the performance of first-time users of PlanningLines in comparison to a suitable traditional visualization technique.

During the design of our empirical study, we needed to choose a planning method which is able to capture similar information as represented in PlanningLines. PERT charts allow to represent temporal information, temporal uncertainty, and task interrelationships explicitly, but do not give a graphical representation of the temporal dimensions. Gantt charts have a more visual representation than PERT Charts, but do not represent temporal indeterminacies at all: they operate on the idea of fixed task duration. Thus they give the impression of exact knowledge about begin, end, and duration of tasks. We chose PERT charts, because they cover the core temporal information we aimed to compare more appropriately.

3 Research Questions

As PlaningLines and PERT use similar attributes to express the time frame of project tasks (see also Figure 2), this allows comparing and possibly integrating these techniques for use in practice. However, empirical studies are necessary to provide evidence on actual strengths and limitations of the approaches.

While the ultimate goal of the technique is to support professional planners, we focus in this initial phase of research on the cognitive aspect of the technique: the ability to understand the representation correctly and to deduct correct answers to typical questions in a sample project of limited size and complexity.

The study addresses three main research questions with ascending planning difficulty:

1. Can subjects read data from a (correct) PlanningLines representation with similar ease as from a PERT representation?

2. Are subjects faster and/or make fewer mistakes when answering detailed questions on single attributes of a project plan using PERT charts rather than PlanningLines?

3. Are subjects faster and/or make fewer mistakes when judging temporal uncertainties (duration, starting, or finishing times of activities) of interrelated tasks with the PlanningLines representation rather than with PERT charts?

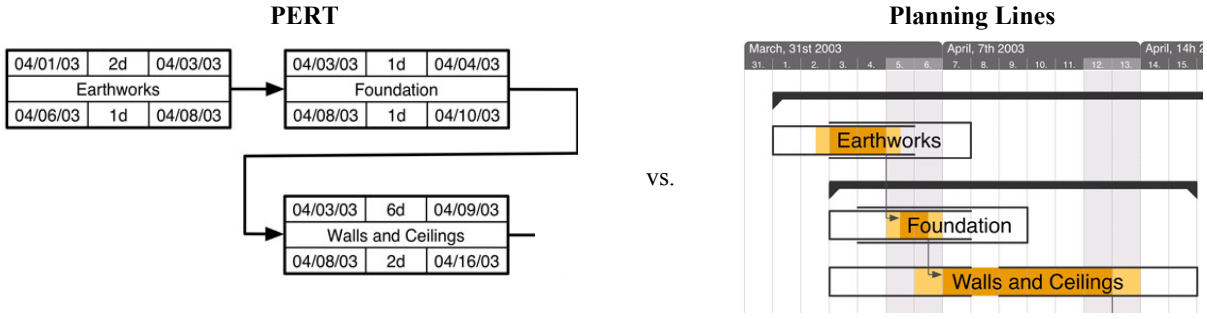

Figure 2. PERT charts and PlanningLines.

In the experiment context we had to use both techniques with all subjects. The resulting 2x2 experiment design allows to compare the results of two rounds for validation and to assess learning effects. For the initial data analysis study [2] we translated the research questions into the following set of hypotheses that we want to re-assess in the context of a second experiment round.

Hypothesis 1: PlanningLines are as simple and intuitive to use as PERT charts. PlanningLines should enable users to read key planning information,

such as calendar dates and durations, from correct graphs in a similar fashion as the alternative technique, in our case PERT charts. We use a set of standard tasks, which can be presented with both techniques, and measure the time needed and the number of mistakes when answering the questions. The initial data analysis study found no significant difference between PlanningLines and PERT users regarding the performance of both mistakes (p=0.468) and duration (p=0.323).

Hypothesis 2: PERT charts are more appropriate than PlanningLines for answering detailed questions on single attributes of a project plan. PERT charts list explicit attributes on calendar dates and duration intervals in tabular form, while PlanningLines use bar notations that need to be linked to calendar dates on the side of the chart (see illustrations in Figures 1 and 2); exact durations have to be computed from these calendar dates. The subjects answered questions referring to specific tasks in a given project plan, mainly multiple choice questions and questions about attributes of selected tasks in this project plan. The initial data analysis study found that PERT users make significantly fewer mistakes than PlanningLines users (p=0.016), while the task duration is not significantly different (p=0.087).

