Chapter 9
Evaluation of Visualization Systems with Long-term Case Studies

BERNHARD PREIM[1], *Otto-von-Guericke-Universität Magdeburg, Germany*
ALARK JOSHI[2], *University of San Francisco, USA*

Abstract
Long-term case studies are an essential tool to capture the adoption of visualization techniques and systems, their usage patterns, and how these change over time. They may provide essential insights that stimulate further research that is focused on actual user needs. Long-term case studies typically involve a few users that use an interactive system frequently and for a longer time at least a couple of weeks [25]. This intense use of a system is analyzed with various techniques, such as documentation including screenshots (*diary of use*), automated logging protocols, regular interviews, and screen capture. They may elicit cognitive processes related to decision-making and problem solving. This chapter explains the concept of long-term case studies, its variants, and examples describing how it was realized in medical visualization, information visualization, and visual analytics. We characterize this evaluation technique with respect to the stages in a development process to which they fit and with respect to scenarios in data visualization where they are often used. The limitations of long-term case studies are also discussed leading to recommendations when this type of evaluation is useful to accompany visualization research.

9.1 Introduction

Many evaluation concepts have been employed to assess the value of individual visualization techniques or whole visualization systems. Isenberg et al. provided an in-depth analysis of evaluation practice in visualization and considered eight major variants [8]. Regarding their terminology, we focus on *empirical* evaluations that include actual users, instead of hypothetical discussion of usage scenarios or formal analysis based on quality metrics. As Isenberg et al. point out, most of the existing

[1] e-mail: bernhard@isg.cs.uni-magdeburg.de
[2] e-mail: apjoshi@usfca.edu

empirical evaluations relate to user preferences and other usability or user experience criteria but do not focus on how visualization systems are actually used to solve complex problems. The case study type of empirical evaluation that we discuss in this chapter is considered as ”particularly strong form of evaluation for understanding work practices and visual data analysis" [8]. However, case study-based visualizations are rare compared to the substantial portion of application-oriented visualization research in the visualization community. In this chapter, we discuss the potential and current practice of case study-based evaluation in visualization research. We emphasize one aspect of expressive case study reports, namely the long-term character. Today's complex visualization systems may involve longer learning periods and problem-solving activities that require substantial time. Thus, care is necessary to provide enough time for experts to use the system and for visualization researchers to observe and analyze the usage of the system.

Long term case studies have their roots in ethnographic research [4], e.g. in cultural anthropology, where researchers live in a different culture, e.g. in an African tribe, take part in the daily activities and carefully document their first-hand experiences (*diary of use*). “The observer tries as much as possible to be unobtrusive”, ideally not affecting what is being observed [3]. With ethnographic methods, a few researchers gain insight and in-depth experiences over long time. *Field tests* and *workplace studies* are alternative names used in human computer interaction (HCI) [3] [6].

Case study-based methods were introduced in HCI early. A survey by Hughes et al. [7] documents success stories both in academic and commercial settings, where time and budget constraints need to be considered as well. Figure 56 puts field studies in the context of other evaluation methods and highlights that they are particularly realistic, but not very precise.

Putting long-term case studies in the context of empirical evaluation

Long-term case studies are a promising instrument of empirical evaluation and “yields realistic and believable narratives” of real users interacting with a visualization tool [5]. They are motivated by shortcomings of the more frequently used controlled laboratory studies as stated by Shneiderman and Plaisant [25]: “laboratory studies became ever more distant from practical problems and broader goals.” Carpendale adds that the use of very small datasets, students as test subjects, and unrealistic tasks lead to the problem that the results of information visualization evaluations are not believable and actually, that the developed techniques are not adopted [3]. In particular, systems that require substantial learning effort and are intended to be used for complex problem solving or discovery activities cannot be adequately assessed within one or two hours in a lab experiment with well-defined tasks. The simple fact that the evaluation takes place in a lab and not in a realistic work context reduces ecological validity, that is the amount to which the results can be translated to realistic settings.

