COMPSCI 715
Advanced Computer Graphics

Writing about Evaluations

Today's Mission

1. How do you report evaluation results?
2. How does a good evaluation section look like?

Typical Research Paper Structure

1. **Introduction**: What is the research problem? Introduce and motivate it. Summarize your contributions.
2. **Related Work:** What have others done? How is it different? Cite, summarize other solutions & compare it with your own.
3. **Design:** Your solution. Describe it in enough detail so others can implement / replicate it. Software architecture (e.g. class diagram)? User interface (e.g. screen diagram)? Algorithms?
4. **Implementation:** How have you implemented your solution? Tools and technologies used? Implementation challenges?
5. **Evaluation:** Explain the methodology you used for evaluation. Present the results. Discuss them.
6. **Conclusion**: Summarize contributions. Point out future work.

Writing about Evaluations (1-4 pages 2-column)

Provide empirical evidence of the quality of your contributions.

1. Methodology subsection: describe how you conducted the study, i.e. variables, tasks, methods etc.
2. Results subsection: summarize the data that was collected (qualitative and quantitative)
3. Discussion subsection:
   a. How could you explain the results? What conclusions could you draw wrt. the quality of your solution?
   b. Critically reflect on your work. What are the limitations?
   c. What are the threats to validity?
      Why might it be difficult to generalize your conclusions to other users/systems/environments?

Exercise

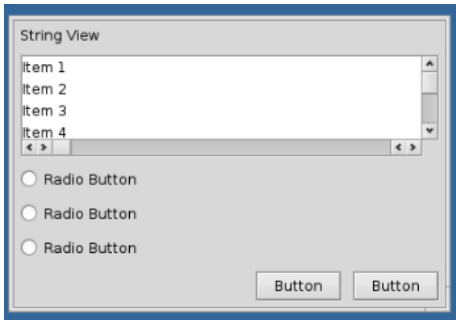
Learn from good and bad example evaluation sections

1. Read them (note: examples are shortened)
2. Identify the following parts:
   a. Methodology: variables, tasks, methods
   b. Results: aggregated data
   c. Discussion: explanations, conclusions, limitations

3. Discuss:
   What are the good and bad points?
   How could it be improved?

Good Evaluation Example 1a

https://www.cs.auckland.ac.nz/~lutteroth/publications/ZeidlerEtAl2013-ALEEvaluation.pdf



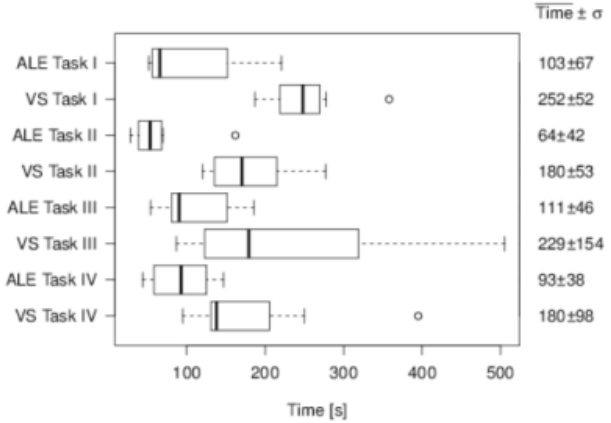
[We investigated] how our approach performs in comparison to a state-of-the art GUI builder. For this comparison we chose the GUI builder in MS Visual Studio 2010 (VS) as a representative for the state of the art, since it was popular at the time [...]

16 participants, mostly software engineering students with experience in GUI development, were asked to perform four GUI creation tasks, each either with ALE or with VS. In each task, they were asked to rebuild a realistic GUI layout from a sample screenshot. Figures 9, 10 and 11 show the four tasks. We measured task completion time as an indicator of efficiency, and used a post-questionnaire to determine participants' preferences.

For both ALE and VS, a training task was given before the respective main tasks to ensure a reasonable amount of training with both tools. To counteract potential learning effects, half the participants were allocated to a group which first performed the training and tasks I and II with ALE, and then the training and tasks III and IV with VS. The second half used the tools in the opposite order.

Good Evaluation Example 1a

https://www.cs.auckland.ac.nz/~lutteroth/publications/ZeidlerEtAl2013-ALEEvaluation.pdf



The measurements were not normally distributed. The medians of ALE and VS were 74 and 188 seconds, respectively. A Wilcoxon signed-rank test identified a significant effect of the GUI builder [...]. Pairwise Wilcoxon signed-rank tests show that ALE was significantly faster than VS for every task [...]. According to the post-questionnaire, 11 of the 16 participants preferred ALE over VS. A separate study is necessary to determine what exactly made ALE perform better than VS [...]