Hypothesis 3: PlanningLines are better suited than PERT to deal with temporal uncertainties regarding the duration, start, or end of activities or plans. While PERT charts explicitly list the attributes of single tasks, they do not show intuitively the flexibility of single tasks in a task network. PlanningLines show the duration range of a task and also the flexibility of start and end dates of single tasks in the context of a task network. Thus this notation allows to quickly and intuitively assess the flexibility of many tasks in parallel and to spot bottlenecks in a task network. When answering questions on temporal uncertainties (on the duration, starting, or finishing time of project tasks), PlanningLines users make relatively fewer mistakes than PERT users.

In the emprical study temporal uncertainties had to be found for specific parts of a given project plan. Those parts were partly simple tasks or sections of the project plan. The initial data analysis study found that PlanningLines users did not make significantly fewer mistakes than PERT users (p=0.086), but they were significantly faster (p=0. 012).

In this paper we use data from the both rounds of the experiment (a) to examine whether hypothesis tests yield consistent or conflicting results in the two rounds and (b) to observe learning effects between the two rounds.

A key aspect of the introduction of a new technique is the learning effort over time [8]. In the experiment we gave a tutorial to all participants and then compare the performance of a well-known traditional technique and the new technique. We discern three different aspects of learning between the two rounds: improved familiarity with the technique, improved familiarity of the problems to be solved, and remembering solutions, if the same problems and data are presented again with a different technique. To minimize the last effect, we used similar problems but with different project data of similar complexity.

As the participants tackle in both rounds problems of similar difficulty, we expect them on average to become faster and to make fewer mistakes both from improving on task familiarity and technique familiarity, similar to experiences in other empirical studies [9]. However, in the tutorial ahead we make participants sufficiently familiar with both the techniques and the tasks to have most of the learning effect within the tutorial and much less between the two rounds of the experiment. Thus we expect a rather modest learning effect and mostly on the task level as the participants solve similar problems with different techniques.

To analyze learning effects we will look at the average difference between the first and second experiment round and compare the change of different groups that use a technique in their first or second round to observe the technique-with-task effect. In the second round participants may focus better on the technique at hand; however, there may be a trade-off between the speed of answering questions and the number of mistakes made.

4 Experiment Description

This section describes the subject background, plan for a controlled experiment with a 2x2 counterbalanced design, the experiment artifacts and procedures, and finally discusses threats to validity.

4.1 Subjects

The experiment participants were 48 students of the workshop ‘Usability Engineering’ in an academic environment. The subjects were graduate students of informatics and business informatics. The typical student worked part-time, and had at least some professional experience with software development. Most students knew the concepts of PERT and Gantt charts from prior university courses on project management. As could be expected, none of the participants had heard of PlanningLines before the tutorial.

4.2 Experiment Design

The experiment was conducted as part of an academic workshop. As both representation techniques are believed to have substantial merit to the students' education, we decided to let all students work in depth with both techniques. This led to a 2x2 experiment design as basic setup. The independent variables in the experiment are:

- *Visualization Technique:* These are *PlanningLines* or *PERT charts*, which we believe to influence subject performance.
- *Experiment round*: First or second round.
- *Project data set*: Two sets "Data set 1" and "Data set 2".

The dependent variables in the experiment are related to subject performance.

- *Number of mistakes* when answering a standard set of questions regarding planning problems; from this variable we can derive the *relative number of mistakes* as share of all possible mistakes in an experiment part (0 to 100%).
- *Duration* for answering these questions in minutes.

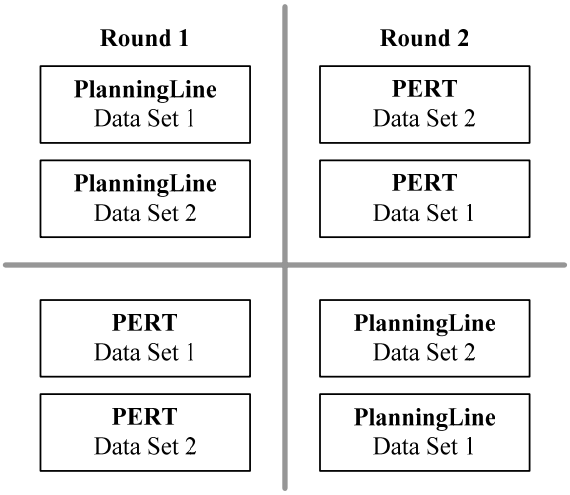

Figure 3. 2x2 experiment design; sequence of execution: from left to right.