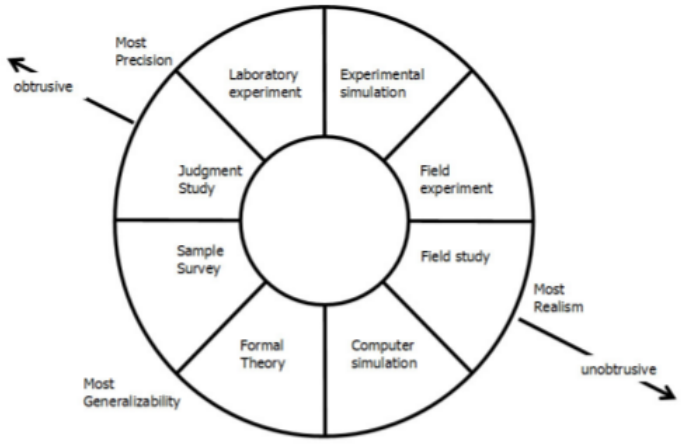

Fig. 56 Field studies, an alternative term for long-term case studies, are unobtrusive and yield realistic observations (From: Carpendale et al. [3]).

Visualizations related to isosurface and volume renderings are often evaluated by means of task-based perceptual experiments typically involving a comparison of methods with respect to shape and depth perception. In the terminology of Carpendale [3], these are referred to as *judgement studies* (recall Fig. 56). Similar to (other) lab experiments, they favor precision or realism. Although valid tasks and methods are available, the evaluations only explain perceptual aspects at a rather low level (see Preim et al. [19] for a survey and Saalfeld et al. [21] for a tutorial-like paper on how to perform such experiments with a focus on medical visualization). In medical visualization for example, the actual purpose is to support advanced diagnostics (Is a muscle infiltrated by a tumor and to what extent?) and treatment planning (Is the patient operable? How much tissue needs to be removed? And how to access the pathology?). For understanding such cognitive activities involving problem solving, decision making, and discovery perception-based experiments are not directly relevant. Moreover, almost all these experiments relate to static rendered images, that is the whole value of interactive exploration, e.g. rotation and clipping, for which 3D visualization techniques are provided, is ignored.

Long-term case studies typically involve a few highly specialized professionals that use a system in their familiar work environment for tasks that are relevant to them, based on data that they have available [2]. The discovery in large scientific, business or finance data, police analysis and medical research based on large and heterogeneous data are examples of such situations. As we will see in this chapter, there are a number of examples how long-term case studies were used for medical visualization, information visualization, and visual analytics applications.

9.2 Goals and Variants of Long-term Case Studies

Ethnographics-inspired evaluations were carried out in human-computer interaction and software engineering as a research method for a deep understanding of processes and the use of interactive products. A deep involvement in the users' activities can provide genuine insight into the processes and daily routines of the users.

These observational methods may be applied early in the development process to analyze current work practices and establish initial requirements [7]. Here, we focus instead on *evaluative ethnography*, that is the evaluation of innovative visualization systems based on a working prototype. Evaluation ethnography includes an assessment of the prototype, the deployment in particular contexts and workflows and the extraction of ideas for redesign – three of the five stages of empirical evaluation as discussed by Lam et al. [16].

9.2.1 Goals

Long-term case study evaluations last at least several weeks and are carefully documented by the users with both verbal notes and screenshots. Regular interviews, logging protocols, screen capture and video analysis may be added to *understand* differences [25]. They may reveal:

- patterns of use, e.g. typical problems as well as actions to tackle them,
- characteristic changes of these patterns over time,
- the social context of system use,
- engagement and motivation,
- the variety of data to be processed and tasks to be solved in practice, and
- unintended usage scenarios.

Such findings may have serious implications for further design which makes these methods appropriate for *formative* evaluation where the major goal is to refine or add requirements for the further development of a system. If for example, some features are not used at all, they may be removed or at least "hidden" in submenus or dialogs that rarely appear and thus do not distract.

