Longitudinal Evaluation Methods in Human-Computer Studies and Visual Analytics

Jens Gerken, Peter Bak, and Harald Reiterer

Abstract— In Human-Computer studies and visual analytics, the majority of the evaluation methods applied, collect data at a single point in time, in form of cross-sectional data. In several studies [e.g. 8] numerous visualization tools were evaluated in controlled experiments. Although the experiments discovered valuable findings, certain drawbacks of the research method were expressed. The time constraints of one-time experiments reduce the amount of training which can be given to the participants. Furthermore, when the studies tried to measure the insight derived from the visualization tools the time constraints didn't allow observing how these insights develop over time or their interdependency. Further problems of cross-sectional studies are well known, like the selection of appropriate tasks, the mostly extrinsic motivation of the participants, the influence of a laboratory environment compared to a realistic work setting and whether a visualization tool does meet the work requirements in the long run. In this position paper we argue for applying longitudinal research methods in human-computer studies as an extension to cross-sectional studies and present a first approach towards a methodological research framework. We suggest a set of research questions and performance measures that would be benefical for extending cross-sectional studies with longitudinal ones. We also describe in two case studies, in which only cross-sectional research methods were used, how they can improved by longitudinal methods.

Index Terms—longitudinal, long-term, methodology, research framework, experiment, visual analytics, evaluation

1 Introduction

Within the social sciences, which have served quite often as the methodological basis for human-computer evaluation methods, collecting and analyzing longitudinal data has emerged during the last 30 years as an important and indispensable research approach. So for example, longitudinal data allows researchers to study how the perceived quality of human relationships with and without children changes over the years [3] or in market research to study how people change their level of consumption of certain products. In research literature this is expressed by an increasing number of publications in this field [11, p. vii]. Longitudinal data could be defined as follows: "basically, 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 it is 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 can have certain benefits compared to a cross-sectional study as well as compared to a longitudinal field study.

2 Longitudinal Research in Human-Computer Studies

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 grown. So for example the UPA conference in 2005 held a seminar on this topic [5] and also on this year's CHI conference a special interest group (SIG) was founded dealing with longitudinal research [11]. 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 [8] 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. [6] 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". MacKenzie & Zhang [7] already stated in 1999 when comparing a optimized keyboard layout with the traditional QWERTY standard "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."

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. The result is that although a couple of longitudinal studies exist, they mainly don't follow a certain research methodology and therefore vary quite a bit in their respective designs. So for example Saraiya et al. [9] used diaries in which insights and screenshots were stored by the participants themselves as methodological basis. The goal was to get a better picture of the whole visual analytics process. On the other hand Shneiderman and Plaisant [10] present an approach that relies on many different data collection methods such as interviews, observations, and logging and is therefore much more complex for both the researcher and the participant. MacKenzie & Zhang [7] rely on a series of laboratory based studies to analyze how much training is necessary for a new soft-keyboard layout to be superior to the QWERTY standard. Kjeldskov et al.[6] also rely on two laboratory

- *Jens Gerken - University of Konstanz, E-Mail: jens.gerken@uni-konstanz.de.*
- *Peter Bak - University of Konstanz, E-Mail: bak@dbvis.inf.uni-konstanz.de.*
- *Harald Reiterer - University of Konstanz, E-Mail: harald.reiterer@uni-konstanz.de.*

studies to analyze whether usability problems might "disappear" after 15 month of system usage. One common aspect of most of these studies is the lack of reasons stated that explain why a certain longitudinal methodology was applied. One rare exception is the work of Shneiderman & Plaisant cited above, which presents an adaptation of multi-dimensional in-depth long-term case-studies (MILCs), initially developed within the creativity research domain, to information visualization. They describe it as a new paradigm for evaluation and present quite elaborated guidelines for conducting such studies. We think that it could very well serve as a basis to harmonize longitudinal research that especially aims to obtain qualitative data of tool usage, e.g. tries to analyze how insight is derived from data over time with the help of visualization tools. Nevertheless we think that the approach taken has two main disadvantages when thinking about a research framework for longitudinal studies. First in our understanding, longitudinal studies can be field studies as well as laboratory studies. The approach of MILCs however only covers field studies. Second we think that the methodology suggested might be too costly for both the researchers and the participants in some cases. More tailor made and thinner methods should help in minimizing a methodological overkill.