Figure 3 shows the assignment of visualization technique (PERT or PlanningLines) and project data set (data set 1 or data set 2) to experiment groups in rounds 1 and 2 of the experiment. The number of 48 participants allowed a counterbalanced design: We randomly assigned the students to one of the four groups in round 1 (left column in Figure 3), 12 students to each group. By randomization we forced unknown source of discrepancy to contribute homogeneously to the treatments, following the suggestion presented in [3]. In the second round, each group worked with the alternative visualization technique and project data set. In this setup we can analyze all possible combinations of visualization technique, project data set, and experiment round. This design allows to investigate whether the visualization techniques perform similarly with the two project data sets.

4.3 Experiment procedures and artifacts

This section provides a short overview of the process steps and experiment artifacts used in the presented study.

The participants had no knowledge on the PlanningLines method. Since the subjects in our study have varying degrees of experience with the PERT method, we conducted a tutorial, held by one of the experiment design members, to ensure a baseline of familiarity with the method. The tutorial briefly repeated how to use PERT, a method known by most participants, and introduced the new method 'Planning Lines', to guarantee the minimal common level of knowledge for the experiment. The participants applied the techniques to small examples to ensure some familiarity; further the tutorial discussed typical difficulties with the types of problems in the experiments, and typical novice mistakes.

The tutorial and training session was followed by filling in a questionnaire on participant experience with project planning and visualization techniques. Then the first experiment round started according to the groups shown in Figure 3; the first round finished with filling in a feedback questionnaire on the ease of use of the visualization technique used. The second round followed after a break in a process similar to the first round, but with the alternative visualization technique. Each round took up to 45 minutes; participants could take a break to refresh. Subjects were asked to take time stamps when starting and finishing a part of the answering sheet or when taking a break. This allowed to measure the time a participant needed to solve the given questions and tasks in a part. A supervisor was assigned to every group to provide assistance and to make sure the participants understood their respective tasks.

The experiment participants received the following experiment materials.

1. *Background Questionnaires:* A one-page questionnaire was provided at the beginning of the experiment. Participants were asked for demographic information and specific information about their experience with PERT and other visualization graphs.

2. *Answering sheets for task solutions;* Four different versions of this part were available, one per experiment group, according to the visualization technique (PlanningLines, PERT) and project data set (1, 2).
   a. *Part A* consisted of a three-page answering sheet for questions and tasks, concerning the usage of PlanningLines or PERT.
   b. *Part B* consisted of a project plan and a five-page answering sheet for questions based specific tasks in a simple project plan.
3. *Feedback Questionnaire:* At the end of the experiment, every participant gavehis or her feedback to the visualization technique used; the questionnaire was adapted from the *Technology Acceptance Method* questionnaire [5].

4.4 Threats to validity

In every empirical study there are possible threats to the validity of the study that need to be acknowledged and mitigated with appropriate countermeasures.

Internal validity. A potential problem in any experiment is that some factor may affect the dependent variables without the researcher's knowledge. This possibility must be minimized.

History: Changes in dependent variables may be due to other events, e.g., communication or collaboration within a group and between groups (plagiarism). During the study we had between 4 to 6 persons from the experiment team who supervised the subjects and answered questions, if necessary. The experiment team paid special attention to communication and plagiarism and motivated the participants to work on their own solutions. We did not give feedback on the solutions to experiment problems.

An issue raised in empirical studies on reading techniques is the possibility that participants use prior know-how to solve their tasks using another technique than prescribed by the experiment. The potential of this kind of threat is low in this study context as the tasks the participants had to complete were considerably less work than to create an alternative visualization representation.

Maturation: Effects coming from processes taking place within subjects like tiredness, boredom, or learning apart from the experiment. During the experiment the subjects could take brakes whenever they felt like they needed one. Some students took a 5-minute brake between rounds 1 and 2 but most of them preferred to continue working.