If certain usage patterns become obvious, the system may be redesigned to provide *guidance*, e.g. to support the user along a certain analysis path. As a social aspect it may become obvious that domain experts cooperate with others in the analysis of data. As an example in medicine, the analysis of medical image data, is a careful cooperation between radiology technicians and radiologists and the results of the diagnostic report are presented to referring physicians from other medical disciplines. If such a collaborative aspect is identified and analyzed, requirements to directly support cooperation may arise. In fact, early uses of ethnographic methods were already focused on analyzing social aspects in office contexts or air-traffic control [1].

Long-term case studies may also reveal *how* engaged users are, often despite struggling with the system, how motivated they are and how they (and perhaps their colleagues) trust a system. These user experience (UX)-related properties and their changes over time are essential for visualization systems to be used in research and industrial practice. The understanding of actual data and tasks often leads to requirements related to the support of more file formats or related to a better support to convert data.

Additionally, unintended usage scenarios are observed that typically involve creative workarounds to achieve a goal, the system was not meant to be used for. As a consequence, a redesign should directly support these usage scenarios. As an example, Whitaker [28] analysed e-mail use and found that mail systems are not only used for communication (as intended) but also for reminding to activities and as an archive of communication and knowledge.

9.2.2 Multi-dimensional In-depth Long-Term Case Studies

Multi-dimensional In-depth Long-Term Case Studies(MILCs) were introduced as a special long-term evaluation technique particularly for InfoVis. This evaluation concept by Shneiderman and Plaisant [25] was introduced to evaluate creativity support tools. The major goal of MILC evaluations is to ”study the creative activities that users of information visualization systems engage in". *Multi-dimensional* relates to the integrated use observations, interviews, logging protocols. Shneiderman and Plaisant also explain what they consider as long-term: a system use in different stages with a minimum duration of several weeks. The following stages are discriminated:

- the *training stage*, where the users get familiar with the system, optionally a written tutorial to assist independent use,
- an *early use stage* where the users are visited also with the goal of assisting in using the system and identifying smaller problems that may be solved soon, whereas
- in the *mature use stage* the system is no longer altered. Thus, changes in usage patterns in the *mature use stage* are not due to changes in the system and may reflect that usage patterns change over time.
- a *final stage* in which the documentation is summarized and a final review is carried out.

The methods of data collection is the same, in the early and mature use stage. Only the stable system state makes the difference between the two. Not all authors that base their evaluations on the MILC principles follow all recommendations. Valiati et al. [27] for example report on three MILC studies, where they have *not* discriminated between early and mature use stage. The system was not improved at all during the whole study. They employed most of the instruments recommended by Shneiderman and Plaisant [25] but did not provide logging functions.

Shneiderman and Plaisant discriminate basically two types of MILCs:

- a moderate MILC as part of a typical research project, where the early use and mature use stage last approximately four weeks and
- long MILCs that may last up to several years where the evaluation is the core activity of a research project.

Most MILC evaluations are moderate variants. In the examples discussed by Valiati et al. [27], the study duration was between six weeks and four months, 5-8 meetings with users were arranged and the overall time of observing users was between 12 and 18 hours. This example confirms the recommendations of Shneiderman and Plaisant to combine different instruments, such as observation, interviews and thinking aloud (recall [25]). They traced the problems identified in the long-term evaluation to the instruments used to detect them: While some problems were explicitly described by the analysts during interviews, a considerable portion were detected based on observation.

9.3 An Overview of Long Term Evaluations in Visualization Research

In the following, we briefly describe selected examples of long-term case study evaluations. They were chosen, since they rigorously report on goals, preparation, conduction, and analysis. The underlying papers do not introduce a visualization framework, but focus on the evaluation of an already presented system. Thus, the evaluation is not a minor part of a large paper. Among the seven scenarios from Lam et al. [16], they all relate to *visual data analysis*. It seems that long-term case studies are particularly important in this scenario. In the scientific literature, there are more long-term case study evaluations of visualization, but they are described in considerable less detail.

9.3.1 Evaluating the Rank-by-Feature Framework

Seo et al. have developed a comprehensive visual analytics framework that enables the efficient analysis of high-dimensional data [23]. Many metrics (*interestingness measures*) are involved to rank individual features and pairs of features to direct the further analysis to potentially interesting aspects, e.g. features, where the distribution strongly deviates from a normal distribution or pairs of features where a strong linear or quadratic correlation exists. As a general unsupervised learning method hierarchical clustering method with an interactive dendogram visualization is provided. The system was primarily used for analyzing gene expression data and was initially presented along with informal evaluations including feedback from domain experts.