One potential threat to validity is the fact that in VS participants did not use a single gridbag layout, but a nested gridbag layout with a column- and row-span of one. According to observations during the experiment, many participants had difficulties when nesting multiple layouts [...] A possible explanation is that a gridbag layout specification has to be understood more thoroughly upfront, and cannot easily be developed on the fly during the design process as with a constraint-based layout approach.

Good Evaluation Example 2a

https://www.cs.auckland.ac.nz/~lutteroth/publications/PenkarLutterothWeber2013-NavigatingHypertextWithGaze.pdf

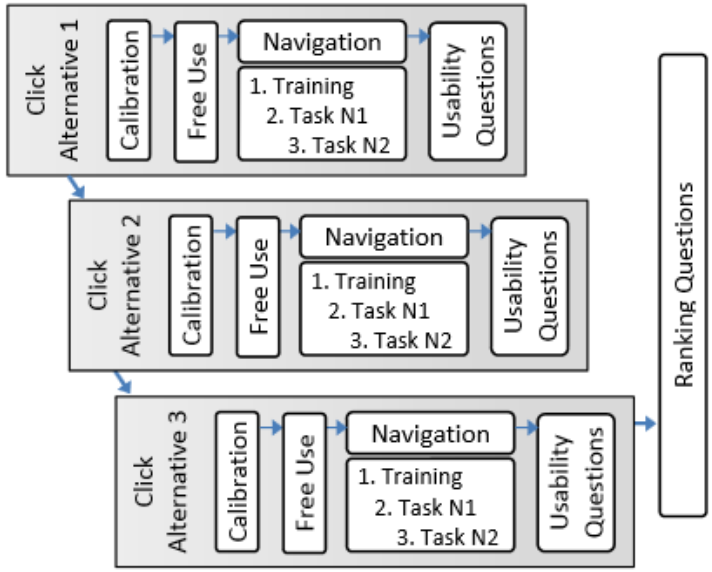


Fig. 8. Experimental procedure for each participant

The experiment was performed using a within-subjects design with “click alternative” as the independent variable, using a nominal scale. The dependent variables measured are “time to click,” “number of incorrect clicks” and “number of failed clicks.” Furthermore, user satisfaction was measured with a questionnaire.

The participants performed the same two navigation tasks using each click alternative (as shown in Fig. 8). [...] Participants were allowed to familiarize themselves with a gaze click alternative for a few minutes before using it in the tasks. For each navigation task, a start page was shown and the participants were told to click four hypertext links one after the other. The first navigation task involved clicking four hyperlinks with comparatively few links in the vicinity, while the second navigation task had a higher hyperlink density in the involved pages. The same two navigation tasks were used for training before performing the actual trials. The order of click alternatives was permuted to mitigate order bias and training effects. [...] After completing the experiment, the participants ranked the click alternatives.

Good Evaluation Example 2b

https://www.cs.auckland.ac.nz/~lutteroth/publications/PenkarLutterothWeber2013-NavigatingHypertextWithGaze.pdf

A total of 19 volunteers performed the experiment, out of which 18 (13 men and 5 women) were successful. The age of the participants varied from [...]

Single Confirm is problematic and not feasible due to the inherent gaze tracker inaccuracy as well as eye jitter. It was difficult for users to activate the confirm button by looking at a particular link. [...] Multiple Confirm, as expected, was not as fast as the mouse. But contrary to expectations, not all participants found the mouse to be the best. Even some of those who judged it best believed that with some more practice they would likely change their ranking [...]

One limitation was the time it took for the web browser to load and display web pages after a click was performed [...] However, as mentioned before, the page load times favored the mouse as participants had time to move the mouse to the expected areas of the screen before the page was fully loaded.

The findings cannot be generalized to all users as most participants had similar demographics [...] Finally, it is not unreasonable to assume that there was a bias in favor of the gaze tracking alternatives due to the novelty factor.