Testing: Subjects get familiar with the tests, e.g., the project plan. We did not provide feedback on experiment results to subjects during the experiment. Moreover, we had different project data sets in rounds 1 and 2; so the subjects had to recalculate durations and could not use data remembered from the previous round. These data sets were similar with respect to structural and mental complexity.

Further, the experiment team collected all questionnaires and material from round 1 before they distributed the material for round 2.

External validity. However, there can still be a number of external threats to validity which we tried to avoid or control as much as possible.

Interaction of selection and treatment: Selection of sample different from target population. As stated in [3], an external threat to validity is not meeting *setting representativeness;* this threat refers to the issue of having a setting or material which is not comparable to an industry setting or material. As control technique we took a technique (PERT) that is widely used in practice. As our setting was aimed at investigating the cognitive understandability of PlanningLines in comparison to PERT charts, the selected tasks seem appropriate. Further, we executed an extensive pilot test of the material to assure correctness.

Sjøberg [12] states that "one should be aware that it may be a methodological problem that the teacher is also the researcher, that is, the technology being subject of an experiment run by a given researcher is also being taught by the same researcher. Consequently, the students might be biased". We were aware that the researcher/teacher can be biased through "wishful thinking" [14]. In order to avoid these biases the evaluation was carried out by colleagues who where not involved in the development of PlanningLines nor were the students from a class of those researchers. Development and evaluation were completely separated.

There is an ongoing discussion in the empirical software engineering community whether student subjects can provide valid results [6,7,12]. In this study, the emphasis is on cognitive tasks; the necessary skills to apply the planning techniques are the ability to read calendar dates and basic calendar arithmetic to calculate duration and calendar dates corresponding to entities of a visualization technique. All participants easily could show sufficient skills as part of the tutorial. We investigated mainly cognitive abilities of subjects rather than their project management abilities. Thus using students for the study seems not to pose a considerable problem.

5 Data Analysis Approach and Results

After counting correct and incorrect answers on parts of the answering sheets, the experiment team calculated the experiment duration per part and experiment round.

We tested the performance of techniques and groups with the two project data sets in both rounds to see whether we can simplify subsequent analysis steps. If we can show that the project data sets are indeed equivalent, we can combine data from groups with the same technique, but different project data sets. Otherwise we would have to analyze the data of the four experiment groups separately and would consequently loose statistical power.

From the experiment design (see Figure 3) we have four groups of combinations of two techniques (PlanningLines and PERT) and two project data sets (1, 2). We tested the performance (relative number of mistakes and duration; see defintions in Section 3) of groups that use the same graphs set but different data sets (PlanningLines1 – PlanningLines2 and PERT1 – PERT2). As the p-values regarding performance differences of data sets range from 0.401 to 0.601, no significant differences between both the two PERT and PlanningLines data sets could be found.

For the remainder of analysis we compare the performance of different graph methods regardless of the data set used, with 2 groups of 24 participants for each technique and round. Note that we encountered problems with missing data in parts B1 and B2 and round 2 for the time stamps with one and two subjects, respectively. Otherwise, all groups were of similar size.

We used the t-test and the Mann-Whitney test. As both tests consistently showed similar results, we report the p-values from the t-test. The statistical tests were performed on an alpha level of 0.05.

The hypothesis tests of round 1 were reported in an initial data analysis [2]. In this Section we report in more detail descriptive data and tests for three key hypotheses that map to the data of answer sheet blocks. Each result block reports descriptive data on duration (in minutes) and mistakes (relative to the total number of possible mistakes when answering questions in that part). We analyze the changes between rounds 1 and 2 for similar techniques to observe the effect of improved familiarity with tasks in combination with the visualization techniques. As next step we looked at the results of hypothesis tests to find out whether the results of rounds 1 and 2 provide a consistent picture.

5.1 Results Part A: Reading Attributes

Tables 1a and 1b report descriptive data on duration, and the relative number of mistakes for part A of the experiment answering sheet. The maximum duration for part A was around 20 minutes with average durations between 10 to 12 minutes in the first round and around 7 minutes in the second round. PERT users were on average somewhat slower that PlanningLines users. The users of both techniques were consistently faster in the second round (by 3.3 to 3.7 minutes). We attribute the shorter average duration mainly to better familiarity with the problems to solve as the differences between the average duration of the participants using the two techniques do not change between the experiment rounds. For PERT users the level of mistakes increased on average in the second round of the experiment contrary to our expectation; however, there may be a trade-off between problem solving speed and quality of the answers, which needs to be investigated in more detail.