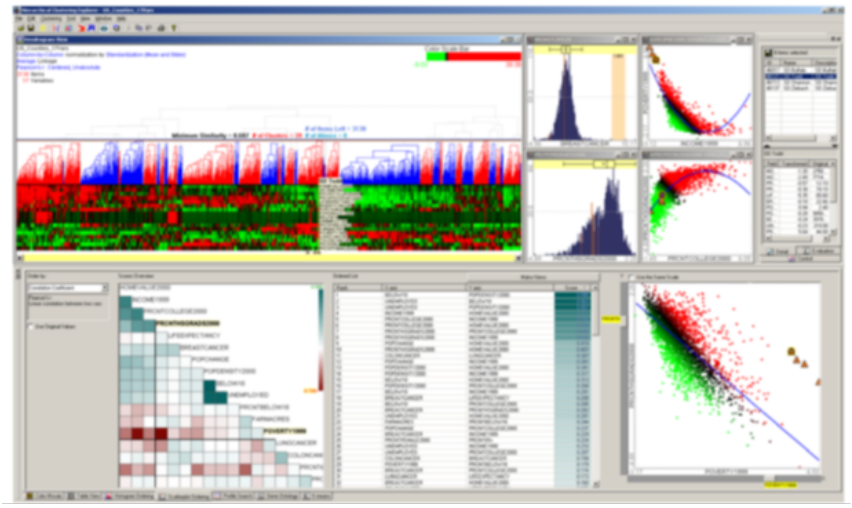

Fig. 57 The Rank-by-feature framework with hierarchical clustering (top left), a matrix view depicting correlations between dimensions (bottom left), an ordered list with most interesting feature combinations (bottom middle) as well as histograms and scatterplots for selected dimensions and combinations thereof (From: [23]).

To get a deeper understanding, if and how the rank- by-feature framework change the researchers exploration process, a MILC evaluation was performed [24]. Six participants were recruited that had used the framework and published scientific results obtained with it. These researchers were from different fields (including statisticians, biologists, metereologists) and were not involved in the design and development of the tool.

9.3.2 Evaluating the Social Action Tool

Few users employed a social network analysis tool with graph-based visualizations and statistics related to graph-based data [18]. The long term case study was performed according to the MILC variant (recall [25]), where the early and mature use stages lasted four weeks. The evaluation was started with a 2 hour training session and a documentation was provided to further support the autonomous use of the system.

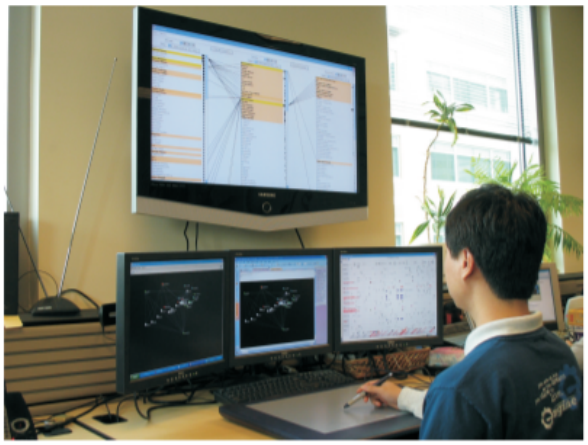

Fig. 58 The Jigsaw system used with a multi-monitor setup. The top view provides a list visualization with connections between people where selected people are highlighted (From: [26]).

9.3.3 Evaluating the Jigsaw Analysis Tool

Another prominent visual analytics tool analyzed with ethnographic methods is JIGSAW, a tool that enables the analysis of large document collections [26, 14]. Clustering is provided where the similarity of documents is analyzed depending on the co-occurrence of words. Documents may be also sorted according to different criteria. Thus, document views, list views, and cluster views are essential components of the systems (see Fig. 58).

