
Enhancing Input Device Evaluation: Longitudinal Approaches

Jens Gerken

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

Hans-Joachim Bieg

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

Stefan Dierdorf

HCI Group, University of Konstanz
Box D-73
78457 Konstanz, Germany
stefan.dierdorf@uni-konstanz.de

Harald Reiterer

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

Abstract

In this paper we present our experiences with longitudinal study designs for input 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. Finally, we will point out future research tasks in the context of longitudinal evaluation methods.

Keywords

Pointing device, laser-pointer, longitudinal data, evaluation, retention task, transfer task

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces – Input devices and strategies, Evaluation/methodology.

Introduction

In general, input device evaluation follows a rather strict experimental approach. This is especially the case with the evaluation of pointing devices which relies on a highly standardized procedure in terms of tasks (uni- or multi-directional tapping tasks) and experimental designs. This goes back to Card et al. 1978 who were the first to show that Fitts' Law could be used to predict pointing performance [2]. In 1954 Fitts discovered 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 also manifests in the appendix to the ISO 9241-9 which provides recommendations on the experimental design and procedure, accordingly. As a result, most experiments in this domain are somehow based on or related to Fitts' pointing paradigm.

Longitudinal Evaluation Studies

The social sciences have served quite often as the methodological basis for human-computer evaluation methods. Within this field collecting and analyzing longitudinal data has emerged as an important and indispensable research approach during the last 30 years. Longitudinal data allows researchers to study how the perceived quality of human relationships with and without children changes over the years. In the field of market research it can reveal how people change their level of consumption of certain products. In research literature this is apparent in an increasing number of publications in this field [11, p. vii]. Longitudinal data could be defined as follows: "longitudinal data present information about what happened to a set of research units [in our case the participants of a study] during a series of time points. In contrast, cross-sectional data refer to the situation at one particular point in time." [11, p.1]. This definition also shows that longitudinal data could be obtained with one single retrospective study, where participants are asked about e.g. their attitudes or behaviour at several distinct time points in the past. Furthermore, although being quite common, it is not necessary to obtain longitudinal data in the field. A set of laboratory based experiments, inviting the same

participants again and again, also qualifies as such and can have certain benefits compared to a cross-sectional study as well as compared to a longitudinal field study.

In human-computer studies, longitudinal data collection is still the exception to the rule but it seems that during the last few years the need for such research methods has constantly grown¹. Besides, several researchers are explicitly stating the benefit that could be derived from such methods. Gonzáles and Kobsa [4] for example state that these methods "are needed to reveal the ways in which users would integrate information visualization into their current software infrastructures and their work routines for data analysis and reporting". In [10] Saraiya et al. suggest that "it would be very valuable to conduct a longitudinal study that records each and every finding of the users over a longer period of time to see how visualization tools influence knowledge acquisition". Kjeldskov et al. [5] analyzed how the usability of a patient record system was perceived over time and concluded that "more longitudinal studies must be conducted into the usability of interactive systems over time, focusing on qualitative characteristics of usability problems".

Longitudinal Studies in Input Device Evaluation

As opposed to system usability evaluation, longitudinal approaches have a long tradition in this domain. The cited paper by Card et al. [2] already incorporated a longitudinal design to assess learning effects of different pointing devices in 1978. In 1999 MacKenzie &

¹ See for example:
<http://www.usabilityprofessionals.org/conference/2005overview.html> &
<http://longitudinalusability.wikispaces.com/CHI+SIG+2007+Summary+and+Transcript>

Zhang [8], while comparing an optimized keyboard layout with the traditional QWERTY standard, stated that “users who bring desktop computing experience to mobile computing may fare poorly on a non-QWERTY layout – at least initially. Thus, longitudinal empirical testing is important.” 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. However, this might require a skill set which is unfamiliar to the users. While the handling of some devices might be easily acquired 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. 9]. Reasons are manifold, the most obvious being of course the amount of time and effort needed for longitudinal studies. Besides, in a strict 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. Several approaches are used to gain an insight into learning. One possibility is to set an expert criterion and measure how long it takes participants to reach this [e.g. 2]. While it is difficult to define this criterion in the first place, some users might reach it earlier, some later, and some probably never will. This makes it difficult to predict the amount of time needed. Another approach is to set the timeframe for the longitudinal study a priori and



Figure 1: Using the laser-pointer in a multi-directional tapping task in front of large high-resolution display

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 [7]. 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 usually follows a power function, which means that people learn at a higher rate in the beginning while the curve bottoms out 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).

