Investigating longitudinal approaches for pointing device evaluation

Jens Gerken
University of Konstanz
HCI Group
Box D-73
78457 Konstanz, Germany
jens.gerken@uni-konstanz.de

Hans-Joachim Bieg
University of Konstanz
HCI Group
Box D-73
78457 Konstanz, Germany
hans-joachim.bieg@uni-konstanz.de

Harald Reiterer
University of Konstanz
HCI Group
Box D-73
78457 Konstanz, Germany
harald.reiterer@uni-konstanz.de

CHI 2009, April 4 – April 9, 2009, Boston, MA, USA
ACM 978-1-60558-247-4/08/04.

Abstract
In this paper we present our experiences with longitudinal study designs for pointing device evaluation. In this domain, analyzing learning is currently the main reason for applying longitudinal designs. We will shortly discuss related research questions and outline two case studies in which we used different approaches to address this issue.

Keywords
Pointing device, laserpointer, longitudinal, evaluation, retention task, transfer task

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces – Input devices and strategies, Interaction styles, Evaluation/methodology.

Introduction
Input device evaluation in general follows a very rigor experimental approach. This is especially the case for pointing devices where the procedure is highly standardized in terms of tasks (uni- or multi-directional tapping tasks) and experimental designs. Card et al. 1978 first showed that Fitts' Law could be used to predict pointing performance [2]. Fitts discovered in 1954 a linear relation between movement time (time to point onto a target) and the difficulty of the task (depending on the target size and the amplitude between the starting position and the target) [3]. Since

then, the applicability of this regularity has been validated several times and the ISO 9241-9 reflects this in an appendix giving recommendations on the experimental design and procedure for the evaluation of pointing devices. As a result, most experiments in this domain are somehow based on or related to Fitts' pointing paradigm. As opposed to system usability evaluation, longitudinal approaches have a long tradition in this domain. The cited paper by Card et al. already incorporated a longitudinal design to assess learning effects of different pointing devices. Since the idea behind Fitts' Law is to assess the information capacity (measured in bits/s) of a device, it is necessary that participants in an experiment are able to use a device to its full performance potential. However, this requires a skill set which might be unfamiliar to the users. While the handling of some devices might be controllable after a few minutes of training, others require more extensive practice, resulting in the design of longitudinal studies to assess their performance. Nevertheless, most experiments still are single session, cross-sectional designs, which means that learning is in many cases not analyzed at all [e.g. 6]. Reasons are manifold, the most obvious being of course the amount of time and effort needed for longitudinal studies. Besides, in a strictly laboratory setting, longitudinal designs inherit some imponderables. A decline in motivation among the participants may severely influence performance and lead to wrong conclusions. Fatigue can be a factor when sessions get too long and the time span between sessions is still a matter of discussion. Besides, addressing learning in an experiment is not a straightforward procedure as well. For example one approach is to set an expert criterion and measure how long it takes participants to reach this [e.g. 2]. The problem is on the one hand to define this criterion in the first place and on the other hand some users might reach it earlier, some later, and some probably never will, which leads to quite an unpredictable amount of time needed. Another approach is to set the timeframe for the longitudinal study a priori and analyze learning post-hoc. This is done by using statistical procedures like a Helmert contrast analysis or simple pair-wise comparisons (e.g. via t-tests) of different sessions [5]. In any case, one will try to find the point in time when there is no more significant learning. Again, the exact way of defining this point is open to scientific discussion and different ways of doing so exist.

Learning normally follows a power function, which means a higher rate of learning in the beginning and a flattening of the curve over time [2]. When comparing different devices, analyzing the power function can also be an appropriate way to analyze whether one device outperforms another (probably already known) device only after some time. It also can be used to see whether further testing could be useful, since one device is about to catch up on the other(s). In the following, we will present two case studies to illustrate our current approaches to address learning by using longitudinal designs for pointing device evaluation. Both of them investigate the usability of a laserpointer as an input device [4].

Study 1 – analyzing laserpointer practice

In our first study, we were mainly exploring two questions: 1) Does using a laserpointer follow a learning curve and 2) Is such learning via practice permanent and independent from the experimental task? More details regarding this study can be found in [1].

Experimental Design and Procedure

We selected six subjects to use the laserpointer on five consecutive days for 30-45 minutes on each day. The task followed a discrete multi-directional tapping paradigm but was enhanced with a feedback component to keep users' motivation high, similar to the study by Card et al. 1978 [2]. Each session consisted of six blocks (à 126 trials) and 3x3 different amplitudes x target size combinations, totaling in 756 trials per session. The first and last session were accompanied with four additional blocks of a continuous one-directional tapping task. This served two additional purposes. First, two of these blocks were performed with the laserpointer and were used as a transfer task (marked as OL in figure 1). Thereby we wanted to see whether practice transfers to a different task and can thus be ascribed to learning the device and not just the task at hand. Second, the other two blocks were performed with a mouse (OM). We assumed that the performance between the first and the last session would not differ for the mouse since practicing the experimental task should not have an effect on the mouse performance in the one-directional transfer task.