The evaluations with three intelligence analysis experts (two from academia, one from industry) lasted between two and fourteen months. Interviews (45-60 min.) were audio-recorded, fully transcribed and carefully analyzed with specialized software to understand core themes [15]. The prepared questions of the semi-structured interviews relate to specific tasks for which Jigsaw is used, the goals of the analysis, the data to be used, features considered essential or superfluous. The analysis with one expert revealed that he employed mostly documents related to a narrow time frame since otherwise the resulting visualizations are overwhelming. It turned out that a feature was missing that allowed users to select/deselect documents for the current analysis. Mostly, the analysis of documents served to understand whether there are relations between two persons and, if so, to better understand what type of relation they have. Graph views that provide a visual interpretation of the data were new to them and appreciated.

The long-term case study provided many insights in the learning process required to use the Jigsaw system and in unexpected pattern of system use.

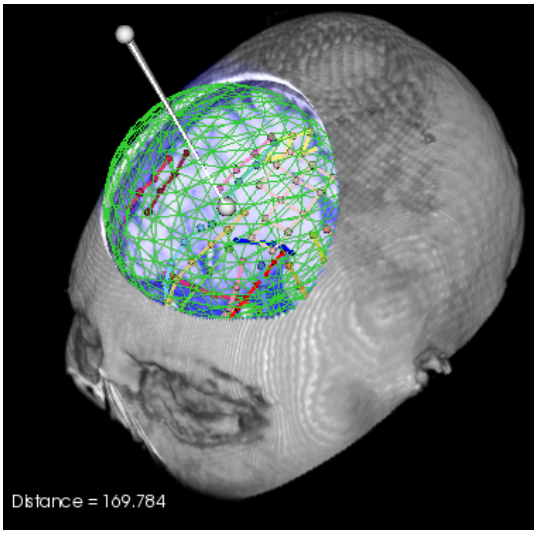

Fig. 59 A figure showing context (EEG electrodes) around the position of the surgical instrument. (From: [13]).

9.3.4 Evaluating the Impact of a Medical Visualization tool

Working in the medical domain requires long-term immersion into a medical facility that frequently leads to a deeper understanding of the primary pipeline for the treatment of a patient. This includes the processes followed at the facility as well as all the individuals involved in the processes. In previous work, Joshi et al. [13] worked closely with neurosurgeons to understand challenges with respect to image-guided surgery. Neurosurgeons, radiologists, neurologists, and technicians are all involved in process of surgical planning and the actual surgery. The researchers identified challenges associated with data representation of all the modalities being used for image-guided surgery such as CT, MRI, EEG electrode strips, and in some cases, PET scans, and DTI imagery. They developed a system that allows contextual-representation of the data during surgery and evaluated it with neurosurgeons and residents [13, 12].

Due to the embedded nature of the researchers involved in the project, other problems related to occlusion in vascular neurosurgery too were identified and addressed [10]. These techniques were incorporated into existing image-guided surgery software and were evaluated over a long-period of time for ease-of-use and adoption. Technicians and surgeons continued to use the technique via the image-guided navigation system.

Expert analysis provided crucial insight into use cases and usability of the system. As the research team continued to work with the surgeons and operating room technicians, other challenges with respect to the ambient lighting in the operating

room were identified and light-sensitive solutions [9] were designed and deployed in the operating room.

These solutions to their problems were identified, resolved, deployed, and evaluated over a two-year period to iteratively improve the image-guided surgery system.

9.3.5 Generalized Experiences

A common result of all long-term evaluations was that experts always started their analysis with clear analytical questions in mind. The visualization researchers have not observed pure exploration activities without any hypothesis. The initial use of the system for the selected data may lead, of course, to interesting or even surprising situations, that stimulate follow-up questions, e.g. to understand a phenomenon in more detail or to confirm a pattern. The analytical questions are very specific for the particular domain but as Valiati et al. [27] point out, most of them can be mapped to rather general visualization tasks, such as gaining an overview, searching for a particular configuration and comparison that were performed at different abstraction levels. This generalization may help to translate the experiences to other areas and enable other researchers to *reproduce* these experiences or find out that the results cannot be confirmed eventually leading to more reliable knowledge about analytical patterns and appropriate computer support. All long-term case studies discussed in this section relied on very few experts. The three evaluations described by Valiati et al. had one expert only. The Rank-by-feature evaluation had six experts, the largest number, we found in such an evaluation.