Table 1a: part A; duration in minutes.
PL stands for PlanningLines

H1	*Round 1*	*p-value: 0.323(-)*
PERT	mean:11.5 min.	std.dev.: 3.7 min.
PL	mean: 10.3 min.	std.dev.: 4.6 min.
H1	*Round 2*	*p-value: 0.412(-)*
PERT	mean: 7.8 min.	std.dev.: 3.8 min.
PL	mean: 7.0 min.	std.dev.: 3.3 min.

Table 1b: part A; number of mistakes (%).
PL stands for PlanningLines

H1	*Round 1*	*p-value: 0. 468(-)*
PERT	mean: 9.0%	std.dev.: 6.0%
PL	mean: 10.4%	std.dev.: 6.4%
H1	*Round 2*	*p-value: 0.007(S)*
PERT	mean: 12.7%	std.dev.: 6.1%
PL	mean: 8.0%	std.dev.: 5.6%

Tables 1a and 1b show the p-values for hypothesis tests (in italics) regarding the performance variables of the two visualization techniques on both experiment rounds; in parentheses you find whether the p-value is significant (S) or not (-). The first hypothesis tests whether there is a difference in duration or level of mistakes when reading attributes from a correct representation. As expected, the tests show no significant differences in round 1 and for duration in general, while there is a significant difference in the

level of mistakes in the second round due to the surprising increase of mistakes from PERT users.

5.2 Results Part B: Uncertainty of Single Attributes

In the same structure as in the previous subsection, tables 2a to 2b investigate the performance of participants in answering detailed questions on single attributes of a project plan. Again, the average duration decreases per round for both techniques. In the second round, participants are significantly faster to find answers using PERT charts. At the same time the PlanningLines groups take on average 1.6 to 2.0 minutes (22% to 44%) longer to answer questions and also tend to make on average 6.96 %-points to 8.6 %-points (50% to 85%) more mistakes. Complementary to the average difference in duration between the two techniques there is also a learning effect between rounds: on average by 1.8 to 2.2 minutes between rounds. Improved familiarity with the problems at hand has in the study context a much stronger impact on the average duration of task than on the average level of mistakes. In this part of the experiment answering sheets, we can observe the highest level of defects, which warrants more detailed research on the source of the mistakes; an average level of 10% to 20% mistakes seems hardly acceptable for use in practice.

Table 2a: Part B-1; duration in minutes.
PL stands for PlanningLines

H2	*Round 1*	*p-value: 0.086(-)*
PERT	mean: 6.7 min.	std.dev.: 2.5 min.
PL	mean: 8.3 min.	std.dev.: 3.9 min.
H2	*Round 2*	*p-value: 0.026(S)*
PERT	mean: 4.5 min.	std.dev.: 1.8 min.
PL	mean: 6.5 min.	std.dev.: 3.3 min.

Table 2b: Part B-1; number of mistakes (%).
PL stands for PlanningLines

H2	*Round 1*	*p-value: 0.016(S)*
PERT	mean: 10.5%	std.dev.: 7.4%
PL	mean: 19.1%	std.dev.: 14.8%
H2	*Round 2*	*p-value: 0.083(-)*
PERT	mean: 11.0%	std.dev.: 10.2%
PL	mean: 17.9%	std.dev.: 16.0%

Hypothesis 2 proposes that the PERT chart is more appropriate for reading single attributes from a project task in a task network than PlanningLines. In round 1 the techniques show similar duration, while PlanningLines users make significantly more mistakes. In round 2, PERT users take significantly shorter, while the comparison of mistakes is not conclusive.

5.3 Part B2: Judging Temporal Uncertainty

This part concerns the key feature of the PlanningLines representation: better understanding of temporal uncertainties and judging the flexibility of single project tasks or project plans. Tables 3a and 3b show that in this part all participant groups made very few mistakes – on average only every 20th to 40th answer was a mistake. However, PlanningLines users are consistently in both rounds faster and tend to make fewer mistakes.