Table 1. Performance of all click alternatives

	Mouse	*Single Confirm*	*Multiple Confirm*
Average time to click a link (sec)	2.1	10.3	7.4
Number of incorrect links clicked	0	4	0
Number of times clicked with mouse	n/a	26	1

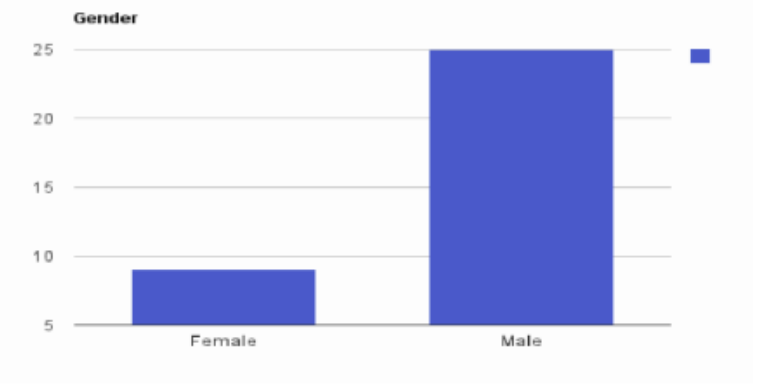
Bad Evaluation Example 1

The study is broken up into three stages. The first is the interviewing stage. The interview's qualitative answers drive the questions and potential answers for a survey to obtain quantitative results. [... The participants] were not told what this paper's research question was in order for them not to insinuate possible answers we may have been looking for [...]

The interview questions are not statistically significant in any way. The results were qualitative and were used to drive the questions and answers for the surveys. For the question "How many hours a day do you use the computer?" Software engineering students tended to have answered between the range of 6 to 16 hours per day. Non-software engineering students have tended to answer around 4 hours per day.

figure 15

Avg satisfaction after customization	Avg satisfaction after customization
6	6.85



Bad Evaluation Example 2

In this study, I try to classify users who utilize computers on a regular basis from three aspects, then create a 3D cube and group the users into categories that are suitable for them [...]
After deciding the framework for classifying users, we need to think about how to use the cube to decide which category a given user should be assigned to. Therefore, a group of open questions need to be designed for an interview, which could provide fundamental information for making a questionnaire [...]
During observational study, a 10-minute interview is also necessary. And we have interviewed 12 participants and asked them a variety of open questions related to their behaviors, preference and reasons [...]
During the process of data collecting, twelve participants submitted their answers and all the answers are valid and effective [...]
Only 16 respondents submitted their answers for the questionnaire, so there was limited information for me to analyze the result. In my opinion, this may have a bad effect on the validness of data for the questionnaire, which means some of my conclusion may violate the fact in reality [...]

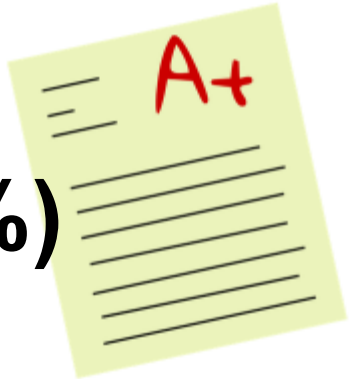
Bad Evaluation Example 3

In this Empirical study first a design and script were prepared and action plan was finalized. Then, we utilized a pre and a post questionnaire along with a formal session sitting of 40 minutes in a quiet room, so as to avoid any confounding variable affecting the study. Before this, emails and fliers were put up to attract participants with a chocolate, so as to reduce any desirability bias.

The volunteering participants were persuaded for an appointment at a convenient time, and on arrival they were explained about the scope of this study. After their approval to carry on, a pre-questionnaire for recording demographics and relevant experience was asked to be filled in. The first 8 participants were given the training on the tasks to be completed using Gridbag layout first and for last 7 the ALM layout based methodology was explained to carry out the tasks so as to reduce the order bias [...]

The results were quite close considering ‘ease to understand’ and ‘ease to explain parameters’, showing that the facilitators were quite impressive in transferring the knowledge and participants accompanied with their experience had now got hold of both the topics but when it came to accuracy...

Assignment: Write Design & Implementation Sections (2.5%)



Write a design and an implementation section for your project (~3-7 pages double-column)

- Individual submission, no group work, worth 2.5%
- Solutions can be hypothetical where necessary: imagine your project is over and was successful
- Be professional: try to imitate well-written papers
- Use LaTeX and BibTex, e.g. https://www.writelatex.com/

Submit PDF by Sunday 28/9 7pm to assignment dropbox: https://adb.auckland.ac.nz

All the best :-)