The overall assessment of long-term evaluations revealed a number of tasks that were not supported well at least by early information visualization systems [27]. Users want to:

- document and record (intermediate) results for themselves or discussion with others,
- emphasize or comment on items, groups of items or relations,
- to verify observations derived from data visualizations with statistical methods

9.4 Planning, Conducting, and Reporting

A long-term case study obviously requires careful planning and sufficient time. It is likely that we rarely see this type of evaluation since it does not nicely fit in the tight schedule of paper publishing where the implementation is often finished only a few weeks before the deadline. The most important aspect is the recruiting of experts to use the system for a longer time. These experts need to either be the target users or be representative, in particular they should have approximately the qualification and experience, of the target users. Sometimes, the few top level experts

for a special domain are not available and are replaced by users with a little lower experience. However, if the system is intended for users with long term experience and responsibility to make decisions, students or junior researchers in this area are not representative enough.

Once the users are selected, the evaluation and documentation procedures organizational issues should be discussed (see hints in [25]). The software needs to be prepared carefully, including a short tutorial/documentation to enable autonomous use, logging capabilities, and testing. Since realistic tasks should be investigated, the selection of specific goals, tasks and data is the responsibility of the domain expert. However, discussions between visualization researcher and the domain expert are required to ensure that the data and tasks are representative.

This involves a considerable effort on the side of the domain expert and consequently, publications in their scientific domain are a typical result [11, 9].

9.4.1 Reporting

Reporting on a case study requires considerable thoughts as well as. According to Isenberg et al [8] the following aspects are crucial in any type of reporting on empirical evaluation:

- be specific about your domain experts (age, gender, qualification, experience in the domain and with similar software, ...),
- be specific about the nature of your relation to them, e.g. Are they co-authors of the paper? Are they independent or part of the same institution/project?
- be careful with definitive statements and try to include proper statements of uncertainty when justified.

In addition, we recommend additional components for reporting based on Valiati et al. [27] who described three MILC evaluations in a standardized manner.

- description of the data used by the experts, e.g. number of dimensions, number of datasets, size of a document collection
- analytics questions that the experts tried to answer
- severe usability problems that may have avoided that experts could analyze the data in the way they originally wanted to perform

9.5 Challenges and Limitations

A major challenge of long-term evaluations is that very self-disciplined users are needed that are willing and able to document over a longer time why they used the system, what they considered satisfying, surprising or frustrating. Users often stop the evaluation earlier than expected [2][17]. Participants of long-term evaluations

are not only few, but often also more tech-savvy than average users leading to a selection bias that further reduces the generalizability of the results.

Another limitation is that due to the small number of test persons, there is more randomness involved than in a lab study, i.e. it is a bit by chance how the system is actually used and which data are used. The environment in which case study work is performed is realistic, but not controllable. Thus, statistical analysis is typically not meaningful. ”The outcomes should not be too generalized" as Elmqvist et al. argue [5]. Since only few users are involved, long term case studies do not help to characterize the use of system for a diverse set of users that differ e.g. in their spatial ability.

To ease the burden on the target users of the system, developers could consider automating the data collection process through system logs and infrequent face-to-face meetings with the users. This would provide insight into whether a deployed system is being used as well as identify pain points for users that may be preventing them from using the system.

9.5.1 Combinations with Other Methods

Long-term case studies have a number of advantages that were stated in the introduction and motivate this chapter. However, as Carpendale [3] points out, no single evaluation method can fully characterize the value of interactive visualization systems. Long-term case studies enable realistic observations, but they are not precise. Even the MILC variant (recall [25]) that combines a number of methods within a case study evaluation remains limited. Therefore, combinations with other methods are relevant.