Case Studies

In the following, we will present two case studies which investigate the long-term usability of a laser pointer as an input device [6]. Within the scope of this paper, we will focus on the methodological approaches and related research questions.

Study 1 – analyzing laser pointer practice

In our first study (see [1] for more details), we were mainly exploring two aspects: 1) does learning to use a laser pointer follow a power function? 2) How can we analyze whether learning transfers to another task?

EXPERIMENTAL DESIGN AND PROCEDURE

Probably the most straightforward longitudinal design for laboratory based experiments is to repeat the experimental session several times. We selected six

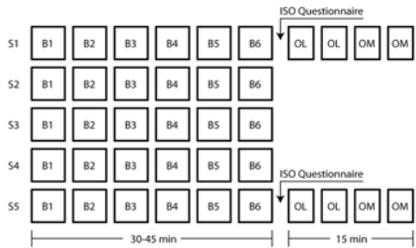


Figure 2: Experimental Design Study 1

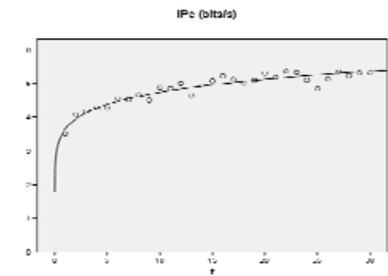


Figure 3: Power function applied to laser-pointer learning

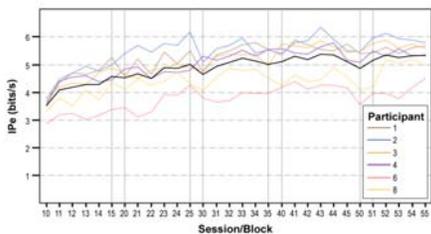


Figure 4: Performance for each participant over time

subjects to use the laser pointer on five consecutive days for 30-45 minutes each day. The task followed a discrete multi-directional tapping paradigm but was enhanced with a feedback component to keep users' motivation high (see figure 1, similar to the study by Card et al. 1978 [2]). Each session consisted of 756 trials in total per participant. To investigate the question of performance transfer, one has to decide on an appropriate transfer task for assessment. Although using a complex, realistic task which furthermore increases the ecologic validity might seem natural such a task actually reduces the effect of practice. The user has to learn this new task at first, which makes it again difficult to distinguish between task difficulty and device difficulty. Therefore, we decided to use a rather similar and easy to learn task: a continuous one-directional tapping task. While two blocks were performed with the laser pointer and were used as a transfer task (marked as OL in figure 2), two additional 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. If this hypothesis would not hold, one could assume that the task was too similar in comparison with the experimental task.

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 pair-wise comparisons, this procedure is more suited to take into account the whole learning process. When using pair-wise comparisons

outlier sessions (e.g. having a bad day) have a stronger effect and make it difficult to interpret the results (e.g. interpreting a phenomenon such as no learning between session 3 and 4 but then again between 4 and 5). Besides, doing a complete set of pair-wise 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. Applying the power law of practice reveals a very good fit of the data with $R^2=0.917$ (for IP_e , $\beta = 3.6$, $\alpha = 0.114$, see figure 3). In figure 4 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. Analyzing the transfer task (using an ANOVA) revealed that participants improved their performance significantly (3.75 to 4.12, $p<0.01$), while the mouse performance remained stable (4.54 compared with 4.58 bits/s). This indicates that quite some performance got lost due to the different task; even though we used a similar and very easy task. Furthermore, the transfer task reveals the impact of the longitudinal design. In a purely cross-sectional design, one might come to the conclusion that a much higher difference between mouse and laser-pointer does exist compared to a more realistic test setting including practice (3.75 compared with 4.54 bits/s vs. 4.12 compared with 4.58 bits/s after one week). For future research it is important to point out that the transfer to different kinds of tasks has to be investigated in more detail.