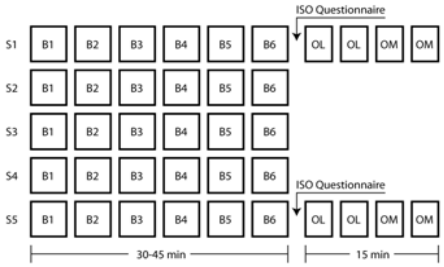

Figure 1: Experimental Design and Procedure

Analysis and Results

We used post-hoc Helmert contrast analysis to assess the point in time when participants were fully trained. This procedure compares the performance of one session with the mean of all following sessions. Compared to simple pairwise comparisons, this procedure better takes into account the whole learning process. Using pairwise comparisons outlier sessions (e.g. having a bad day) have a stronger effect and make it difficult to interpret the results. Besides, doing a complete set of pairwise comparisons requires an alpha level adjustment (e.g. Bonferroni), resulting in a probably overly conservative interpretation – and thus concluding too early that learning has been "completed". In our case, we could see an increase of nearly 1 bit/s in the first session alone – but another 1 bit/s until the peak was reached in session four. Helmert contrast analysis revealed that the improvement in performance was significant up to session three. In figure 2 a typical phenomenon is visible – a performance drop at the beginning of nearly each session. It is a matter of discussion whether it is useful to exclude the first trials of each day to address this issue.

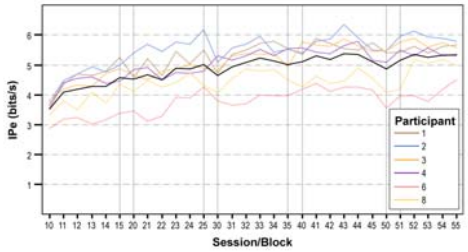

Figure 2: Performance for each participant over time

Analyzing the transfer task (using an ANOVA) revealed that participants improved their performance significantly ($p<0.01$), although the absolute amount dropped to 0.37 bits/s, indicating that quite some performance got lost due to the different task - although the basic task is still the same. Targets are simply in a one-dimensional horizontal plane instead of the two dimensions in the multi-directional task and furthermore a continuous tapping paradigm was applied.

Study 2 – Test-Retest Design

Since doing such a longitudinal study requires time and reliable participants, both of them being rather rare goods, we recently followed a different approach.

Experimental design and procedure

Instead of different sessions we included more trials in one single session experiment with 32 participants. Each of them concluded 1260 trials, divided into seven blocks. In addition, twelve of our participants came back into the lab about 50 days later to perform a retention session, allowing us to assess the permanence of the performance increase. To enhance the efficiency we combined the retention session with a different second experiment. We furthermore added twelve “freshman” who also performed the retention task, although this being the first time for them using the laserpointer. This allowed us to have an additional control group. For this second experiment, all participants had to perform another task, which was again similar to the transfer task. However, it was designed in a more difficult mode with smaller targets, which we accordingly used as a transfer task for this analysis.

Analysis and ResuIts

The first experiment with 32 participants was again analyzed by using Helmert contrasts, revealing that learning was still significant up to the very last block of trials ($p=0.024$). Participants improved about 0.78 bits/s or 26% (2.99 compared to 3.77 bits/s). This means that we weren’t successful in incorporating enough trials in one session to address the learning issue adequately. Since our session already took participants about two hours, it becomes clear that at least one additional session is needed.

For the analysis of the retention task we only looked at those twelve participants (minus one, who was excluded based on extremely high error rates). Here we can see a similar improvement from 3.0 to 3.67 bits/s up to the seventh block of the first experiment. After the 50 days time span the performance dropped again to 3.5 bits/s in the retention task, equaling roughly the performance of the fifth block in the first experiment. However, variances during the retention task went up (0.66 bits/s compared to 0.4-0.56 in the first experiment). With regard of the rather small sample size it is not surprising that pairwise comparisons show that the performance difference to the first experiment is only significant when compared to the very first block of trials ($p=0.00$). Our control group of novice users reached 3.0 bits/s – a difference compared to the retention group of about 0.5 bits/s which turned out to be slightly not significant ($p=0.069$). Interestingly, this difference vanishes in the second part of this experiment. During that, the same kind of task (multi-directional tapping paradigm) was used but with smaller (=more difficult) targets. In this setting, both the experienced retention group as well as the novice group performed about equal ($p=0.346$). This might

indicate that learning is not only dependent of the kind of task but also the difficulty level.

Conclusion

We think longitudinal designs are a must for pointing device evaluations. Learning or practice effects clearly are not easy to come by and only longitudinal designs provide the flexibility to distinguish between learning of the device and the task. However, more research is needed regarding useful transfer and retention tasks that are able to address this issue. Besides, other influence factors such as the motivation of users have to be considered. Since these kinds of experiments rely on very basic tasks, they are very dependent on a controlled environment. Therefore, future research should also provide insight how to evaluate pointing devices in the field using longitudinal designs.

Acknowledgements

We thank Martin Rieger and Stefan Dierdorf for their help in planning & conducting the case studies. This work was supported by the DFG Research Training Group GK-1042 ”Explorative Analysis and Visualization of Large Information Spaces”.

Citations

[1] Bieg, H.-J. (2008). Laserpointer and eye gaze interaction - design and evaluation. Master’s thesis, University of Konstanz, 2008. http://hci.uni-konstanz.de/intehrdis/Bieg2008.pdf

[2] Card, S. K., English, W. K., & Burr, B. J. (1978). Evaluation of Mouse, Rate-Controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT. *Ergonomics*, 21(8), 601-613.

[3] Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381-391.

[4] König, W.A.,Bieg,H.-J., Schmidt, T., Reiterer, H. (2007). Position-independent interaction for large high-resolution displays. In *Proc. IHCI’07*, 2007.

[5] MacKenzie, I. S., & Oniszczak, A. (1998). A comparison of three selection techniques for touchpads. CHI 1998, p336 -343.

[6] Myers, B. A., Bhatnagar, R., Nichols, J., Peck, C. H., Kong, D., Miller, R., et al. (2002). Interacting at a Distance: Measuring the Performance of Laser Pointers and Other Devices. CHI 2002.