Instead of discussing all possible combinations, we will focus on one combination that is particularly relevant for visual analytics systems that often aim at discovery processes. The combination with an explicit recording of *insights* is a natural choice and provides a clear focus for long-term case studies. The number and quality of such insights, e.g. whether insights are surprising and can be verified, is considered as an evaluation measure in *insight-based evaluations* [22]. The original insight-based evaluations were lab experiments where analysts should freely use the system (after appropriate training) to find interesting relations. Seo et al. [24] combined the MILC evaluation with insight-based analysis of the rank-by-feature framework. This combination is promising since discovery processes often are not very effective when restricted to a limited amount of time. This combination, however, does not solve the major problem of long-term case studies, namely that they comprise only a very few participants. Therefore, Seo et al. [24] added a broader survey where they asked a larger number of users, again authors of publications that employ their tool, to take part in an interview [24]. This interview cannot provide such a rich description of system use as in the long-term evaluation, but since much more participants are involved, more reliable and generalizable statements about usage patterns and usefulness can be derived.

The conduction of insight-based long-term case study evaluations has to consider many aspects (see [3] for a discussion). A crucial question is when analysts are interviewed with respect to what they have learned about the domain using a visual analytics system and a selection of datasets. Insights may occur suddenly, but also hours or even days after a system was used. Of course, the insights that were gained strongly depend on the domain knowledge of the analyst, her motivation and creativity. Preim et al. [20] provide a discussion of the evaluation practice in medical visualization, where long-term case studies and their combination with other methods are discussed.

9.6 Conclusions

Long-term case study is a viable empirical evaluation method for visualization systems that enables an understanding of cognitive activities, such as problem-solving and decision making. Long-term case studies in visualization research has some unique aspects compared to applications in human computer interaction, e.g. discovery processes in visual analytics applications. Thus, we discussed primarily such visualization examples and hope to stimulate further attempts in this direction.This qualitative and observational evaluation method overcomes many limitations of lab-based studies and enables a deep understanding of system use. It can be adapted to different time-frames and budgets ranging from several weeks to a few years. The observation of users doing real work in their familiar (work) context is a key aspect. The MILC variant described by Shneiderman and Plaisant [25] provides guidance how to perform such evaluations in an informative manner. Since only a few users are involved and the working environment cannot be controlled, long term case studies are limited. A combination with other evaluation methods, e.g. questionnaires, allows to derive quantitative assessments. Long-term case studies were successfully used in a number of InfoVis and visual analytics applications. In other areas, particularly, in scientific visualization applications, the method is underutilized but promising as well.

Long-term case studies evolve into continuous use of a deployed system only if the researchers are immersed and have clearly addressed an existing problem in the workflow of the target users. The maintenance and iterative development of the system in conjunction with the end users results in successful outcomes. If you would like your system to be used for a long period of time, you have to be willing to maintain and support it for that same duration as well.

References

1. Bentley, R., Hughes, J.A., Randall, D., Rodden, T., Sawyer, P., Shapiro, D., Sommerville, I.: Ethnographically-informed systems design for air traffic control. In: Proc. of the 1992 ACM