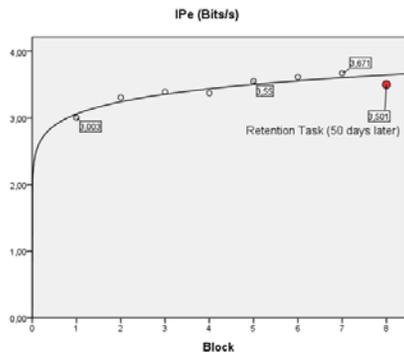


Figure 5: Power function applied to cross-sectional design for laser-pointer learning

Study 2 – Test-Retest Design

In this second study, we explored the following questions: 1) is it possible to assess learning appropriately in a cross-sectional design? 2) How can we investigate the permanence of learning? 3) How can we investigate transfer to a different task more in detail?

EXPERIMENTAL DESIGN AND PROCEDURE

We relied on a test-retest design to address these issues. First we conducted an experiment with 32 participants. Each of them concluded 1260 trials, divided into seven blocks, which allowed us to analyze the learning similar to the first study but in one single session. 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. In case of the laser-pointer this time interval represents an infrequent usage scenario which we assumed would be most likely when using the device for presentations or collaborative work in front of a large high-resolution display. Future research has to investigate this aspect more systematically, of course. In this second experiment, we also invited twelve novice users as a control group to further validate possible learning. In addition, all participants of this second experiment had to perform another task. Contrary to the first study presented, we used exactly the same task, but with smaller targets making the task more difficult. This allowed us to investigate the transfer of learning more systematically than in study 1.

ANALYSIS AND RESULTS

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 (2.99 compared with 3.77 bits/s). Since our session already took participants about two hours we conclude that at least one additional session would be necessary to stabilize learning, again stressing the need for longitudinal designs. Nevertheless, we could obtain a similar power function compared to the first study with a high fit of $R^2=0.972$ and a very similar slope (IPe, $\beta = 3.022$, $\alpha = 0.111$, see figure 5). Our twelve participants who also participated in the retention experiment showed a similar improvement from 3.0 to 3.67 bits/s up to the seventh block of the first experiment. After the 50 days interval the performance dropped 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). Consequently, pair-wise comparisons show that the performance difference to the first experiment is only significant when compared with the very first block of trials ($p=0.00$). Our control group of novice users reached 3.0 bits/s – a difference compared with the retention group of about 0.5 bits/s which turned out to be slightly not significant at the 0.05 level ($p=0.069$). Interestingly, this difference vanishes in the second, more difficult part of this experiment. In this setting, both the experienced retention group as well as the novice group performed about equal ($p=0.346$). This indicates that learning is not only dependent on the kind of task but also the level of difficulty.

Conclusion & Future Work

We have presented two different approaches to address learning and practice in pointing device evaluations. We

think longitudinal designs are a must in such studies. 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. Longitudinal designs could also give more insight into usage strategies and how these might change over time. Besides, other influencing factors such as the motivation of users have to be considered. Since input device experiments rely on very basic tasks, they are currently very dependent on a controlled environment. Thereby, they often lack ecologic validity and so future research should also investigate how to evaluate input devices in the field using longitudinal designs (e.g. in combination with diaries [10]). Finally, the overall goal should be to develop a framework of longitudinal evaluation methods in HCI. While the social sciences have developed a methodological framework for longitudinal research during the last decades, distinguishing between several different approaches and stating which type of research questions demand which kind of approach or method and the appropriate analysis method [e.g. 11], such a framework is still missing for human-computer studies.

Acknowledgements

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] Gonzáles, V. and Kobsa, A (2003). A workplace study of the adoption of information visualization systems. In *Proceedings of I-KNOW'03*, Graz, Austria, p92-102.

[5] Kjeldskov, J., Skov, M. B., Stage, J. (2005). Does time heal?: a longitudinal study of usability. In *Proceedings of the 19th Conference of the Computer-Human interaction*, ACM.

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

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

[8] MacKenzie, I. S. Zhang, S. X. (1999). The design and evaluation of a high-performance soft keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: the CHI Is the Limit*, New York: ACM Press, p. 25-31.

[9] Myers, B. A., Bhatnagar, et al. (2002). Interacting at a Distance: Measuring the Performance of Laser Pointers and Other Devices. CHI 2002.

[10] Saraiya, P. North, C., Duca, K. (2004). An evaluation of microarray visualization tools for biological insight. In *Proc. of IEEE Symposium on Information Visualization*, p 1-8, 2004.

[11] Taris, Toon W. (2000). A Primer in longitudinal data analysis. London: SAGE Publications.