PERT users take on average 1.8 to 2.2 minutes (25% to 40%) longer that PlanningLines users and make on average almost twice as many mistakes.

As in part A, the shorter average duration seems to come mainly from better familiarity with the problems to solve as the differences between the average durations do not change between the experiment rounds. The average difference in duration between the two techniques is on average 2.0 to 2.4 minutes between the experiment rounds.

Table 3a: Part B-2; duration in minutes.
PL stands for PlanningLines

H3	*Round 1*	*p-value: 0.007(S)*
PERT	mean: 9.5 min.	std.dev.: 2.7 min.
PL	mean: 7.3 min.	std.dev.: 2.6 min.
H3	*Round 2*	*p-value: 0.001(S)*
PERT	mean: 7.1 min.	std.dev.: 2.2min.
PL	mean: 5.3 min.	std.dev.: 1.5 min.

Table 3b: Part B-2; number of mistakes (%).
PL stands for PlanningLines

H3	*Round 1*	*p-value: 0.086(-)*
PERT	mean: 4.5%	std.dev.: 3.8%
PL	mean: 2.7%	std.dev.: 3.3%
H3	*Round 2*	*p-value: 0.012(S)*
PERT	mean: 4.4%	std.dev.: 3.4%
PL	mean:2.2%	std.dev.: 2.3%

Hypothesis 3 proposes that the PlanningLines representation is better suited to deal with temporal uncertainties regarding the duration, begin, or end of activities or plans. This hypothesis is confirmed regarding duration in both rounds and regarding the relative number of mistakes in the second round.

6 Conclusion and Further Work

In this paper, we reported on an empirical study to compare the performance of two visualization techniques in a two-round experiment. Main results of the study are: (a) while PERT is well suited for reading single attributes, PlanningLines better support users in judging temporal task uncertainty; (b) the second experiment round shows consistent results regarding the strengths and limitations of the techniques; and (c) some learning effects as the participants took less time to complete their tasks and/or made less mistakes.

In general, the results confirm the set of hypotheses derived from our research questions consistently in both rounds of the experiment. The results of the experiment confirmed the superior performance of PlanningLines regarding judging temporal uncertainty of tasks in a project plan task network, an important asset for more realistic software engineering project planning. The complementary strengths of the investigated techniques warrant research on their combined application for managing temporal uncertainties in task networks

The learning effects in the experiment showed no disadvantage for PlanningLines compared to the widely used alternative technique, PERT charts; a surprisingly good result. The level of effort and the learning effect between rounds is comparable to the well-known PERT charts; this seems a remarkable performance of experiment participants with a technique they did not know before. As result from participant feedback we will add questions on the experience with related techniques to the background questionnaire to get a broader view on the experience of the participants.

In the experiment results, we can observe consistent relationships between the performance variables of the visualization techniques. Further we can also see a fairly consistent learning effect in the shorter duration when comparing groups who use the same technique in different experiment rounds. Due to the experiment design that limits interaction with the technique and the project data sets, we attribute these learning effects to the improved familiarity of the participants with the type of problems they had to solve. For further work we suggest more formal modeling of the learning effects to determine the effect size as input to planning the effect size of future experiment designs.

An interesting issue is the considerable variation of the level of mistakes between the parts of the experiment: Average mistake rates range from around 3% to 20%. These differences clearly warrant a more detailed investigation into the cause of these mistakes as the mistake levels are consistent in both experiment rounds and show only little learning effect, contrary to task duration. The highest level of defects can be observed with PlanningLines users in part B-1; there may be a systematic difficulty to relate the position of a PlanningLines graph with the possibly distant time axis on the side of the overall model. A combination with PERT charts can help to avoid this source of mistakes. However, also PERT users have a rather high rate of mistakes in this part of the experiment; which needs to be considered when using the technique in other environments.

The results of the study further suggest that the techniques under study have complementary strengths and limitations: While PERT is well suited for interacting with explicit single attributes of project tasks, PlanningLines are better for intuitively understanding the flexibility of tasks in a project plan context.

As the techniques share a common set of attributes to describe project tasks, it is possible to combine the strengths of both techniques: PERT can be used for input and change of data attributes of individual tasks in a network, while PlanningLines can be used to view the data and analyzed potential risk areas in the task network.