Conference on Computer-supported Cooperative Work, pp. 123–129 (1992)
2. Breakwell, G.M., Hammond, S., Fife-Scha, C.: Research methods in psychology. Thousand Oaks, Calif. : Sage Publications (1995)
3. Carpendale, S.: Evaluating information visualizations. In: Information visualization, pp. 19–45. Springer (2008)
4. Crabtree, A., Rodden, T., Tolmie, P., Button, G.: Ethnography considered harmful. In: Proc. of the ACM SIGCHI conference on human factors in computing systems, pp. 879–888 (2009)
5. Elmqvist, N., Yi, J.S.: Patterns for visualization evaluation. Information Visualization **14**(3), 250–269 (2015)
6. González, V., Kobsa, A.: A workplace study of the adoption of information visualization systems. In: Proc. I-KNOW (Vol. 3), pp. 92–102 (2003)
7. Hughes, J., King, V., Rodden, T., Andersen, H.: The Role of Ethnography in Interactive Systems Design. interactions **2**(2), 56–65 (1995)
8. Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., Möller, T.: A systematic review on the practice of evaluating visualization. tvcg **19**(12), 2818–2827 (2013)
9. Joshi, A., Papanastassiou, A., Vives, K., Spencer, D., Staib, L., Papademetris, X.: Light-sensitive visualization of multimodal data for neurosurgical applications. In: IEEE International Symposium on Biomedical Imaging (ISBI) (2010)
10. Joshi, A., Qian, X., Dione, D.P., Bulsara, K.R., Breuer, C.K., Sinusas, A.J., Papademetris, X.: Effective visualization of complex vascular structures using a non-parametric vessel detection method. In: IEEE Transactions on Visualization and Computer Graphics (VIS 2008), vol. 14(6) (2008)
11. Joshi, A., Scheinost, D., Okuda, H., Murphy, I., Staib, L.H., Papademetris, X.: Unified framework for development, deployment and testing of image analysis algorithms. In: MICCAI Workshop on Systems and Architectures for Computer Assisted Interventions (2009)
12. Joshi, A., Scheinost, D., Spann, M., Papademetris, X.: Evaluation of multi-viewport based visualization for electrode navigation during stereotactic image guided neurosurgery. In: International Brain Mapping and Interoperative Surgical Planning Society's 6th World Congress for Brain Mapping and Image Guided Therapy (2009)
13. Joshi, A., Scheinost, D., Vives, K.P., Spencer, D.D., Staib, L.H., Papademetris, X.: Novel interaction techniques for neurosurgical planning and stereotactic navigation. In: IEEE Transactions on Visualization and Computer Graphics (VIS 2008), vol. 14(6) (2008)
14. Kang, Y.a., Gorg, C., Stasko, J.: Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In: Proc. of IEEE Visual Analytics Science and Technology, pp. 139–146 (2009)
15. Kang, Y.a., Stasko, J.: Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts. tvcg **18**(12), 2869–2878 (2012)
16. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: Seve scenarios. tvcg **18**(9), 1520–1536 (2012)
17. Ohly, S., Sonnentag, S., Niessen, C., Zapf, D.: Diary studies in organizational research: An introduction and some practical recommendations. Journal of Personnel Psychology **9**, 79–93 (2010)
18. Perer, A., Shneiderman, B.: Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In: Proceedings of the ACM SIGCHI conference on Human Factors in computing systems, pp. 265–274 (2008)
19. Preim, B., Baer, A., Cunningham, D., Isenberg, T., Ropinski, T.: A survey of perceptually motivated 3d visualization of medical image data. Computer Graphics Forum **35**(3), 501–525 (2016)
20. Preim, B., Isenberg, P., Ropinski, T.: A critical analysis of the evaluation practice in medical visualization. In: Proceedings of the EG Workshop on Visual Computing in Biology and Medicine (2018)
21. Saalfeld, P., Luz, M., Berg, P., Preim, B., Saalfeld, S.: Guidelines for quantitative evaluation of medical visualizations on the example of 3d aneurysm surface comparisons. Computer Graphics Forum **37** (2018)

22. Saraiya, P., North, C., Lam, V., Duca, K.A.: An insight-based longitudinal study of visual analytics. tvcg **12**(6), 1511–1522 (2006)
23. Seo, J., Shneiderman, B.: A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In: Proc. of IEEE Symposium on Information Visualization, pp. 65–72 (2004)
24. Seo, J., Shneiderman, B.: Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. tvcg **12**(3), 311–322 (2006)
25. Shneiderman, B., Plaisant, C.: Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In: Proc. of the Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (2006)
26. Stasko, J.T., Görg, C., Liu, Z., Singhal, K.: Jigsaw: Supporting investigative analysis through interactive visualization. In: Proc. of the IEEE Symposium on Visual Analytics Science and Technology, pp. 131–138 (2007)
27. Valiati, E.R., Freitas, C.M., Pimenta, M.S.: Using multi-dimensional in-depth long-term case studies for information visualization evaluation. In: Proc. of the Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization, p. 9 (2008)
28. Whittaker, S., Sidner, C.L.: Email overload: Exploring personal information management of email. In: Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 276–283 (1996)