3 RESEARCH METHODOLOGY

Therefore, we are taking a different approach within this paper. We think that a research framework for longitudinal studies is needed covering all shapes of longitudinal methods and assisting in applying these methods in practice. In order to develop such a framework, first research questions should be defined that are either not well addressed in today's cross-sectional studies or even not considered at all. Next to the research questions, measurements and data gathering techniques needed for each of these should be defined as well as methods for data analysis. From there on, methodological procedures should be defined such as whether a lab-based or field-based approach is better suited for the research question at hand. Obviously such an approach heavily relies on case studies where researches not only apply longitudinal methods but also report about methodological advantages and disadvantages observed and lessons learned.

In the following we want to add to such a research framework by defining a simple research methodology including research questions, measurements and data analysis suggestions, based on the literature cited above and own considerations.

3.1 Research Questions

We suggest the following research questions, in which longitudinal studies would be indispensable:

- *How is users' domain knowledge expending over time as a function of different tools or visualizations?*
  Domain knowledge can't be obtained in the time constraints of a cross-sectional design. In longitudinal field studies it could be observed how different tools might facilitate domain knowledge acquisition and others don't.

- *How is the usability of the system perceived over time?*
  Learning how to use the system is an important quality aspect of visual analytics tools. If the time it takes users to master the tool is too long, they might rely on their traditional methods, despite them probably being less effective. Different interaction design approaches could also result in different learning times. It is most of the time impossible to investigate learning behaviour in a cross-sectional study because of time constraints, which is why longitudinal studies, in this case probably within the lab, could help answering this research question.

  Another question is whether usability problems might disappear over time, meaning that users find workarounds to avoid them. Analyzing the nature of such problems in longitudinal studies might help in identifying "false positives" in future cross-sectional studies.

- *How do users define the level of interestingness of their findings?*
  Measuring individual interests is still difficult. Moreover, individuals develop and change their interest over time. Longitudinal field studies could help observing the users' definition of interestingness of a task or finding as a function of changes in their personal preferences and extension of their domain knowledge.

- *How is performance influenced over time by users' individual differences?*
  Individual differences can influence users' system performance. The goal should be to find possibilities to accommodate for individual differences. However, within cross-sectional designs the influence of individual differences can be overlapped for example by learning issues. Therefore longitudinal lab studies are needed to further investigate the influence of individual differences in order to apply them appropriately in the design of interactive visualizations.

- *How is the visual information extraction process – expressed by tool usage – optimized over time?*
  Analyzing logging files of cross-sectional studies often leads to more questions than answers. So for example interaction patterns discovered in a lab might be misleading, since the tool could be used quite differently in the real world. Therefore longitudinal field studies should be used to investigate this research question.

3.2 Measurements

In order to answer these questions the research framework has to provide quantitative and qualitative measurements. We suggest putting more effort into analyzing the potential of logging the users' system-usage behaviour both in the field and during longitudinal lab studies taking all ethical considerations into account. Although logging is quite common, especially in the Web, more effort should be taken to formalize both the logging and the analysis of the data. In our understanding such a log file should record sequence and timing of tool use, transaction matrix of tools and the time spent between tools used in order to assess the effectiveness of different visualizations and users' strategy to extract information. In combination with traditional lab-based performance and effectiveness measures as well as qualitative data obtained through interviews, observations and diary techniques (the latter two mainly in the field) a more complete picture of the visual analytics process could be obtained. Furthermore, retrospective data gathering methods from social sciences, such as special standardized questionnaires or interview techniques should be further tested for their applicability to our domain. Some measurements such as learning time could be derived from existing measures such as session durations for a series of lab studies [see e.g. 7]. As an equivalent to "task completion time"-measurement in the lab one could measure the time it took a participant to find the first insight when using real data in a longitudinal field study.