As the results are promising and consistent in both experiment rounds, we suggest further research in this direction: building up on present research on the cognitive level in an academic environment we will extend the research to more complex project planning contexts and work with professional software project managers.

With this step we can find out whether the visualization approach actually scales up for project analysis and planning. Work with professionals needs better tool support that seamlessly fits to the environment used by the target population. Currently, we develop a prototype to generate PlanningLines representations from industry-strength project plans, such as Micosoft Project. With this input we can examine in empirical studies with practitioners as subject whether they perform similar the academic subjects.

These studies can give valuable insight for project planning improvement in industry and research environments. As our partner companies are aware of the added value and important insights they can gain from empirical studies, several future studies are planned. Those studies aim at integrating the study results in the development cycle and foster a bottom-up approach [3] for the application of empirical data in industry.

Acknowledgements

Many thanks to Peter Messner, for reviewing the experiment design, to Christoph Gesperger, Manuel Ganglberger, and Christoph Fleury, for their support in preparing the experiment material, and to the experiment participants. This project is partially supported by "Fonds zur Förderung der wissenschaftlichen Forschung - FWF" (Austrian Science Fund), grant P15467-N04.

References

[1] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl, "*PlanningLines Usability Studie – User Studie zum Vergleich von PlanningLines und PERT Darstellung*", Technical Report. Vienna University of Technology, Institute of Software Technology and Interactive Systems (in German), 2005.

[2] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. "PlanningLines: Novel Glyphs for Representing Temporal Uncertainties and their Evaluation", in *Proceedings IEEE IV05*, IEEE Comp. Soc., London, 2005. pp. 457–463.

[3] V.R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Soerumgaard, and M. Zelkowitz, "The Empirical Investigation of Perspective-Based Reading", Empirical Software Engineering: An International Journal 1, 2 (1996), pp. 133–164.

[4] G.E.P. Box, W.G. Hunter, and J.S. Hunter, *Statistics for Experimenters*, John Wiley & Sons, 1978.

[5] F.D. Davis, R.P. Bagozzi, and P.R. Washaw, "Perceived Usefulness. Perceived Ease of Use, and User Acceptance of Information Technology" *MIS Quarterly, 13:3*. p. 319-340, 1989.

[6] C. Denger, M. Ciolkowski, and F. Lanubile, "Investigating the Active Guidance Factor in Reading Techniques for Mistake Detection", in *Proceedings of the 3rd International Symposium on Empirical Software Engineering (ISESE'04)*, 2004, pp 219-228.

[7] B. Kitchenham, S. Linkman, and J. Fry, "Experimenter Induced Distortions in Empirical Software Engineering", in *Proceeding of the Workshop for Empirical Software Engineering (WSESE'04)*, Rom, 2004.

[8] S. Makridakis, S.C. Wheelwright, and R. J. Hyndman *Forecasting methods and applications,* New York: John Wiley & Sons, 1998.

[9] M. Morisio, D. Romano, and C. Moiso. "Framework-Based Software Development: Investigating the Learning Effect", in *Proceeding of the 6th International Software Metrics Symposium*, 1999, pp. 260–268.

[10] G. Patzak and G. Rattay, *Projektmanagement*, Linde, Wien, 4th edition, 2004.

[11] C. Plaisant, "The Challenge of Information Visualization Evaluation", in *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2004)*, ACM, 2004, pp. 109-116.

[12] D. Sjøberg, B. Anda, E. Arisholm, T. Dyba, M. Jorgensen, A. Karahasanovic, E.F. Koren, and M. Vokac, "Conducting Realistic Experiments in Software Engineering", in *Proceedings of the 1st International Symposium on Empirical Software Engineering (ISESE'02)*, 2002, p. 17.

[13] D. White and J. Fortune, "Current Practice in Project Management – An Empirical Study", *International Journal of Project Management*, vol. 20(1), pp. 1-11, 2002.

[14] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering – An Introduction*, Kluwer Academic Publishers, 2000.

[15] P. Zhang and D. Zhu, "Information Visualization in Project Management and Scheduling", in *Proceedings of the 4th Conference of the International Society for Decision Support Systems (ISDSS'97)*, 1997, pp. 1-9.