3.3 Analysis methods

In order to analyze the findings different methods of data analysis might be necessary to be used. Traditional methods of ANOVA and regression might not be sufficient with respect to the amount of data and non-numeric data types. Especially the extraction of e.g. strategies of tool usage from log files is quite a challenge. We think that sequence analysis algorithms from the field of data mining could

be tremendously helpful in finding certain patterns in tool usage such as interaction and navigation patterns. Besides, methods of clustering (of users and tool use), classification (of strategies), and text mining could provide the expected insights. The results achieved in the longitudinal studies could serve as an input for further validation in traditional cross-sectional studies. In these studies improvements regarding the usability of the system could be tested regarding shorted search paths, better sequences and timing of tools.

4 Case Studies

In the following we want to illustrate some of the aspects mentioned above such as research questions and measurements on the basis of two cross-sectional visual search studies. We hypothetically extend those studies to longitudinal studies and discuss possible benefits.

The study described in [1] aimed at assessing the degree to which users can effectively use visual querying to perform a data filtering task. In order to perform such tasks efficiently, users need to identify the most effective ways to filter the data and to avoid biases from irrelevant properties of the displayed information. Thereby users may differ in their ability to do so, which may partly depend on their cognitive styles. Participants with higher rationality scores performed more accurately and seemed to have been more aware of the relative efficiency of filtering with different segment types. Participants with a more experiential style initially relied more on browsing than on filtering, as indicated by slower performance and by fewer selections of segments for filtering. However, experiential cognitive style seemed to play a role only at the beginning of the experiment and its effect lessened towards the end. For rationality however, which was the main predictor of performance, the effects continued to exist throughout the experiment [1].

However, for the moment we lack on the knowledge whether the effects of cognitive style can be moderated through learning. More specifically the current research failed to show whether cognitive style is able to predict performance for expert users and not only for novices. Furthermore, in the described setting user behaviour was analyzed for a duration of 30 minutes, therefore the question whether users change their search patterns could not be answered. Longitudinal studies would be beneficial to address these questions and issues, and therefore would allow researchers to assess the usefulness and efficiency of their tools. We would suggest to measure users' preferences for different pattern types and how it is changing over time. We would also hypothesize that cognitive style (referring to the results on rational cognitive style) may be a good predictor of performance for novice users, but there is no evidence that the same results can be achieved when the same users achieved a wide reaching experience with the environment and the tasks.

The study also reported that users preferred to search the graphs sequentially from the left to the right. This finding however was not strong enough to bias performance. This kind of biases might be better investigated in much more complex environments where search paths need to be acquired and remembered for a longer period of time and by using real-world data sets. Our suggestion is that these types of research questions and hypotheses are better investigated in a longitudinal design, which would allow users to develop their individual search behaviour.

In [2] users had to perform visual search tasks using a star-field display on a mobile device. The experiment compared two different presentation techniques, one using solely a zoomable star-field interface while the second one combined this detail view with an additional overview that furthermore offered several interaction possibilities. However as a result of the design of the study as a cross-sectional study, the results led to new questions. Users performed significantly better using the detail-only interface compared to the overview-detail interface in terms of task completion time. This is insofar surprising since the overview offered several possibilities to speed up the tasks. However users seemed to have difficulties in deciding how to use these possibilities – the straight forward approach of the detail-only interface therefore caused less thinking. A longitudinal lab-based study could help to analyze whether these negative effects are just a matter of training and if a cross-over point exists, where the overview+detail interface performs better. In a similar way the preference voting, which was 13:11 in favour of the detail-only interface, might change over time, when users adapt to the additional features. During a longitudinal field study, one could analyze with logging techniques how users switch between overview and detail view and if there are certain interaction patterns that could be better supported. Seldom used functions could also be identified and therefore improved in future designs.

5 Conclusion

In this paper we argued for applying longitudinal evaluation methods as an extension to cross-sectional studies. Currently cross-sectional research methods are widely applied in the field of human-computer studies and visual analytics. However, these kinds of methods are unable to provide a comprehensive picture of users' performance with complex visualizations regarding effects of learning, domain knowledge or users' task solving strategies. In our understanding one major goal should be to develop a comprehensive research framework for longitudinal evaluation methods. To achieve this, research questions, measurements and data gathering techniques as well as data analysis methods have to be defined. Furthermore, appropriate methodological procedures for longitudinal studies should be specified. Besides studies in our own domain, research work within the social sciences, psychology and market research should be considered.

Drawbacks of longitudinal methods should be considered and analyzed as well, such as the increased effort needed or known problems such as drop out of participants and the serial dependence of measurements.

At the end, such a comprehensive longitudinal research framework could serve as a decision help for researchers and practitioners to choose the appropriate and tailor-made methods and procedures when dealing with a specific research question.

Acknowledgements

This work was supported by the DFG Research Training Group GK-1042 "Explorative Analysis and Visualization of Large Information Spaces"

References

[1] P. Bak and J. Meyer, "The Efficiency of Visual Querying of Temporal Data", submitted to the International Journal of Human Computer Studies, 2007.

[2] T. Büring, J. Gerken, and H. Reiterer, "Usability of overview-supported zooming on small screens with regard to individual differences in spatial ability", in Proceedings of the Working Conference on Advanced Visual interfaces, AVI '06. New York: ACM Press, p. 233-240, 2006

[3] Christine Bleich, "Veränderungen der Paarbeziehungsqualität vor und während der Schwangerschaft sowie nach der Geburt des ersten Kindes", *Übergang zur Elternschaft. Aktuelle Studien zur Bewältigung eines unterschätzten Lebensereignisses,* B. Reichle and H. Werneck, eds., Stuttgart: Enke, p. 167-184, 1999.

[4] V. Gonzáles and A. Kobsa, "A workplace study of the adoption of information visualization systems", Proceedigns of I-KNOW'03: 3rd International Conference on Konwledge Management, Graz, Austria, p92-102, 2003.

[5] Lada Gorlenko, "Long-Term (Longitudinal) Research and User Experience Design", UPA Conference 2005, Advanced Topic Seminar, http://www.usabilityprofessionals.org/conference/2005overview.html, (online at Sept. 2007), 2005.

[6] Kjeldskov, M. B. Skov, and J. Stage, “Does time heal?: a longitudinal study of usability”, In Proceedings of the 19th Conference of the Computer-Human interaction Special interest Group (Chisig) of Australia on Computer-Human interaction: Citizens online: Considerations For Today and the Future, ACM International Conference Proceeding Series, vol. 122. Computer-Human Interaction Special Interest Group (CHISIG) of Australia, Narrabundah, Australia, 1-10. 2005

[7] I. S. MacKenzie and S. X. Zhang, „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, 1999.

[8] P. Saraiya, C. North, and K. Duca, “An evaluation of microarray visualization tools for biological insight”, Proc. of IEEE Symposium on Information Visualization, p 1-8, 2004.

[9] P. Saraiya, C. North, V. Lam, and K. Duca, “An Insight-Based Longitudinal Study of Visual Analytics”. IEEE Transactions on Visualization and Computer Graphics 12, 6, p. 1511-1522, Nov. 2006.

[10] B. Shneiderman and C. Plaisant, “Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies”, in Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods For information Visualization, BELIV '06, New York: ACM Press, p. 1-7, 2006.

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

[12] M. Vaughan and C. Courage, “SIG: capturing longitudinal usability: what really affects user performance over time?”, In CHI '07 Extended Abstracts on Human Factors in Computing Systems, New York: ACM Press, p. 2149-2152, 2007.